Covid-19 Document Classification

Applying BERT-style Transfer Learning to Covid-19 Related Publications

Helena Balabin, Natural Language Processing (Summer Semester 2020)

Introduction

LitCovid Database [1]

https://www.ncbi.nlm.nih.gov/research/coronavirus/docsum

LITCOVID

Ex: Remdesivir

NIH NLM

GENERAL MECHANISM TRANSMISSION DIAGNOSIS TREATMENT PREVENTION CASE REPORT FORECASTING

Chemicals

- Hydroxychloroquine (510)
- Chloroquine (263)
- remdesivir (192)
- lopinavir-ritonavir drug combination (159)
- tocilizumab (149)

Journals

- BMJ (619)
- J Med Virol (427)
- Nature (255)
- Lancet (250)
- J Infect (220)

Countries

- China (3601)
- United States (1288)
- Italy (1021)
- United Kingdom (465)
- Spain (327)

Showing 1 to 10 of 27040 publications.

Page 1 of 2704

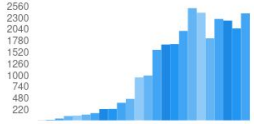
1. JUN 28, 2020 (ADDED TO PUBMED)
   Comparative analysis of protein synthesis rate in COVID-19 with other human coronaviruses.
   Dasari, Chandra Mohan; Bhukya, Raju • Infect Genet Evol
   MECHANISM • TREATMENT • TRANSMISSION

2. JUN 28, 2020 (ADDED TO PUBMED)
   Direct SARS-CoV-2 infection of the heart potentiates the cardiovascular sequelae of COVID-19.
   Bose, Rajendran Jc; McCarthy, Jason R • Drug Discov Today
   TREATMENT • MECHANISM

3. JUN 28, 2020 (ADDED TO PUBMED)
   Convalescent plasma for persisting Covid-19 following therapeutic lymphocyte depletion: a report of rapid recovery.
   Clark, E; Guilpain, P ... Le Moing, V • Br J Haematol
   CASE REPORT

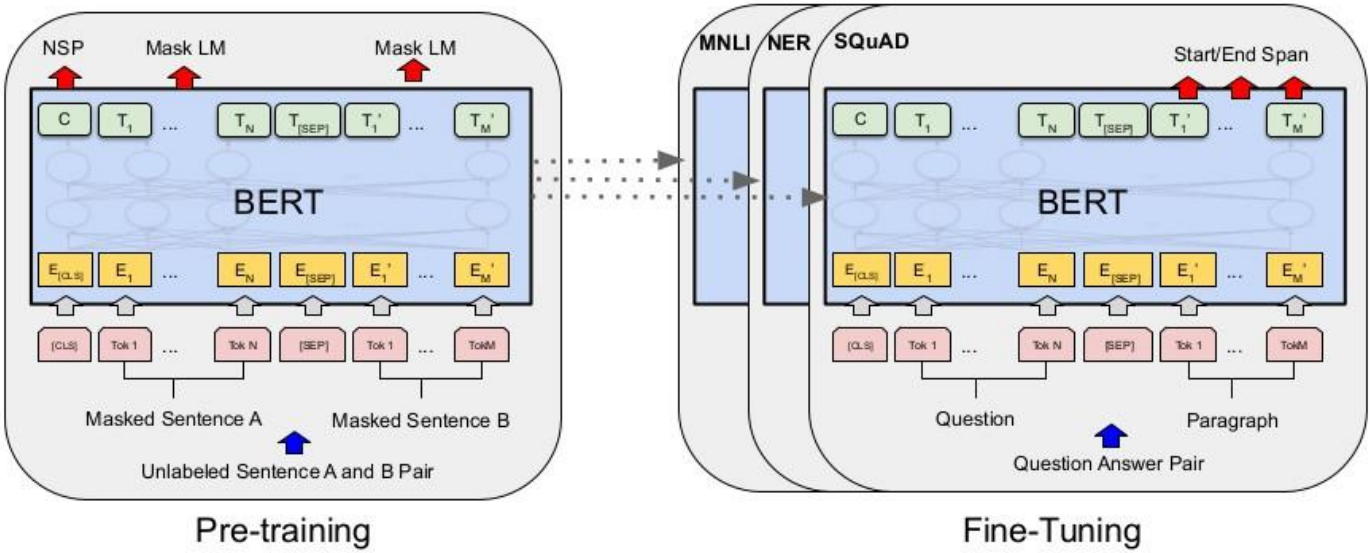
DOWNLOAD RSS

WEEKLY PUBLICATIONS



WORLD MAP

Transfer Learning in the Context of Document Classification [2]



BERT-style Models

Model	Pre-Training Corpora
BERT [2]	BooksCorpus (800M words), English Wikipedia (2,500M words)
BioBERT [3]	BooksCorpus (800M words), English Wikipedia (2,500M words) + PubMed abstracts (4500M words)
CovidBERT [4]	BooksCorpus (800M words), English Wikipedia (2,500M words) + CORD-19 dataset (≈13000 full-text publications)

Hypothesis

Hypothesis

Increased Covid-19 Relatedness of Training Corpora

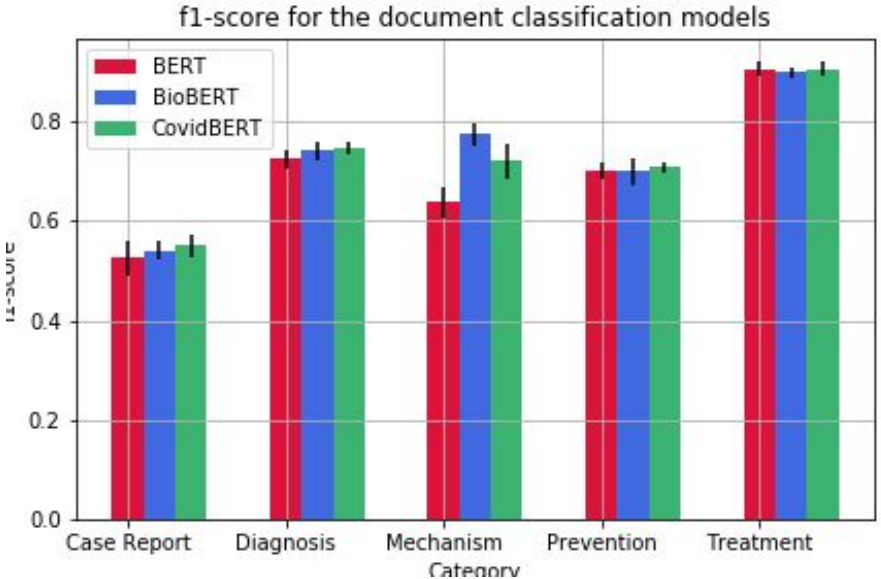
BERT **BioBERT** **CovidBERT**

Performance

Results

F1-score per Category

→ The performances of the three models do not differ significantly (except for the “Mechanism” category)



roughly 6500 abstracts

Modified Hypothesis

Modified Hypothesis

Including pre-trained word representations

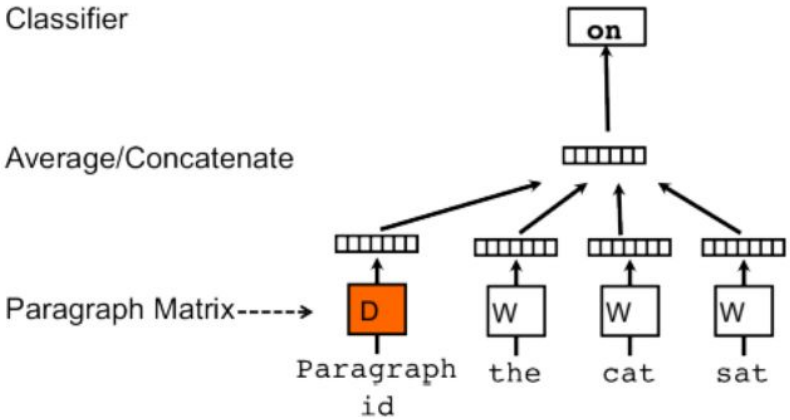
Building an entirely task-specific model

Fine-tuning of pre-trained models

Performance

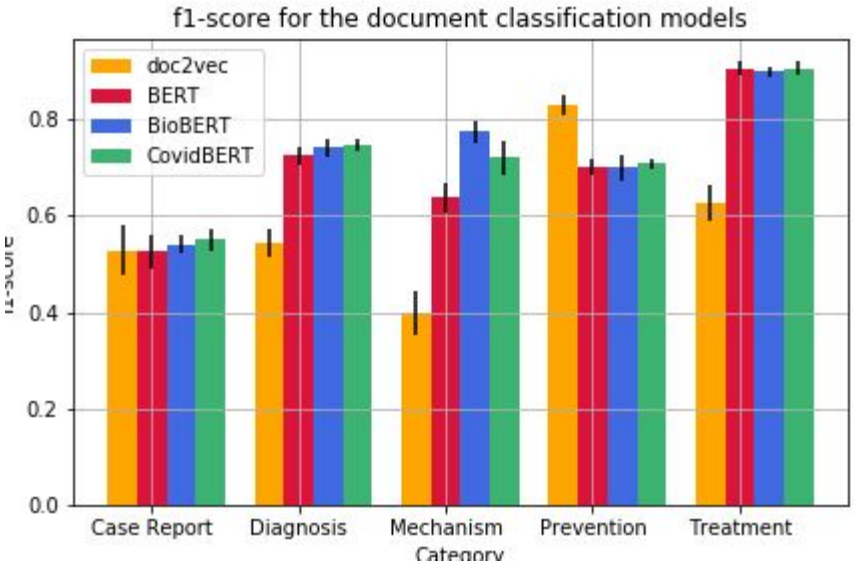
Excursus: Doc2Vec [5]

- Similar to Word2Vec
- Additionally Use a Paragraph ID vector



F1-score per Category (Including the Doc2Vec Baseline)

- **Better** performance of the BERT-style models: Diagnosis, Mechanism, Treatment
- **Equal**: Case Report
- **Worse**: Prevention



Trying to Find an Explanation

Excursus: t-distributed Stochastic Neighbor Embedding (t-SNE) Dimensionality Reduction [6]

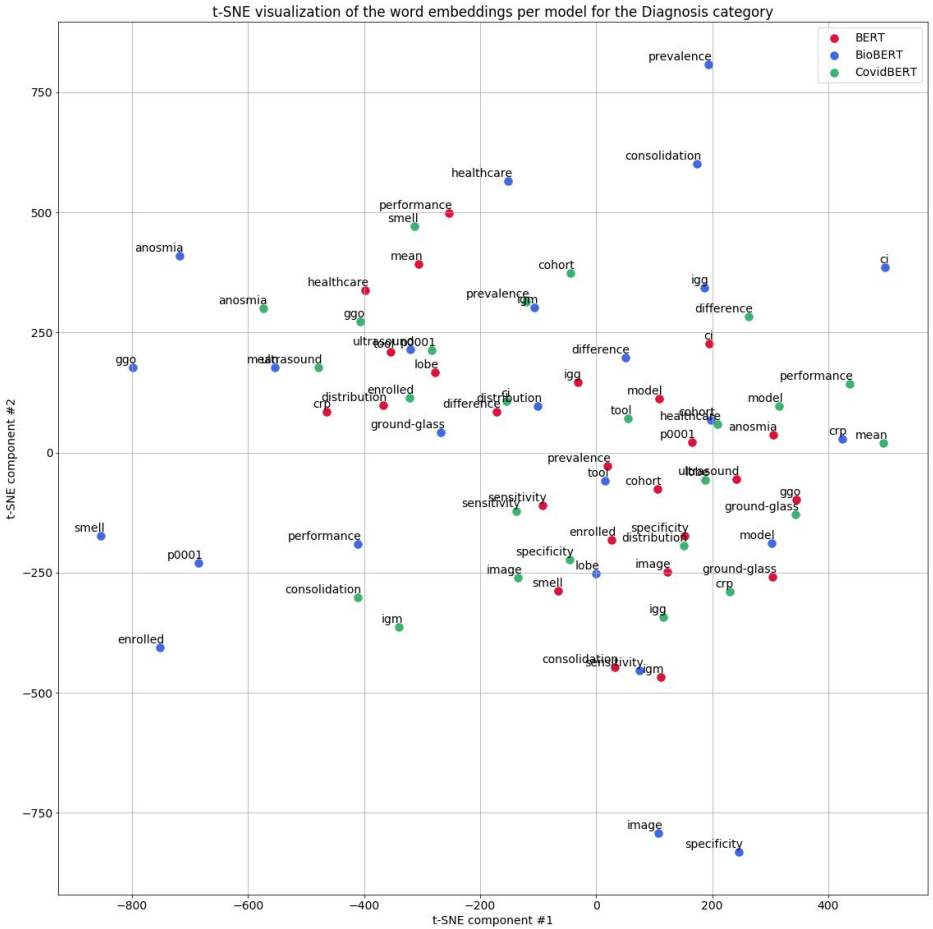
*"The similarity of datapoint $\boldsymbol{x_j}$ to datapoint $\boldsymbol{x_i}$ is the **conditional probability** $\boldsymbol{p_{j|i}}$ that $\boldsymbol{x_i}$ would pick $\boldsymbol{x_j}$ as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at $\boldsymbol{x_i}$."*

→ similarity measure based on the **Student t-distribution**

→ non-linear projection of data points into lower dimensional space with the objective of **minimizing the Kullback-Leibler divergence**

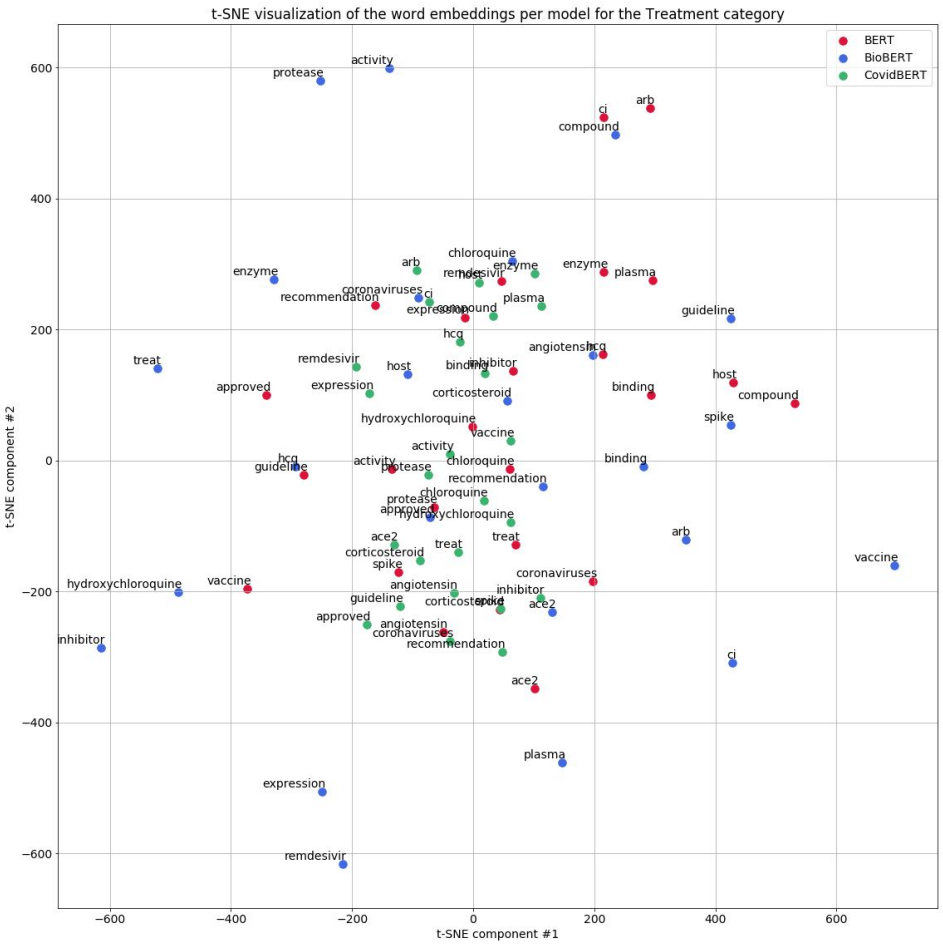
Exemplary t-SNE Visualization for the “Diagnosis” Category

→ Covid-BERT and BERT have a smaller convex hull than BioBERT



Exemplary t-SNE Visualization for the “Treatment” Category

→ Covid-BERT has the smallest convex hull again



Other Explanations

- Abstracts are not distinctive enough → Performance might differ on full-text
- Quality of Covid-19 publications is an issue
- Additional fine-tuning corpora are too small and/or do not add much value

Takeaways

Takeaways

- Domain-specific fine-tuning might not be as effective as it seems
- Using pre-trained (BERT-style) models increases the classification performance
- t-SNE is sensitive on hyperparameter settings and hard to interpret

Questions?

References

References

[1] "LitCovid - NCBI - NLM - NIH." https://www.ncbi.nlm.nih.gov/research/coronavirus/ (accessed May 26, 2020).

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," p. 16, May 2019.

[3] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, p. btz682, Sep. 2019.

[4] "deepset/covid_bert_base · Hugging Face." https://huggingface.co/deepset/covid_bert_base (accessed May 26, 2020).

[5] Q. Le, and T. Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning,* pp. 1188-1196, Jan. 2014.

[6] L. van der Maaten, and G. Hinton. "Visualizing data using t-SNE." *Journal of machine learning research,* pp. 2579-2605, Nov. 2008.