natural language processing for machine learning-assisted text annotation in the social sciences

Franziska Weeber (she/her)
textada | Stifterverband | HU Berlin
01.08.23

the next 40 minutes

- current trends
- two reasons for improved performances

- challenges for NLP
- potential solutions

- semi-automated workflow
- demo project

- discussion
- contact details

interest in NLP has skyrocketed

The New Chatbots Could Change the World. Can You Trust Them?

Siri, Google Search, online marketing and your child's homework will never be the same. Then there's the misinformation problem.

Google now understands more conversational search queries

The tech giant says it's one of the biggest Search updates in the product's history.

AI models like ChatGPT and GPT-4 are acing everything from the bar exam to AP Biology. Here's a list of difficult exams both AI versions have passed.

- https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1
- https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html
- https://www.engadget.com/2019-10-25-google-search-bert-update.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAEdPIAPZLVhHE8_85NkJOqID43bgq8U_cZDxLbwx29T_sJM9Y8ggSt2Lx1Gkkk_D1TqdIbbLa-XSdrpkvmiptQ8qEdJxSymC3nMEBT4KTkgUFx8nqiN0xWZLbQ5qa8cMpXZEC0kjbHeiOAqUUE2sXtPkGnmNiBbgvlYzIqK35YPV

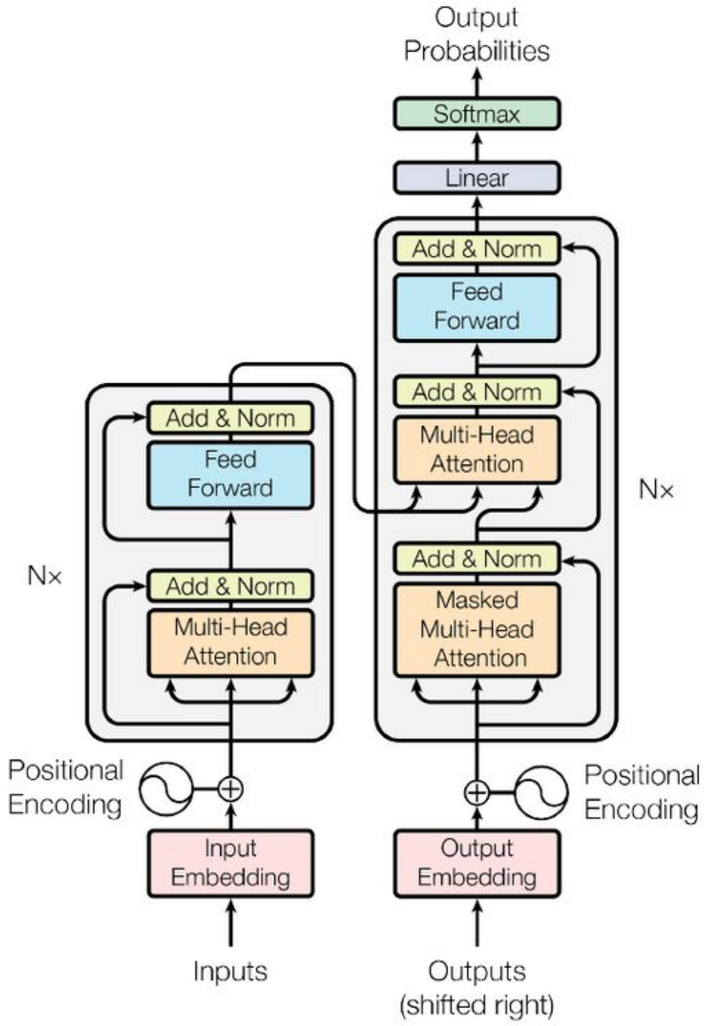
NLP benchmark performances

taken from: Kiela et al. 2019 [2]

improved architecture: transformer

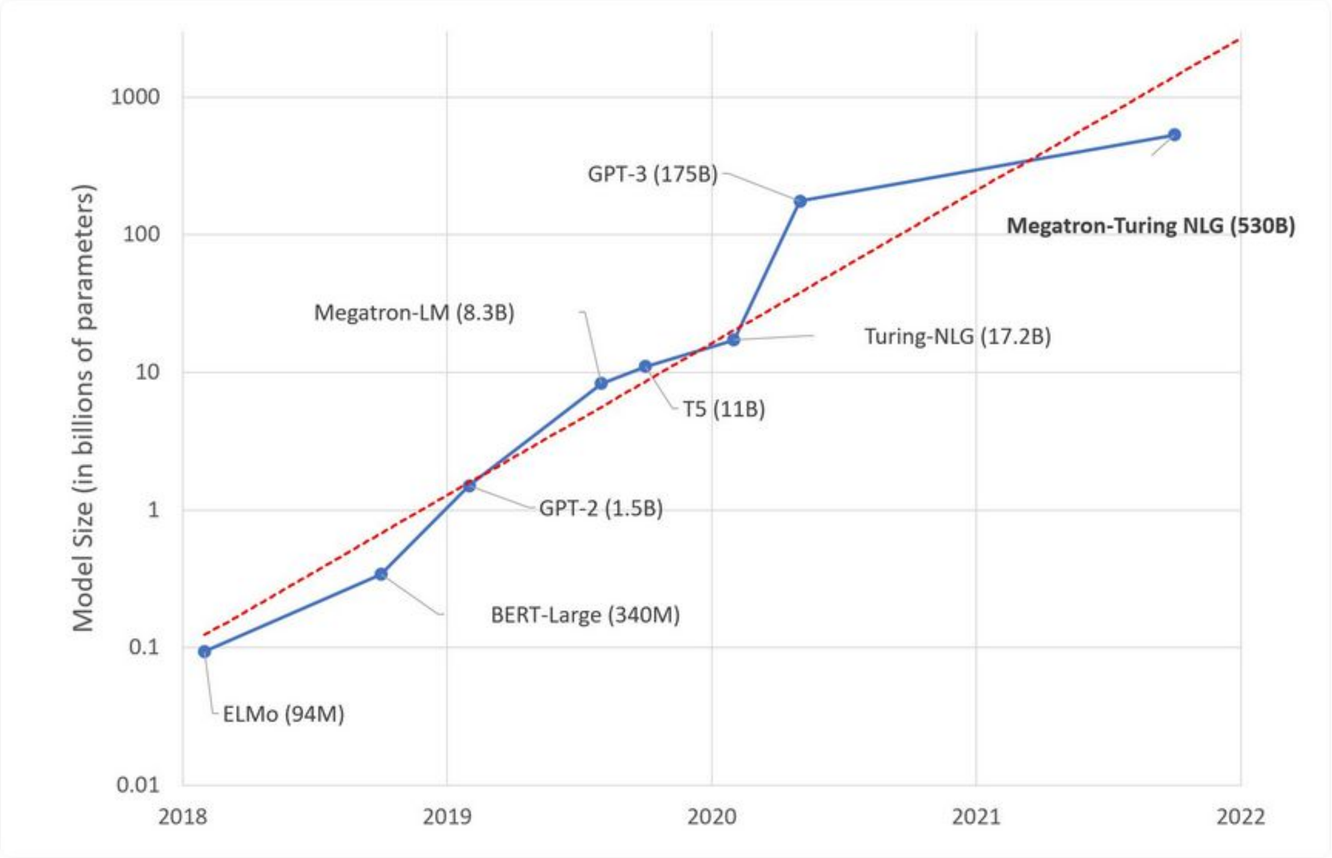
Vaswani et al. 2017 [7]

- process an entire sequence at once
- context-specific embeddings
- attention mechanism to focus on important words
- used in most top-performing NLP models



improved architecture: size

- deep neural networks with (m|b)illions of parameters
- trend towards larger models



- https://huggingface.co/blog/large-language-models

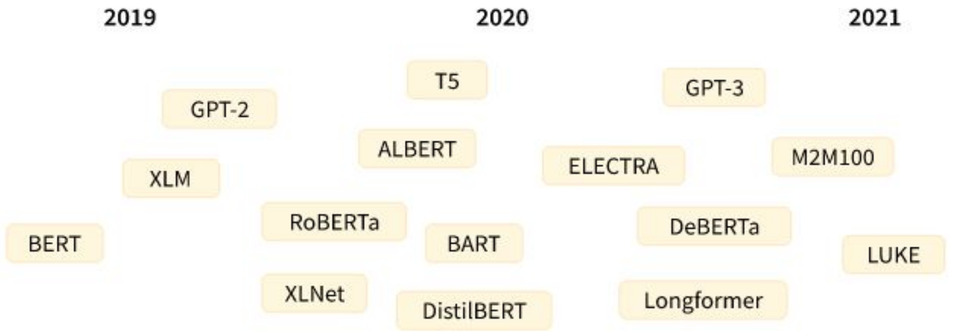
pretrained language models

pretraining → publication → finetuning

- extensive resource-intense pretraining on large corpora
- save weight checkpoints
- make them available
- adapt to specific domain and task-specific finetuning

- July 31, 2023: pretrained model on HuggingFace
  - ~17.000 for text generation
  - ~30,000 for text classification

- https://huggingface.co/models?pipeline_tag=text-generation&sort=trending



- https://huggingface.co/learn/nlp-course/chapter1/4

what do we get from NLP?

improved architectures

- deep neural networks
- attention and transformers

+

pretrained language models

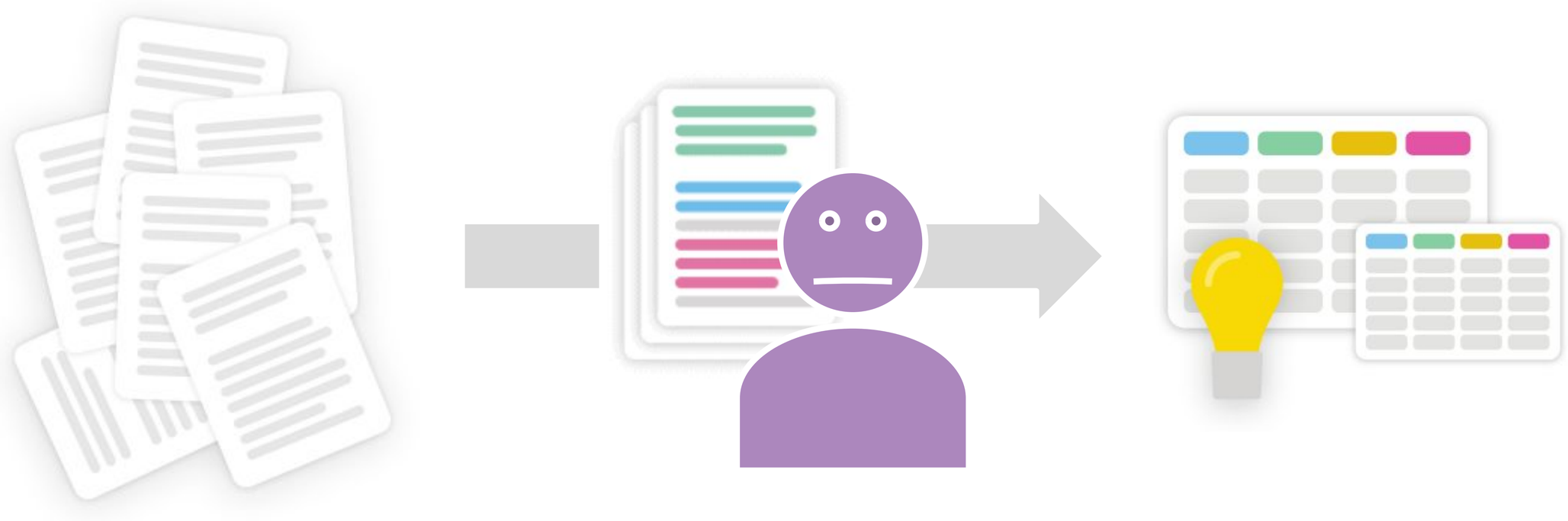
- weight checkpoints
- finetuning on specific tasks

→

- **semantically meaningful and context sensitive embeddings**
- **blueprints to use embeddings for different tasks**

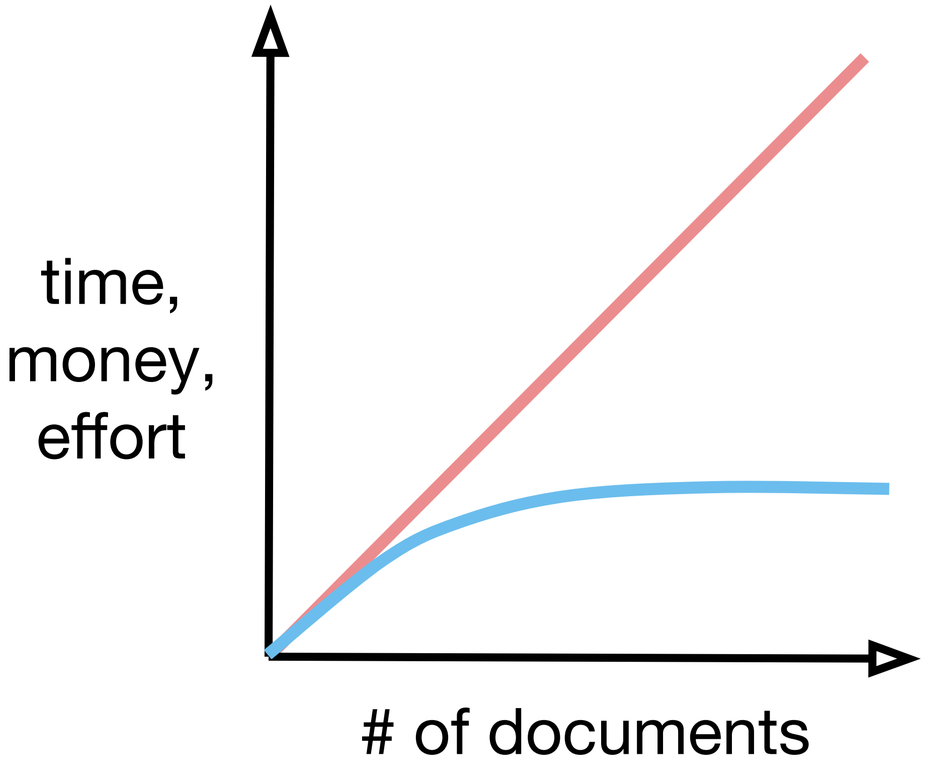
Example use case: text annotation

annotation extracts info from text



effective, but inefficient

- human linguistic & expert knowledge
  - irony
  - metaphors
  - context
  - …
- individualized projects
  - text type
  - categories
  - domain
  - …

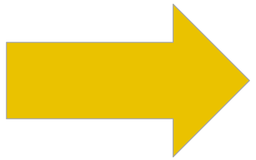


what is our task?

- **sequence tagging**
  - input documents too long (BERT: 512 token limit)
  - typically on one or very few tokens (NER, POS tagging)
- **text classification**
  - convert to fixed units (sentences/paragraphs)

where to get training data?

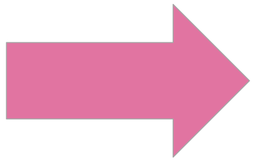
- wide range of domains
- complex category schemes
  - training data required to finetune model and to train classification head
  - purely automated solutions not feasible



get some manual annotations and apply low-resource techniques

when does a model perform well?

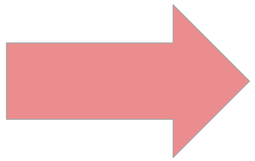
- traditional performance calculation requires large test sets



  - define gold standard documents for testing
  - random splits
  - qualitative evaluation from manual inspection
  - alternative metrics (e.g. prediction uncertainty [5])

what if it performs badly?

- annotated data is used to gain insights for research etc.
  - high performance requirements

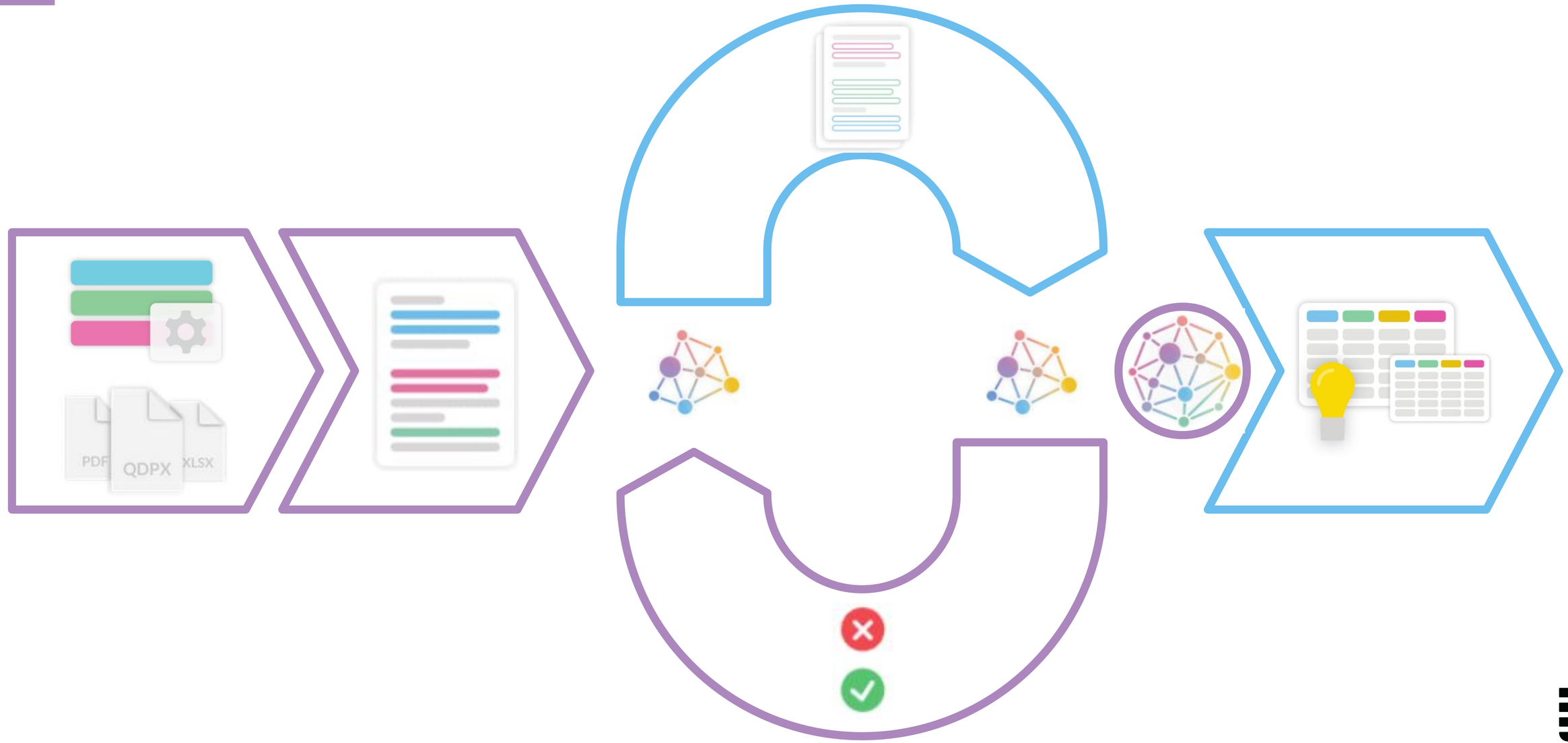


improve training set and then retrain model

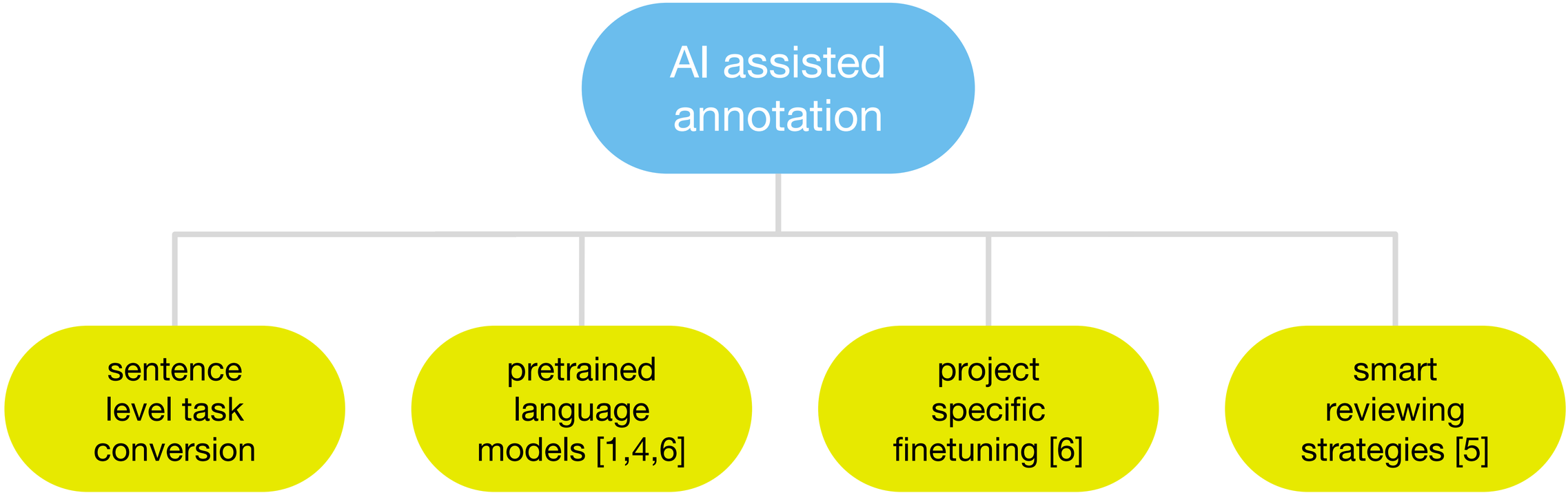
- review suggested annotations
- add more annotations

textada: a no-code webtool with a machine learning annotation assistant

working with textada's AI assistant



the underlying ML framework



Demo time!

identifying six types of framing bias in news articles

Dallas Card et al. 2015

“The Media Frames Corpus: Annotations of Frames Across Issues.”

In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China: Association for Computational Linguistics, 438–44.

some takeaways

try textada ;)

NLP has great power but needs more interdisciplinary research

make your research methodology available to others

let's get in touch!

info@textada.org

textada.com

@textada

@textada__

@textada__

happy to connect to talk about annotation projects, research ideas or anything else!

Franzi

franziska@textada.org

References

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Sep. 22, arXiv, 2019. doi: arXiv:1810.04805
- [2] D. Kiela et al., Dynabench: Rethinking Benchmarking in NLP. 2021.
- [3] P. Mayring, “Qualitative Inhaltsanalyse,” in Handbuch qualitative Forschung: Grundlagen, Konzepte, Methoden und Anwendungen, U. Flick, E. von Kardorff, H. Keupp, L. von Rosenstiel, and S. Wolff, Eds., Munich: Beltz, 1991, pp. 209–213.
- [4] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” presented at the EMNLP, 2019.
- [5] B. Settles, “Active Learning Literature Survey,” University of Wisconsin-Madison Department of Computer Sciences, Computer Sciences Technical Report, 2009.
- [6] L. Tunstall et al., “Efficient Few-Shot Learning Without Prompts.” arXiv, 2022. doi: arXiv.2209.11055.
- [7] A. Vaswani et al., “Attention Is All You Need.” presented at the NIPS, 2017.