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).

OpenFact at CheckThat! 2023: Head-to-Head GPT vs. BERT - A Comparative Study of Transformers Language Models for the Detection of Check-worthy Claims

Notebook for the CheckThat! Lab at CLEF 2023

Marcin Sawiński Krzysztof Węcel Ewelina Księżniak Milena Stróżyna
Włodzimierz Lewoniewski Piotr Stolarski Witold Abramowicz

Poznań University of Economics and Business

2023-09-21

Experiments

- Tasks: *Check-That! Lab, Task 1B-English*
- Dataset: ClaimBuster (23,533 statements extracted from all U.S. general election presidential debates). Splits:
  - train & dev - ClaimBuster crowd-sourced
  - dev_test - ClaimBuster ground-truth
- Methods:
  - GPT
  - BERT
  - Ensemble

Curating dataset - volume vs quality vs usefulness

- 'Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics' - Swayamdipta 2020
- 'Scaling Laws for Neural Language Models' - Kaplan 2020
- 'Textbooks Are All You Need' - Gunasekar 2023

Curating dataset - fewer, better data

- Original train data set size - 16821
- Curated train data set size 2:1 NCS/CS - 7692

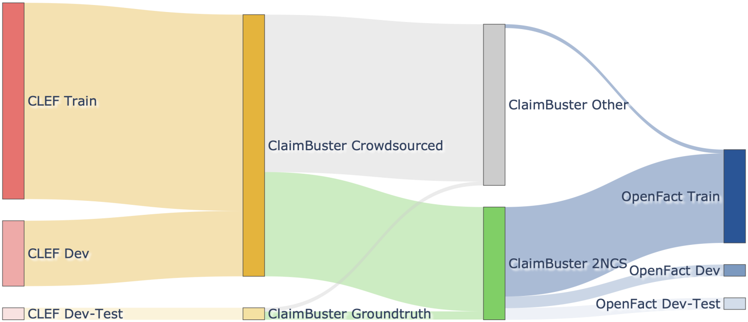

Figure: Reshuffling of 2:1 dataset

GPT in-context learning

- Zero-shot learning using GPT-4
  - System prompt
  - Explanation of task
  - Multiple phrasing variants
  - Mini-batches
- Few-shot learning using GPT-4
  - System prompt
  - Aassistant prompts
  - 4 yes and 4 no examples based on cosine similarity of all-mpnet-base embeddings
- Few-shot learning with Chain-of-Thought using GPT-4
  - multi-step assistant prompts (claim, opinion,topic, topic type, harmful)

Fine-tuning OpenAI GPT-3

- The Curie model - 13 billion parameters, trained using 800GB of text data.
- The Davinci model - 175 billion parameters trained using 45TB of text data3.
- Using only 50% of training data

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

BERT - Model Fine-tuning and Technical Constraints

- Models fine-tuned: DistilBERT, DeBERTa, RoBERTa, XLM-RoBERTa, ALBERT, RemBERT, CamemBERT, ELECTRA, YOSO
- Technical constraints: Local machine setup - four NVIDIA GeForce RTX 2080 Ti GPU cards, 11 GB of memory per card
- Techniques to reduce memory usage: batch size adjustments: 16 to 8 or 4, gradient accumulation float precision: FP16 and FP32

BERT - Hyperparameterization

- Optimizer: AdamW, Adafactor for RemBERT model
- Learning rates: 2e-5 (default), 1e-5, 3e-5
- Fine-tuning duration: 5 epochs (20-30 minutes), extended to 10 epochs if necessary
- Fine-tuning with Layer-wise LR Decay
- Training objective: F1 macro average, F1 positive optimized

Light GBM – Ensemble Approach

- Approach: Combine predictions from fine-tuned models
  - Predicted labels and probabilities
  - Emotion and sentiment probabilities from models BERTemo model
  - Logits returned by ELECTRA discriminator (logit of the first token, the logit of the last token, the minimum logit value, the mean logit value, the maximum logit value, the number of odd tokens (when logit is bigger than zero), and the percentage of odd tokens)
- Best F1 score: 0.79 (despite various hyperparameter settings)
- The most important feature: probabilities from fine-tuned DeBERT-a

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

Perspectives for future work

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

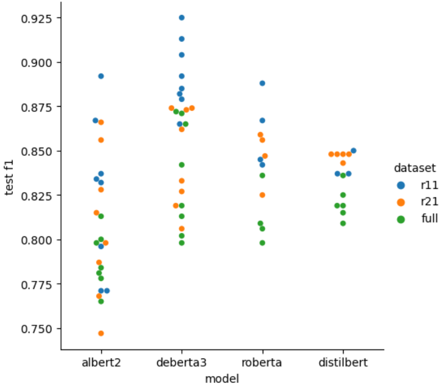

Figure: Further exploration of impact of dataset/annotation quality