sebis

TUM

Evaluation of Different Approaches to Training BERT for Classification Tasks in the Engineering Domain

Sergii Poluektov, 26.04.2021, Kick-off Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Outline

- Motivation
- Background
  - BERT
  - Domain-specific BERTs
- Research Questions
- Methodology
- Current progress
  - Datasets
  - Vocabulary discrepancy
  - Established baselines

Outline

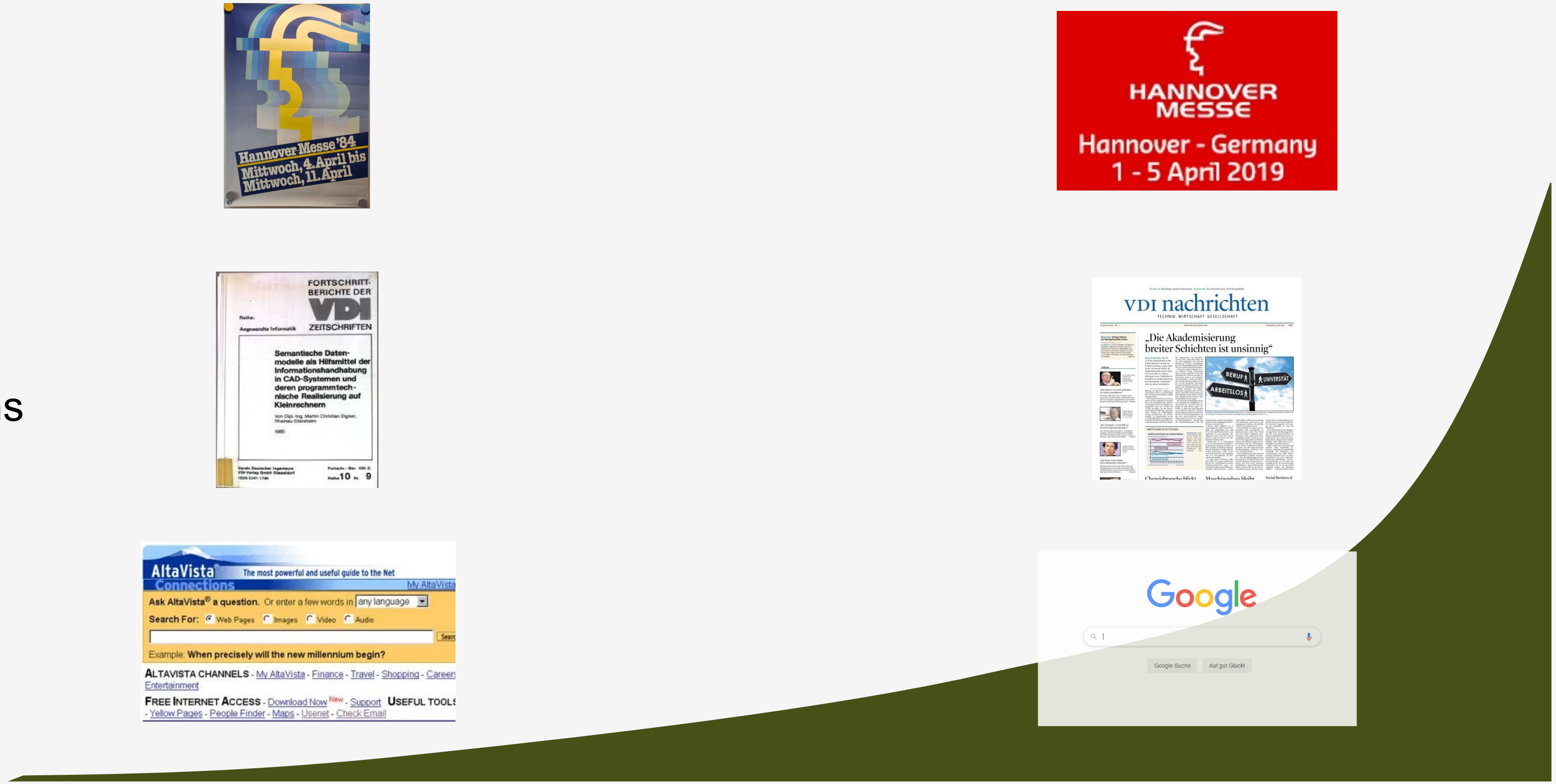
- Motivation
- Background
  - BERT
  - Domain-specific BERTs
- Research Questions
- Methodology
- Current progress
  - Datasets
  - Vocabulary discrepancy
  - Established baselines

Motivation

Trade fairs

Trade publications

Search engines



Data Volume

1990

2020

[1]

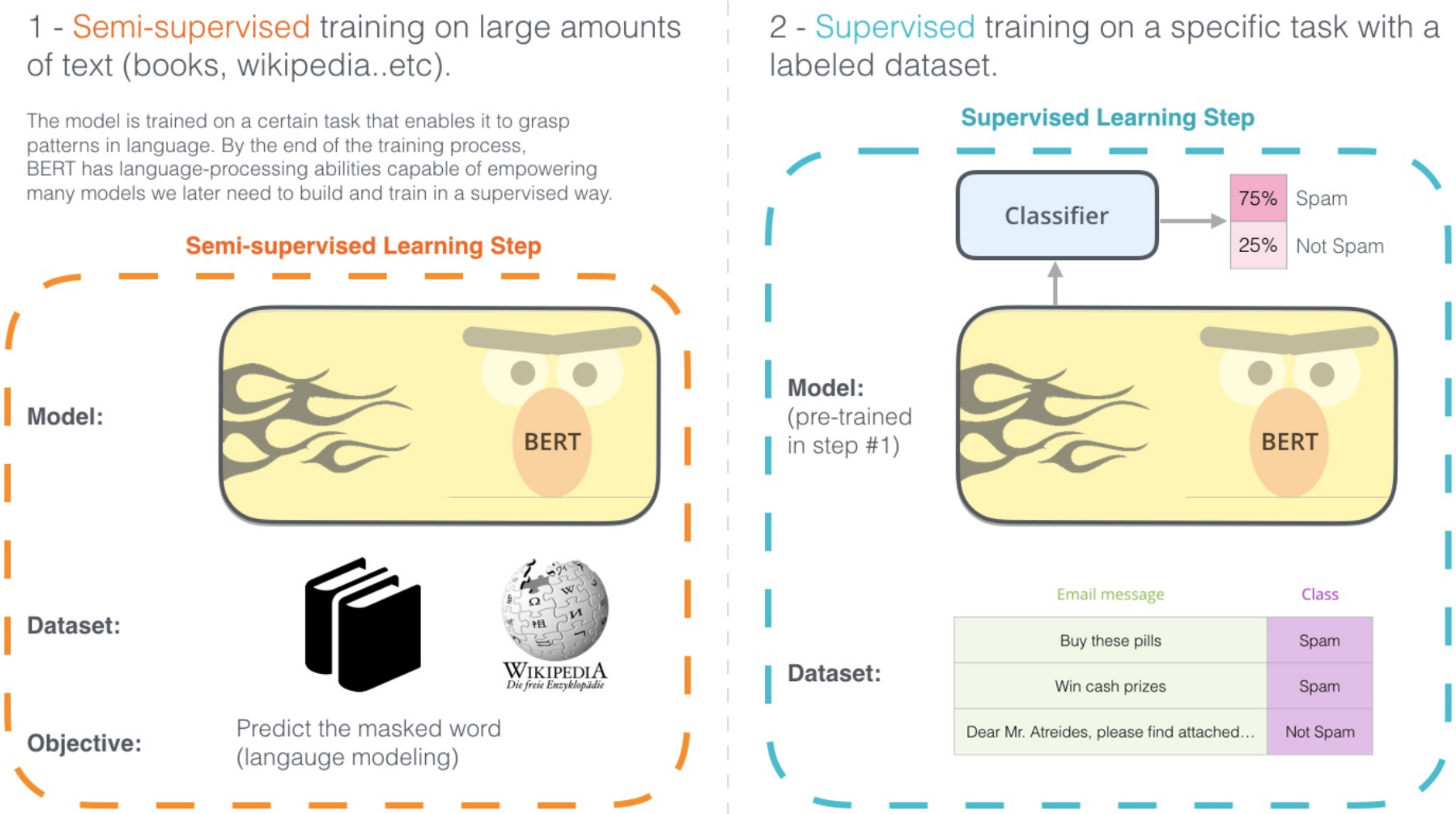
Motivation

[1]

Outline

- Motivation
- Background
  - BERT
  - Domain-specific BERTs
- Research Questions
- Methodology
- Current progress
  - Datasets
  - Vocabulary discrepancy
  - Established baselines

Background: BERT



[2]

Background: Domain-specific BERTs

SciBERT

Training corpus: Semantic Scholar (18% computer science, 82% biomedical)

Approach: train from scratch

Results: outperforms BERT in classification tasks on all considered datasets, achieves SOTA on some of them

[3]

ClinicalBERT

Training corpus: MIMIC-III dataset (health records of hospital admissions)

Approach: train from scratch

Results: outperforms BERT at readmission prediction

[4]

FinBERT

Training corpus:TRC2—financial (subset of Reuters' TRC2, which consists of news artcles filtered for financial keywords)

Approach: further pre-train

Results: outperforms other transfer learning methods in sentence classification.

[5]

Outline

- Motivation
- Background
  - BERT
  - Domain-specific BERTs
- Research Questions
- Methodology
- Current progress
  - Datasets
  - Vocabulary discrepancy
  - Established baselines

Research Questions

1 Does a BERT model pre-trained on texts from engineering domain perform better on given classification tasks compared to the standard BERT?

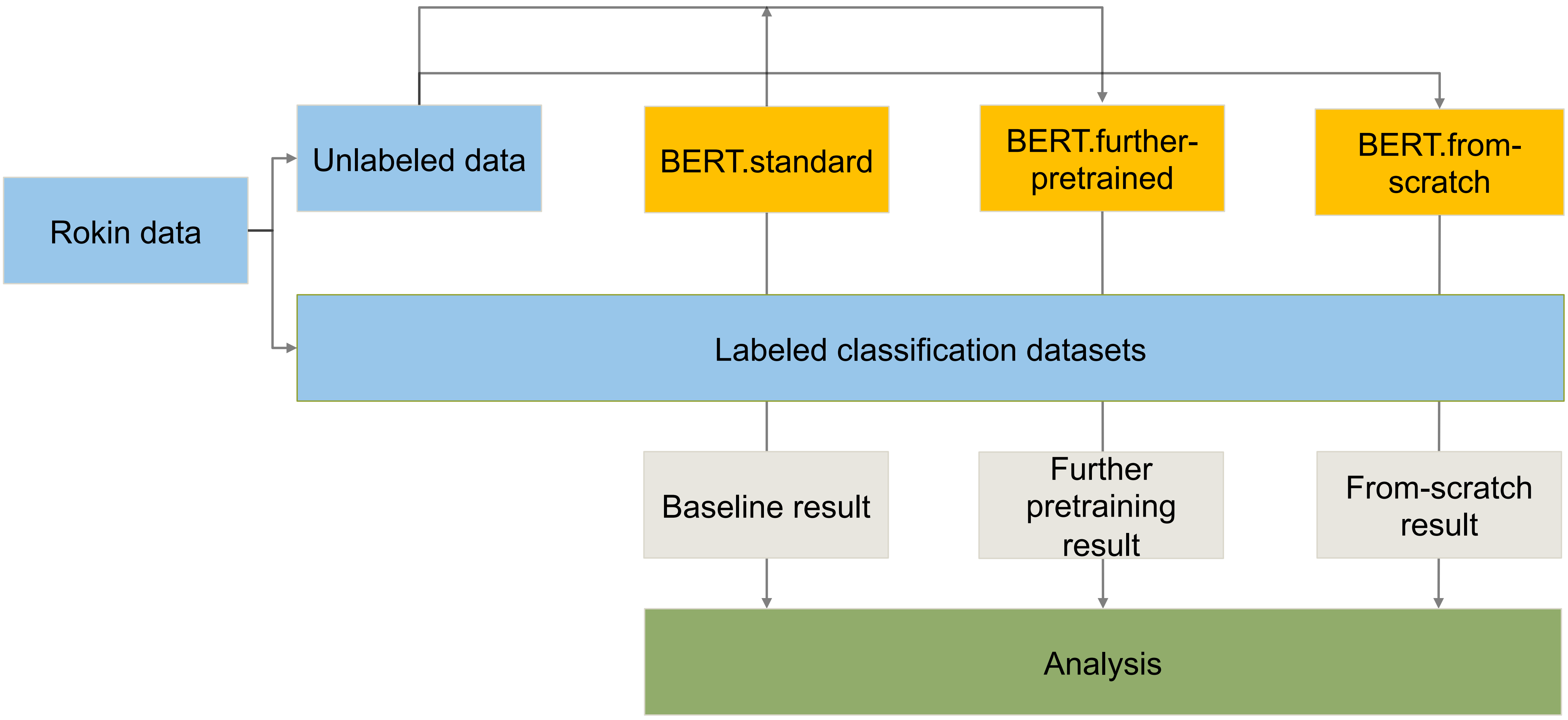
2 Can similar results be achieved by further-pre-training with less data?

3 Can similar results be achieved by fine-tuning standard BERT on bigger labeled datasets?

Outline

- Motivation
- Background
  - BERT
  - Domain-specific BERTs
- Research Questions
- Methodology
- Current progress
  - Datasets
  - Vocabulary discrepancy
  - Established baselines

Methodology



Outline

- Motivation
- Background
  - BERT
  - Domain-specific BERTs
- Research Questions
- Methodology
- Current progress
  - Datasets
  - Vocabulary
  - Baselines

Current progress: Datasets

Articles:

unlabeled

~1.8 M articles from engineering magazines.

Article classification:

labeled

2800 articles labeled as ‘good’ (includes information about a new technology) and `bad` (doesn’t include information about new technology)

Article topic classification:

labeled

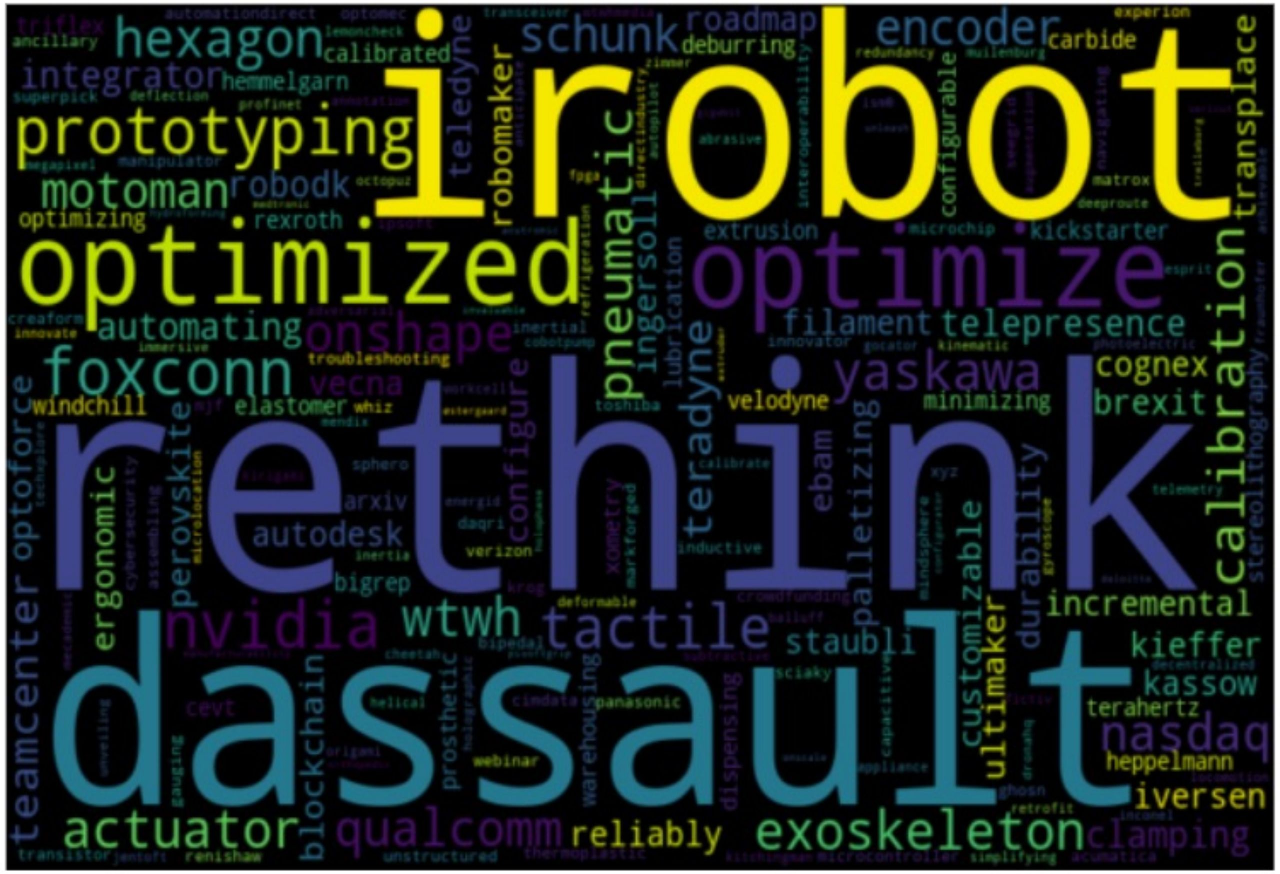
250 articles with assigned topics (i.e. Robotics, Sensors, AR, etc.)

NER:

labeled

200 articles with annotated named entities (Organisation, Product, Person, Material, Event)

Current progress: Vocabulary (not in BERT)



Current progress: Baselines

Article classification:

Model: BERT-base-uncased
+ classification head

Accuracy: 0.89
Precision: 0.67
Recall: 0.75
F1: 0.71

Article topic classification:

Model: BERT-base-uncased
+ classification head

Accuracy: 0.82

NER:

in progress…

https://wandb.ai/gjke/Thesis
https://github.com/Rokin-Tech/Rokin_Dev/tree/RTDev_Sergii

Sources

[1] Rokin https://en.rokin.tech

[2] The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) http://jalammar.github.io/illustrated-bert/

[3] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

[4] Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

[5] Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Sergii Poluektov

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

sergii.poluektov@tum.de
wwwmatthes.in.tum.de