Exploring the benefits and challenges of automated classification of political short texts

A case study of conspiracy theory and antisemitic narratives

Helena Mihaljević, Hochschule für Technik und Wirtschaft Berlin
Joint work with Elisabeth Steffen and Milena Pustet
SICSS 2023, 3.7.2023, WZB Berlin Social Science Center

Agenda

- Motivation & projects
- Text classification based on fine-tuning LLMs
- Experimental results
- Data-based challenges
- Is Few-Shot Learning the right way to go?

Motivation

Antisemitic and racist conspiracy theories

- COVID-19 pandemic triggered not only an unprecedented global health crisis, but also increased the spread of disinformation and conspiracy theories (CTs)
- „The Great Exchange“: the pandemic is the strategy of a global elite to replace a Christian white population with non-white and Muslim people. The alleged massive immigration is a strategy of a powerful group imagined as Jewish.
- Racist, anti-Muslim narratives are thus linked to the long-standing notion of a „Jewish world conspiracy“
- The war against the Ukraine brought similar conspiracy narratives (e.g. The war is part of an imagined Jewish world conspiracy, Those in power use the war to distract from Corona etc.)
- CTs often imbued with antisemitic and racist narratives; lead to hate speech and attacks in the digital but also the offline world (drastical examples: Hanau, Christchurch or Halle)

Role of social and online media

- Conspiracy narratives are spreading at an unprecedented scale and speed, made possible by social networks like Facebook and Twitter, video platforms like YouTube, or messenger services like Telegram.
- Example: in the third quarter of 2022, Meta had taken action on 10.6 million pieces of content considered to be hate speech on Facebook. Of these posts, over 90% were found and acted on proactively, prior to users reporting them (Meta 2022)
- Volume of hate speech online, including antisemitic speech, makes automated detection of respective content indispensable, for content moderation, research and monitoring.

Ongoing research projects

- Digitaler Hass: 2 year interdisciplinary project. One task: development of models for automated detection of conspiracy theories in the context of the COVID-19 pandemic
- Decoding Antisemitism: 4 year interdisciplinary, international project. One task: development of models for automated detection of antisemitic content in mainstream online communication

ASH Berlin

HKW Haus der Kulturen der Welt

AMADEU ANTONIO STIFTUNG

ZfA Zentrum für Antisemitismusforschung

HATE LAB

KING'S College LONDON

No detection models available

- In computational research of political texts, conspiracy theories and antisemitism are rather underrepresented
- There is more research on detection of HOT (hateful, offensive, toxic) content
- CT often used synonymously with rumors or misinformation. Existing datasets and annotation schemes not based on a common theoretical foundation. Only around a third of computational works on CT provide a definition, thus analyzing “conspiracy theories online without explicitly defining the main object of their research” (Mahl et al. 2022)
- No datasets connecting antisemitism and CT (~2% of recent empirical studies addressing single conspiracy narratives focus on antisemitic narratives according to Mahl et al. 2022)
- For antisemitism, only recently annotated data from Twitter were published (Jikeli et al. 2023)

No suitable web services available

- There are no openly accessible services explicitly addressing the detection of CT or antisemitic content
- We tested Perspective API for detection of antisemitism, concluding that it does not work for it (Mihaljević & Steffen 2022).

text	insult	identity attack	threat	toxicity	severe toxicity
The Holocaust was unique in its contempt for humanity and its consequences for the world community.[4]	0.67	0.93	0.85	0.79	0.89
The Holocaust did not happen.[5]	0.05	0.07	0.15	0.13	0.04

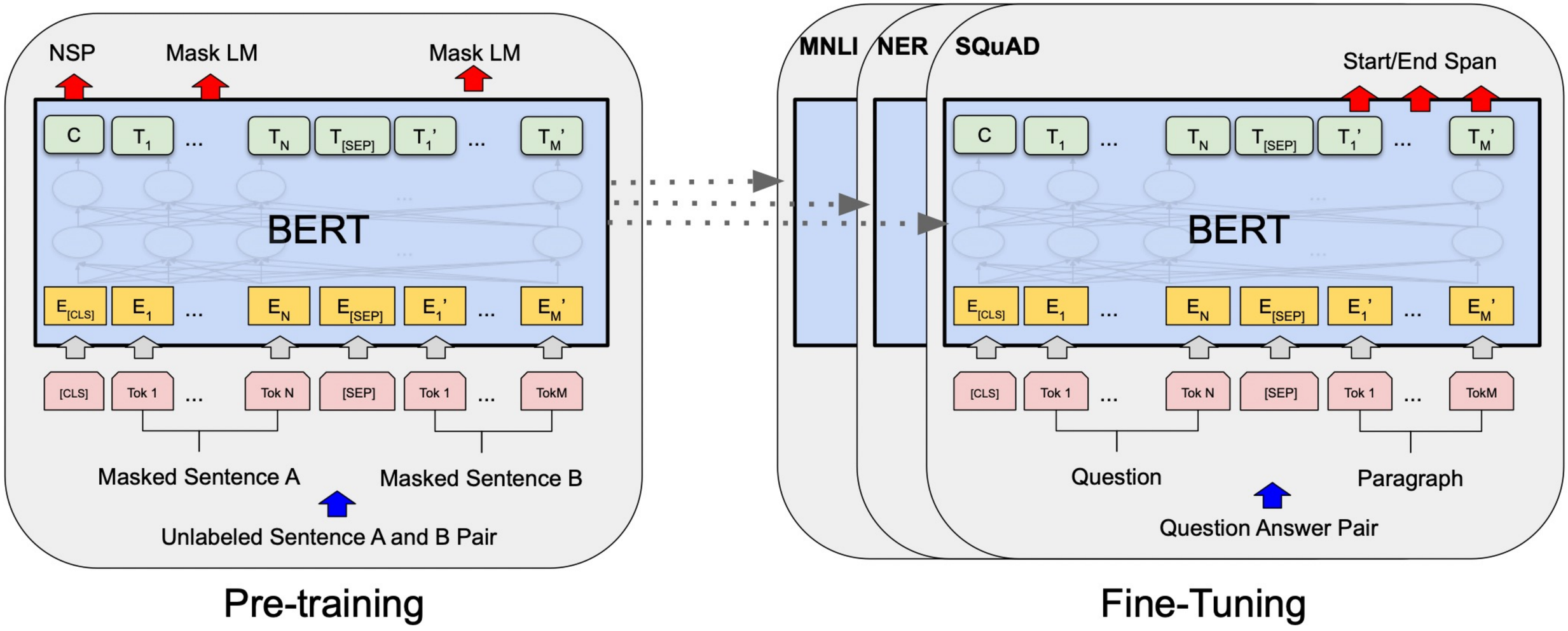
“The US has finally cut bait on the occultist blood suckers. Obama and Trump just drop kicked bibi down to size. This has been a long time coming and that is why the zionists wanted Hitlery to win and start ww3.”

annotated by 53 crowdworkers, with an average toxicity score of 0.33 (Thain et al. 2017)

Text classification through fine-tuning of LLMs

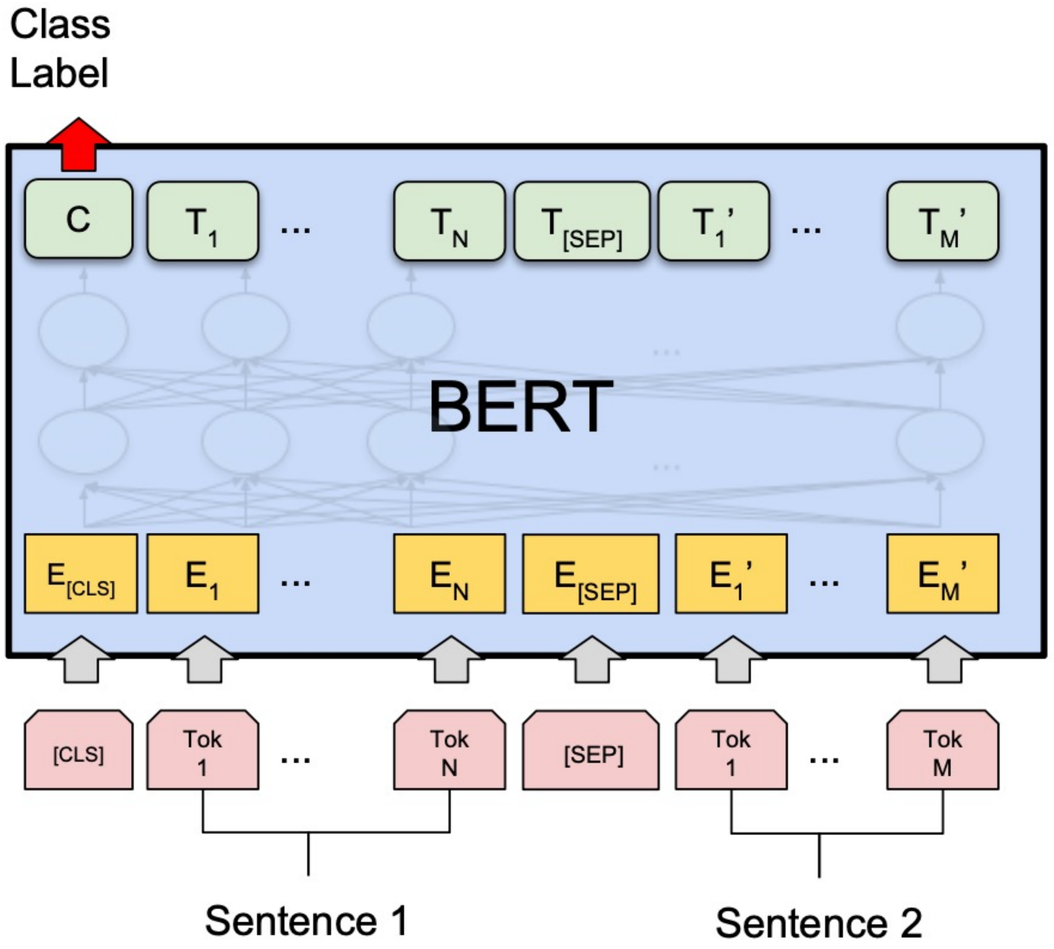
Pre-training and fine-tuning

- Since 2018, the so-called pre-training and fine-tuning approach has substantially improved the training of classification models
- Fine-tuning leverages language models (LMs) that were pre-trained using massive amounts of diverse data on generalist language tasks such as predicting the next word or a masked word in a sentence.
- In pre-training, a LM learns rich representations of language that capture a variety of linguistic phenomena such as word- and sentence-level semantics, syntactic structures, discourse-level phenomena, as well as subtleties of human language like sarcasm or slang
- A pre-trained LLM is adapted in the fine-tuning step to a specific task such as tagging each token in a sentence with respect to a grammar scheme or to classify texts
- Advantage w.r.t. classical supervised learning: better performance (especially on out-of-distribution data) less labeled data

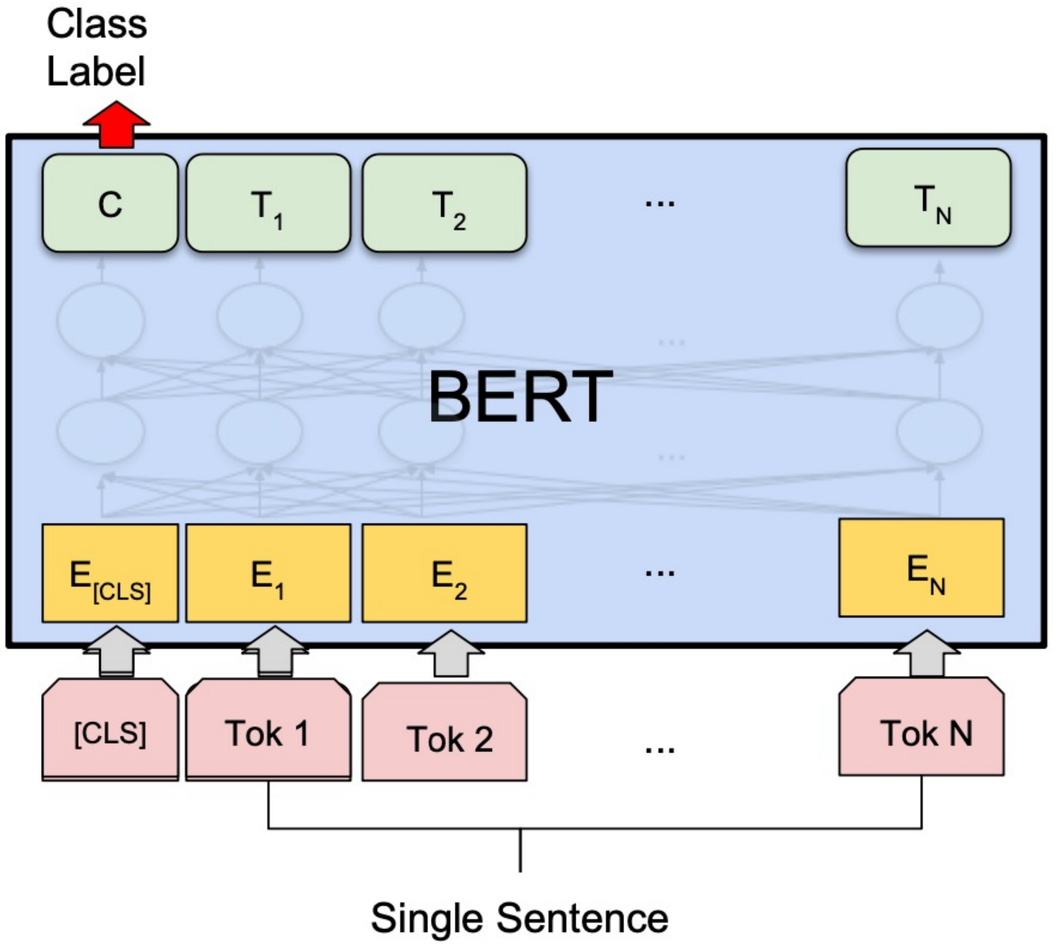


Typically, all weights are updated in fine-tuning (incl. LLM weights); in some case (partial) freezing of layers can be useful

(Devlin et al. 2018)



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

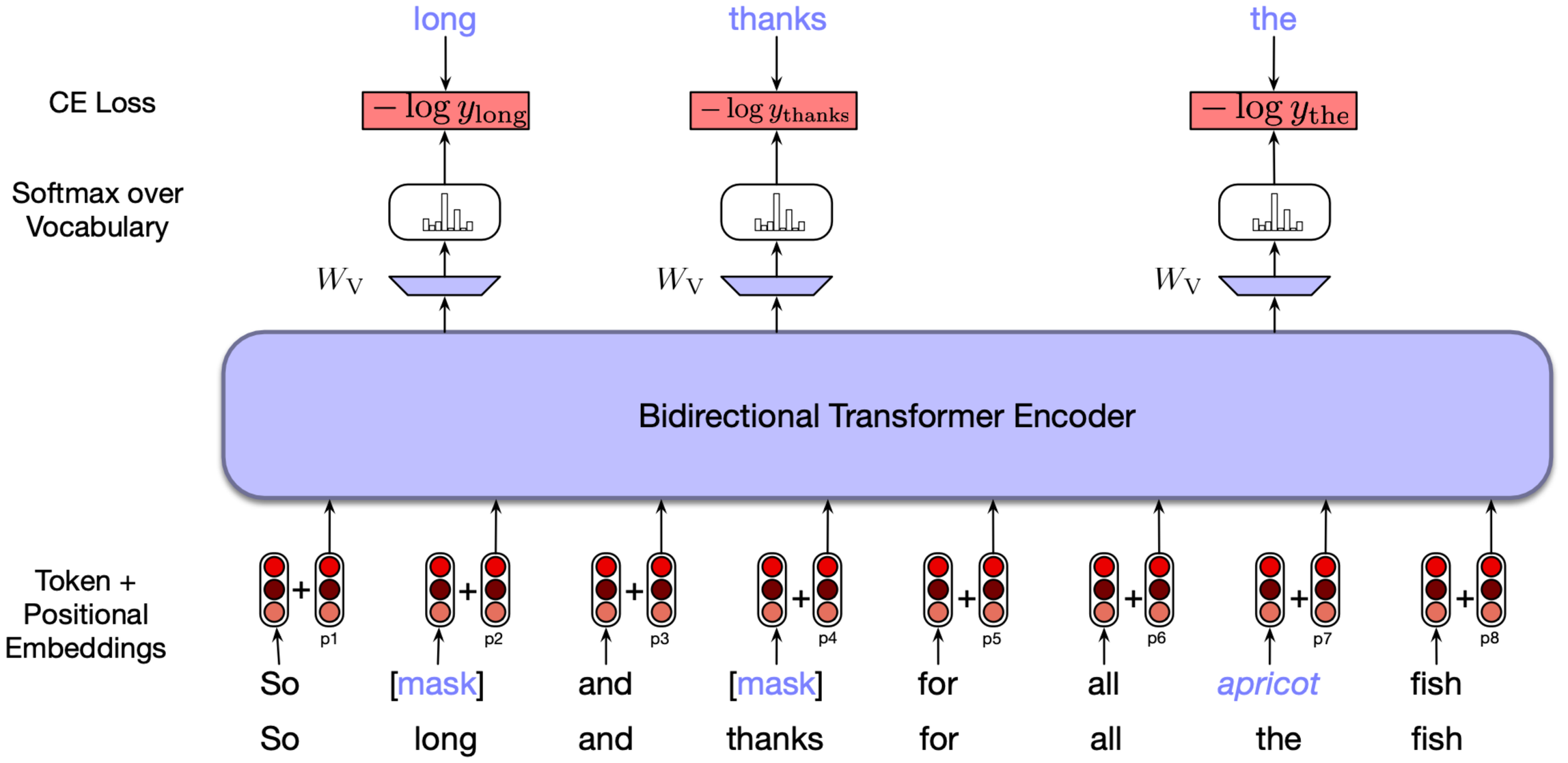


(b) Single Sentence Classification Tasks: SST-2, CoLA

(Devlin et al. 2018)

BERT: some details

- Bidirectional encoder representations from transformers
- First bi-directional transformer-based LM (2018, Google AI Language) → bi-directional architecture allows for parallelization
- Was the state-of-the-art solution for 11+ NLP problems
- Trained on >3 Billion words from Wikipedia and Google Books for Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)
- 12 transformer-block layers, each with 12 multi-head self-attention layers → 110 Million parameters
- Inputs processed through WordPiece Subword Tokenizer (30,000 types)
- Several further developments based on BERT available, e.g. RoBERTa, DistillBERT



(Jurafsky & Martin 2022)

Tasks 1 | Libraries | Datasets | Languages | Licenses | Other

Filter Tasks by name | Reset Tasks

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering

Models 26,234

Filter by name

new Full-text search

ProsusAI/finbert
Text Classification • Updated May 23 • 931k • 250

nlptown/bert-base-multilingual-uncased-sentiment
Text Classification • Updated Apr 18, 2022 • 3.46M • 133

SamLowe/roberta-base-go_emotions
Text Classification • Updated Sep 15, 2022 • 3.6M • 13

distilbert-base-uncased-finetuned-sst-2-english
Text Classification • Updated Mar 21 • 2.52M • 253

arpanghoshal/EmoRoBERTa
Text Classification • Updated Feb 11 • 212k • 63

siebert/sentiment-roberta-large-english
Text Classification • Updated Apr 2 • 200k • 64

JessyTsu1/comment_opinion_extract_ChatGLM_base
Text Classification • Updated 8 days ago • 25 • 3

OpenAssistant/reward-model-deberta-v3-large-v2
Text Classification • Updated Feb 1 • 31.1k • 79

Examples of political text classification tasks

- hate speech (e.g. Basile et al. 2019, Aluru et al. 2020, Mathew et al. 2022)
- offensive language (e.g. Wiegand et al. 2018, Zampieri et al. 2019 and 2020, Mandl et al. 2021)
- (pre-specified) conspiracy theories (e.g. Moffitt et al. 2021, Elroy & Yosipof 2022, Phillips et al. 2022)

The majority of respective datasets is in English language, and heavily focused on Twitter as data source (cf. Poletto et al. 2021) → the latter is already changing

Experimental results

TelCovACT dataset

- 4,000 randomly sampled messages from manually ranked public Telegram channels known for mobilization against state-measures related to pandemic
- Annotated with regard to content and stance
- Top-level content labels: conspiracy theory and antisemitism
- Comprehensive annotation guidebook with a lot of examples (25 pages)
- On a random subsample: two Annotator:innen per text example (Data Science and Critical Discourse Theory backgrounds)
- All annotators had solid knowledge regarding CT & A; not all annotators had knowledge regarding training of models

(Steffen et al. 2023a)

Example

Bitte teilen und handeln. !!!!! Für die Menschenwürde !!! Fast ein Jahrhundert überlebt?

Und jetzt in den Fängen der Todesengel im Corona Guantanamo der Pharmasatanisten von denen Joseph Mengele nur geträumt hätte?

Alten und Pflegeheime als Todeszelle, wo das Leiden, Verzweiflung und die Hilfeschreie aus dem Fegefeuer der noch lebendigen mit einem Fluss von Drogen und Beruhigungsmitteln verstummen sollen. Warten auch deine Eltern, die dich liebevoll großgezogen haben, darauf qualvoll und möglichst menschenunwürdig und allein und ohne Hoffnung und ohne warme Hand den letzten Atemzug machen zu dürfen um, möglichst schnell als Corona Toter für das diktatorische Regime dienen zu dürfen. Folgend ein Beispiel. URL

Conspiracy

- [x] Yes[1]
- [ ] No[2]
- [ ] Uncertain[3]

Stance

- [x] Belief[4]
- [ ] Authenticating[5]
- [ ] Directive[6]
- [ ] Rhetorical Question[7]
- [ ] Disbelief[8]
- [ ] Neutral/Uncertain[9]

Content

- [x] Actor[0]
- [x] Strategy[q]
- [x] Goal[w]
- [ ] Reference[e]

Antisemitism

- [x] Yes[t]
- [ ] No[a]
- [ ] Uncertain[s]

Stance

- [x] Affirmative[d]
- [ ] Critical[f]
- [ ] Neutral/Uncertain[g]

Content

- [ ] Encoded[z]
- [x] Post-Holocaust[x]
- [ ] Other[c]

Others

- [x] Pandemic Reference[v]
- [ ] GDR Reference[b]
- [ ] Memorize Task[y]
- [ ] Task unsuitable[i]

Model evaluation

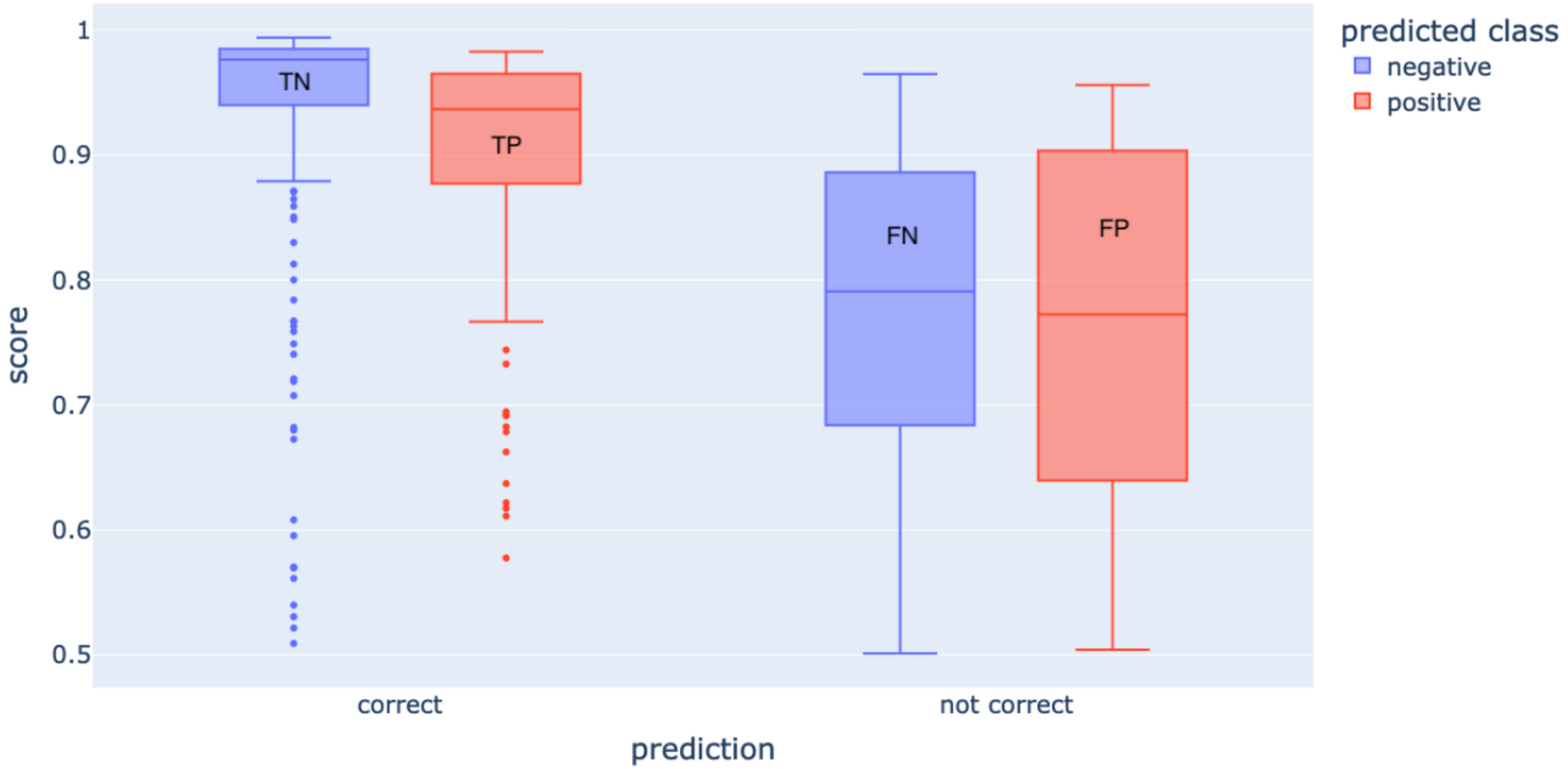
Dataset	Negative Class	Positive Class
Train / *Train down*	1873 / *886*	886
Evaluation	241	104
Test	230	115
Total	2344	1105

- Class 1: 30% of data
- Dataset small (<1,000 examples for class 1)
- Many model- and data-related hyperparameters → grid search too expensive
- Bayesian optimization competitive
- Model performs significantly better on class 0 than on class 1 → typical → **easy negatives**
- Performance on test data even slightly better (F1 for class 1: 0.75)
- F1 for class 1 improves when LLM **retrained** on in-domain data
- Self-Adjusting Dice Loss does not help

	Bayesian Model	Grid Model
downsampling	False	False
remove_emojis	False	False
remove_footer	**False**	**True**
hidden_droput	0	0
attention_probs_dropout_prob	0.3	0.3
epochs	2	2
batch_size	**32**	**16**
learning_rate	**5.5e**-5	**5e-5**
weight_decay	**0.2**	**0.1**
precision_0	0.87	0.88
recall_0	0.86	0.87
f1_0	0.87	0.88
precision_1	0.7	0.71
recall_1	0.71	0.72
f1_1	0.70	0.71
accuracy	0.82	0.83
eval_loss	0.47	0.49

(Pustet 2023)

Model evaluation



- The model struggles in particular with **fragmented** narratives
- True positives: frequent and explicit mention of power, control, and the desire of someone or something to take over the world; words related to plans, explicit mentions of actors
- Passive constructions like “was fired” difficult to detect as parts of CT

Decoding Antisemitism corpus

- Data collected from different plattforms through web scrapers. Corpus meant for qualitative and quantitative analyses, & for training of classification models
- Discourse-trigger driven corpus (discourses that will potentially trigger antisemitic speech such as World Cup or Kanye West)
- Hierarchical annotation scheme; top-level codes: antisemitic, contextually antisemitic, not antisemitic, counterspeech
- Additional codes for linguistic markers, antisemitic stereotypes, object of attack, hate speech, etc.
- All annotators are researchers focusing on antisemitism, working mainly in applied linguistics, communication studies, political science. No Data Science background

Model evaluation

	precision	recall	F1 score	# records	accuracy
class 1	**0.75** (0.73)	**0.65**	**0.7** (0.69)	225 (249)	0.94
class 0	0.96	0.97	0.97 (0.96)	2,084 (2,061)	

- Model trained on 80% (16,539 records in class 0 and 1,936 in class 1) of the data (10% validation, 10% test)
- Model performs a lot better on class 0 than on class 1 → an even more pronounced difference than for CT model
- Performance for class 1 decreases (F1: 0.6) when applied to a **new discourse**; performance on class 0 stable
- Performance for both classes decreases even further when applied to a **new corpus** (similar platform, keyword bias, similar but different annotation)

(Steffen et al. 2023b)

Data-based challenges

Kanye West now "likes Jewish people again" - thanks to Jonah Hill in '21 Jump Street'

Sortieren nach: **Beste** 1679 Kommentare

+ Kommentar hinzufügen

ProLicks • vor 3 Monaten

Man, if only Hitler had been able to see *Superbad*...

7097 Antworten Teilen ···

PMMEBITCOINPLZ • vor 3 Monaten

No one can listen to that monologue about drawing dicks and hate Jewish people.

1309 Antworten Teilen ···

63 weitere Antworten

56 weitere Antworten

Substantial-Pass-992 • vor 3 Monaten

So aside from the obvious is he also saying that he just now saw 21 Jump S

2246 Antworten Teilen ···

AtlasShrunked • vor 3 Monaten

Wait til he discovers *22* Jump Street, it'll blow his mind

955 Antworten Teilen ···

94 weitere Antworten

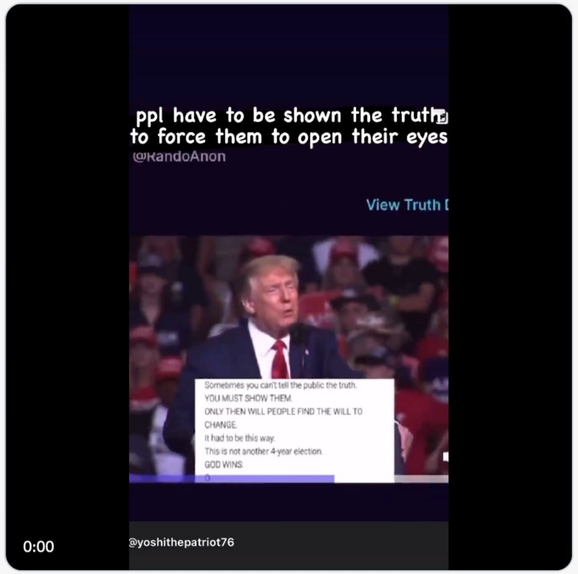
AvalancheReturns • vor 3 Monaten

Theres a lot to unpack here.

452 Antworten Teilen ···

JoseMario"D Great Pumpking"(Gab/Gettr ... @JoseMar39... · Jun 29

the reason why the deep State, elite, globalist, warmonger swamp creatures want to take him out; 45 expose their plans for the "**Great Reset**"



34

Tweet

Karlitaflor @Karlitaflor1 · Jul 1

Well that solidifies it; I love you.
😂🤣🥰😘

1 7 5,391

Kevin - WE THE PEOPLE❤️ - DAD🦁🐉🔥 - @bambkb · Jul 1

Love you back❤️❤️❤️❤️

3 4,967

TimeForTruth @TruthSearching_ · Jul 1

Very logical insights - agree and thanks.

1 10 3,736

Kevin - WE THE PEOPLE❤️ - DAD🦁🐉🔥 - @bambkb · 23h

Thank you😊❤️

2 3,274

The Truth Hurts @TheTruthHurtz82 · 5h

Dear Kllaus, WEF,Global Leaders and Elite's. I want to thank you all from the bottom of my heart. You all collectively have done something no one else could,YOU have brought the world together ❤️
You are no longer needed now,your DESTRUCTION is IMMINENT.Enjoy it while it lasts 🖕

1 2 15 215

Kevin - WE THE PEOPLE❤️ - DAD🦁🐉🔥 - @bambkb · 5h

❤️🔥🖕

1 172

2nGlenn @trynot2lookatit · 10h

Can't wait for NWO season period

1 1 1 464

Kevin - WE THE PEOPLE❤️ - DAD🦁🐉🔥 - @bambkb · 6h

NWO hunting season

Political texts from social/online media

- **Short**
- Increasingly **Multi-modal**
- **Contextual**: references to embedded images, videos, audios; URLs; previous comments; (political) events. The right context not easy to identify (e.g. when trying to resolve current ambiguities as e.g. in „I think you have been told to do this“ → does it make sense to attempt to classify single comments?
- **Codes**: antisemitism but also many forms of hate are often expressed using codes. E.g. “Dog whistles”: „convey one meaning to a broad audience and a second one, often hateful or provocative, to a narrow in-group“ (Mendelsohn et al. 2023). Not well detected by LLMs such as GPT-3 or toxicity scorers like Perspective API. Codes depend on platforms and communities

Annotation: fundamental but difficult

- Datasets annotated by a single research team too small; different research teams utilize different annotation schemes
- Crowd labeling through non-experts seems very limited: too simple definitions cannot be used to recognize fragmented and coded expressions of CT and A. (see Perspective API) → are there ways to join forces between related projects? How can we combine similar annotated datasets?
- What role should context play when annotating messages? Can humans annotate a thread, while ignoring previous messages? Knowledge available for humans needs to be well aligned with that for the model → close collaboration between modelling work and annotation work
- Shouldn't we work on multi-modal models considering the fact that image and video have become dominant?

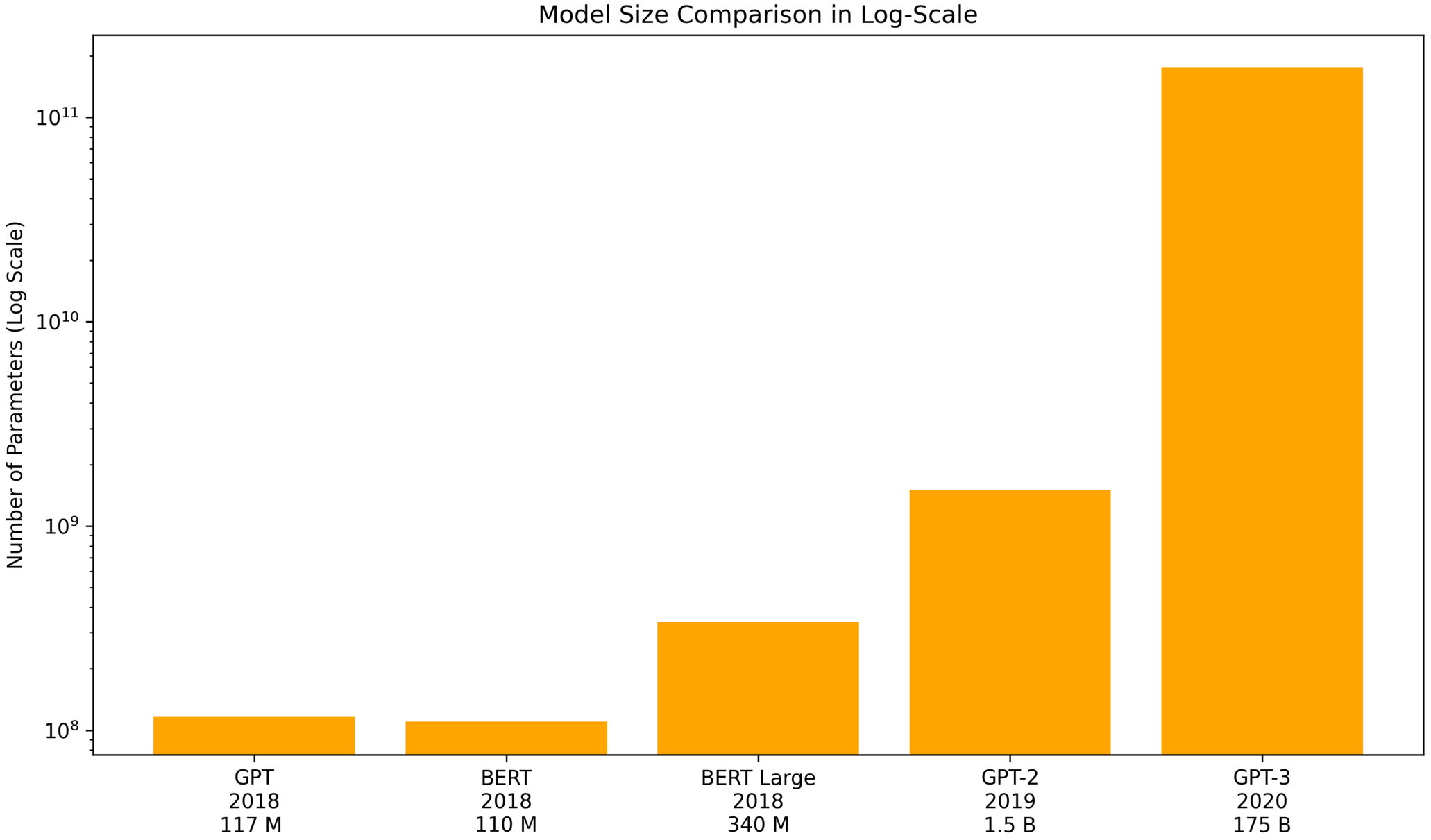
Is Few-Shot Learning the right way to go?

Few-Shot Learning

- Pre-training & fine-tuning effective (achieves SOTA results on many NLP tasks) but fine-tuning is expensive since we still need relatively large amounts of labeled data
- Depending on task complexity and diversity of data, one should prepare thousands or even tens of thousands labeled samples per task
- Few-Shot Learning: Model receives at inference time a few examples of the task together with a task description → training data. Model tries to solve the task but without updating model weights
- One-Shot Learning: only one example; Zero-Shot Learning: no example, only task description
- Zero-/One-/Few-Shot Learning are considered much closer to the way humans solve respective tasks

GPT-3

- GPT-2 can be seen as proof-of-concept for few-shot learning. GPT-3 has shown competitive performance in few-shot scenario, and for some applications even surpassed SOTA fine-tuning based models (Brown et al. 2020)
- GPT-3 is trained for Next Word Prediction → focus on Natural Language Generation and less on NLU/NLP as with BERT
- Transformer-based Architecture, but only context left from current token considered (masked self-attention instead of bidirectional self-attention)
- Instruction (task description plus training examples) are passed to the model as start context (this limits the size of examples to typically <100). Model generates word by word. Model weights are not updated.
- GPT-3 paper shows that large models have significantly better performance. GPT-3 is three orders of magnitude larger than BERT


Model Size Comparison in Log-Scale
Number of Parameters (Log Scale)
10^{11}
10^{10}
10^{9}
10^{8}
GPT
2018
117 M
BERT
2018
110 M
BERT Large
2018
340 M
GPT-2
2019
1.5 B
GPT-3
2020
175 B

GPT-models for political text classification?

- First empirical evaluations indicate huge potential for increasing the efficiency of text classification
- (Gilardi et al. 2023): zero-shot accuracy of ChatGPT exceeded that of crowd-workers in 4 out of 5 tasks related to content moderation, while being about twenty times cheaper
- (Chiu et al. 2022): few-shot learning with GPT-3 for the classification of sexist and racist texts yields solid results (accuracy of 85%)
- (Li et al. 2023): compare the performance of ChatGPT to that of crowd-workers for classification of texts regarding hateful, offensive, and toxic (HOT) content. ChatGPT achieves an accuracy of roughly 80% when compared to the annotations of crowd workers. Prompt design influences the performance of the model → systematic analyses of prompt design for GPT-3 based HOT classification needed

References

- Aluru, Sai Saketh/Mathew, Binny/Saha, Punyajoy/ Mukherjee, Animesh, 2020. Deep Learning Models for Multilingual Hate Speech Detection. Arxiv: arXiv: 2004.06465.
- Basile, Valerio/Bosco, Cristina/Fersini, Elisabetta/ Nozza, Debora/Patti, Viviana/Rangel Pardo, Francisco Manuel/Rosso, Paolo/Sanguinetti, Manuela, 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In: Pro- ceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 54–63. https://doi.org/10.18653/v1/S19-2007.
- Brown, Tom B. et al., 2020. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
- Chiu, Ke-Li/Collins, Annie/Alexander/Rohan, 2022. Detecting Hate Speech with GPT-3. Preprint. http://arxiv.org/abs/2103.12407.
- Elroy, Or/Yosipof, Abraham, 2022. Analysis of COVID-19 5G Conspiracy Theory Tweets Using SentenceBERT Embedding. In: Artificial Neural Networks and Machine Learning – ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part II. Berlin, Heidelberg: Springer-Verlag, 186–196. https://doi.org/10.1007/978-3-031-15931-2_16.
- Gilardi, Fabrizio/Alizadeh/Meysam/Kubli, Maël, 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Preprint. https://doi.org/10.48550/arXiv.2303.15056.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jikeli, Günther/Karali, Sameer/Miehling, Daniel/Soemer, Katharina, 2023. Antisemitic Messages? A Guide to High-Quality Annotation and a Labeled Dataset of Tweets. Preprint. arXiv: 2304.14599.
- Jurafsky, Dan / Martin, James H. Speech and Language Processing. 3rd edition (draft). (2022)
- Li, Lingyao/Fan, Lizhou/Atreja, Shubham/Hemphill, Libby, 2023. 'HOT' ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. Preprint. http://arxiv.org/abs/2304.10619.

References

- Mahl, D.; Schäfer, M. S.; and Zeng, J. 2022. Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research. New Media & Society, 1–21.
- Mandl, Thomas/Modha, Sandip/Shahi, Gautam Kishore/Madhu, Hiren/Satapara, Shrey/Majumder, Prasenjit/Schaefer, Johannes, et al., 2021. Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages. arxiv: 2112.09301.
- Mathew, Binny/Saha, Punyajoy/Yimam, Seid Muhie/ Biemann, Chris/Goyal, Pawan/Mukherjee, Animesh, 2022. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 17, 14867– 14875. arXiv: 2012.10289.
- Meta, 2022. Community Standards Enforcement | Transparency Center. https://transparency.fb.com/data/community-standards-enforcement (last accessed on 14 February 2023).
- Mendelsohn, Julia, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. "From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models." https://doi.org/10.48550/arXiv.2305.17174.
- Mihaljević, Helena/Steffen, Elisabeth, 2022. How Toxic Is Antisemitism? Potentials and Limitations of Automated Toxicity Scoring for Antisemitic Online Content. In: Proceedings of the 2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022), 1–12. Potsdam, Germany.
- Moffitt, J. D./King, Catherine/Carley, Kathleen M., 2021. Hunting Conspiracy Theories During the COVID- 19 Pandemic. Social Media + Society, Vol. 7, No. 3. https://doi.org/10.1177/20563051211043212.
- Nithum Thain, Lucas Dixon, and Ellery Wulczyn. 2017. Wikipedia Talk Labels: Toxicity. figshare. Dataset. https://doi.org/10.6084/m9.figshare.4563973.v2
- Phillips, Samantha C./Ng, Lynnette Hui Xian/Carley, Kathleen M., 2022. Hoaxes and Hidden Agendas: A Twitter Conspiracy Theory Dataset: Data Paper. In: Companion Proceedings of the Web Conference 2022. WWW '22. New York, NY, USA: Association for Computing Machinery, 876–880. https://doi.org/10.1145/3487553.3524665.
- Poletto, Fabio/Basile, Valerio/Sanguinetti, Manuela/ Bosco, Cristina/Patti, Viviana, 2021. Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review. In: Language Resources and Evaluation, Vol. 55, No. 2, 477–523. https://doi.org/10.1007/s10579-020-09502-8.
- Pustet, Milena. Transformer Based Detection of Conspiracy Narratives in the Context of COVID-19 on German Social Media. Bachelor's Thesis. Hochschule für Technik und Wirtschaft Berlin. 2023

References

- Steffen, Elisabeth/Mihaljević, Helena/Pustet, Milena/Bischoff, Nyco/Varela, María do Mar Castro/Bayramoğlu, Yener/Oghalai, Bahar, 2022. Codes, Patterns and Shapes of Contemporary Online Antisemitism and Conspiracy Narratives -- an Annotation Guide and Labeled German-Language Dataset in the Context of COVID-19. In: Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023). June 5–8, 2023, Limassol, Cyprus. AAAI Press, Palo Alto, California USA. https://doi.org/10.1609/icwsm.v17i1.22216

- Steffen, Elisabeth/Mihaljević, Helena/Pustet, Milena, 2023. Algorithms against antisemitism? Towards the automated detection of antisemitic content online. Preprint.

- Wiegand, Michael/Siegel, Melanie/Ruppenhofer, Josef, 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In: Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018). https://epub.oeaw.ac.at/0xc1aa5576_0x003a10d2.pdf.

- Zampieri, Marcos/Malmasi, Shervin/Nakov, Preslav/ Rosenthal, Sara/Farra, Noura/Kumar, Ritesh, 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 75–86. https://doi.org/10.18653/v1/S19-2010.

- Zampieri, Marcos/Nakov, Preslav/Rosenthal, Sara/ Atanasova, Pepa/Karadzhov, Georgi/Mubarak, Hamdy/Derczynski, Leon/Pitenis, Zeses/Çöltekin, Çağrı, 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020. https://doi.org/10.18653/v1/2020.semeval-1.188.