OpenFact – AI tools for verification of veracity of information sources and fake news detection.
Financed by National Center for Research and Development in Poland (INFOSTRATEG-I/0035/2021-00).

Finetuning Llama 2 with PEFT QLoRa method for detecting Check-worthy Claims

Follow-up on CheckThat! Lab at CLEF 2023

Marcin Sawiński

Poznań University of Economics and Business

2023-10-25

Experiments

- Setup cloud infrastrcuture for running custom Llama 2 Models
- Prepare training and infernece pipeliens for 7B, 13 B and 70B model variants.
- Use 3 datasets variants (full, 2:1 and 1:1)

Curating dataset - fewer, better data

- Original train data set size - 16821
- Curated train data set size 2:1 NCS/CS - 7692
- Downsampled train data set size 1:1 NCS/CS (randomly picked)

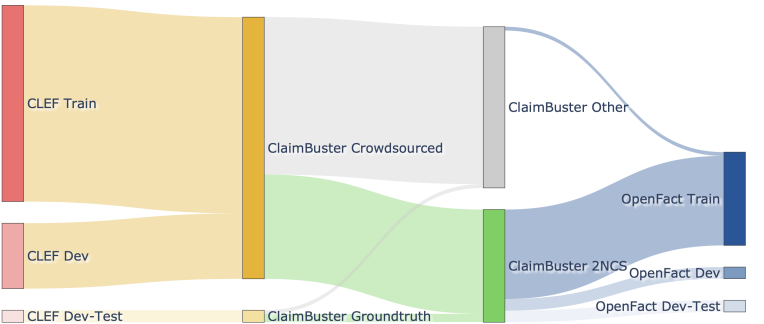
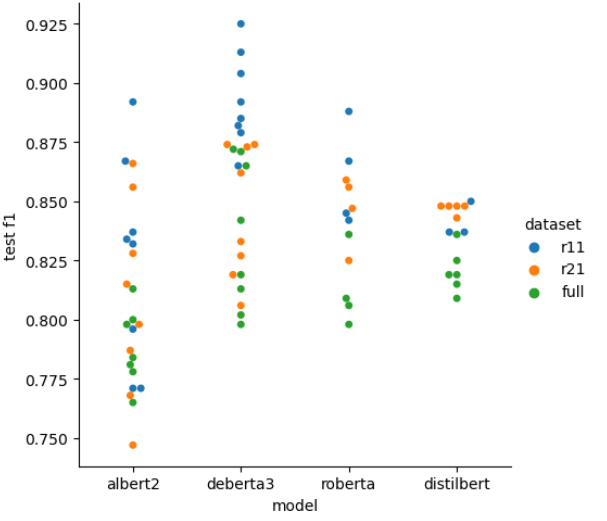


Figure: Reshuffling of 2:1 dataset

Stating point - Perspectives for future work

- More resources / bigger models / smaller models
- Examine dataset curation impact
- Chain-of-Thought and beyond

Downsampled datasets further increased detecion f1 score



Fine-tuning Llama 2

Llama 2 is pretreained model from Meta[1], trained on corpus of 2 Trillion tokens. Further finetuning was performed usign 1'000'000 human annotations. Inference infra chosen in GCP:

- abc

[1] https://ai.meta.com/llama/

Hyperparamters

Hyperparameter	Value
Batch size	8
Learning rate multiplier	0.1
Epochs	4
Prompt loss weight	0.01
Compute classification metrics	True

Table: Hyperparameters used for fine-tuning GPT-3 models

Experiments results

Model	F1	precision	recall	accuracy
GPT-3 curie fine-tuned curated	0.898	0.948	0.852	0.934
DeBERTa v3 base fine-tuned	0.894	0.978	0.824	0.934
GPT-3 davinci fine-tuned curated	0.876	0.946	0.815	0.921
RoBERTa base fine-tuned	0.862	0.966	0.778	0.915
RoBERTa base fine-tuned with custom optimizer layer-wise learning rate decay	0.860	0.976	0.769	0.915
LightGBM ensemble of all BERT-based models and additional embeddings	0.854	0.976	0.759	0.912
ELECTRA fine-tuned	0.851	0.954	0.769	0.909
AlBERT large v2 fine-tuned	0.848	0.976	0.750	0.909
DistilBERT base uncased fine-tuned	0.827	0.952	0.731	0.896
GPT-3 curie fine-tuned random	0.826	1.000	0.704	0.899
GPT neo 125M fine-tuned	0.800	0.961	0.685	0.884
GPT-4 few-shot learning	0.788	0.867	0.722	0.868
GPT-4 zero-shot learning	0.778	0.710	0.861	0.833
GPT-4 Chain-of-Thought	0.722	0.574	0.972	0.745

Experiments results on curated dataset

Model	f1	precision	recall	accuracy
GPT-3 curie fine-tuned curated	0.898	0.948	0.852	0.934
RoBERTa base curated	0.896	0.968	0.833	0.934
DeBERTa v3 base fine-tuned	0.894	0.978	0.824	0.934
GPT-3 davinci fine-tuned curated	0.876	0.946	0.815	0.921
RoBERTa base fine-tuned	0.862	0.966	0.778	0.915
GPT-3 curie fine-tuned random	0.826	1.000	0.704	0.899
DeBERTa v3 base curated	0.818	0.900	0.750	0.887