Enhancing Text Classification for Accurate Statistics: Leveraging LLM at the Ghana Statistics Service

Laurent Smeets and Josephine Baako-Amponsah

Ghana Statistical Service (GSS)

Why?

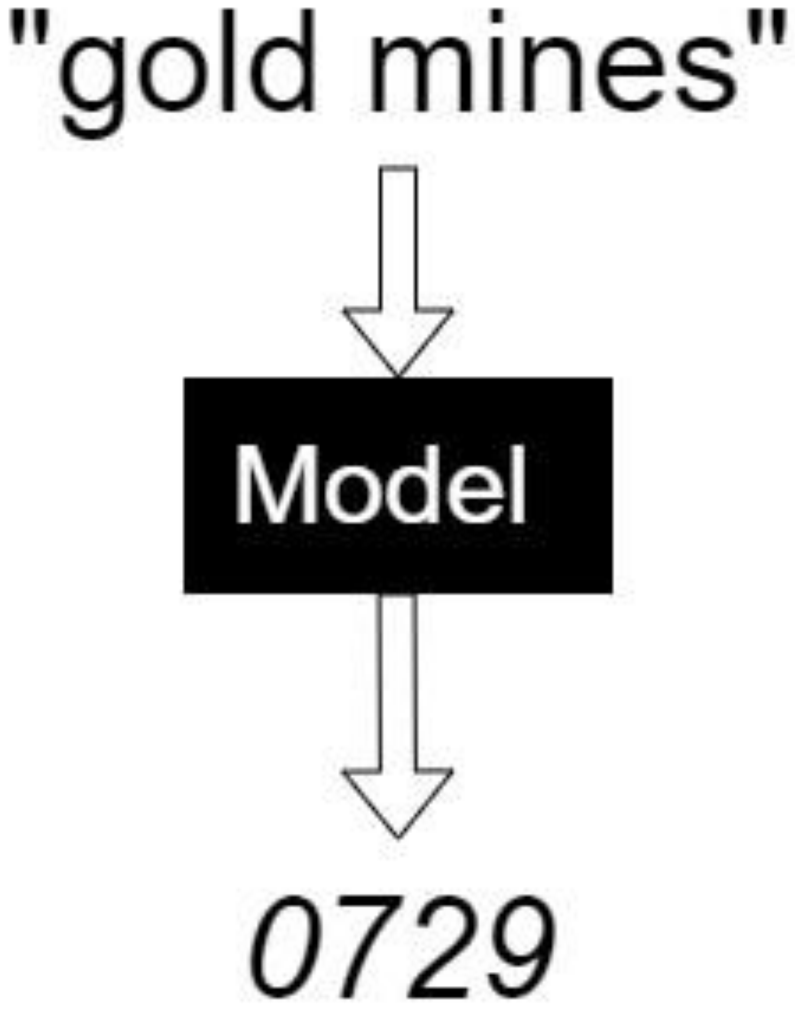
- Aim is to empower the GSS with a tool that extracts of relevant information from open-ended questions.
- To make automate tedious and error prone work and to speed up the validation of field work by minimizing the need for manual coding.
- Get validated data public quicker.
- build cost-efficient system that is easy to maintain.

When?

- ISIC during IBES
- ISCO and ISIC during GLSS
- HS code for cross border trade
- To reclassify “other, specify” questions into answer categories
- The first level of application of this project stage involves comparing the outputs of the trained model against codes assigned by field officers during data collection.

What do we want?

- A model that can automatically predict the ISIC code from open text
- Should be robust to typos
- This tool should use “English” as used by field enumerators and not United Nations English
- https://isic-prediction-27a3d67ecbf2.herokuapp.com/



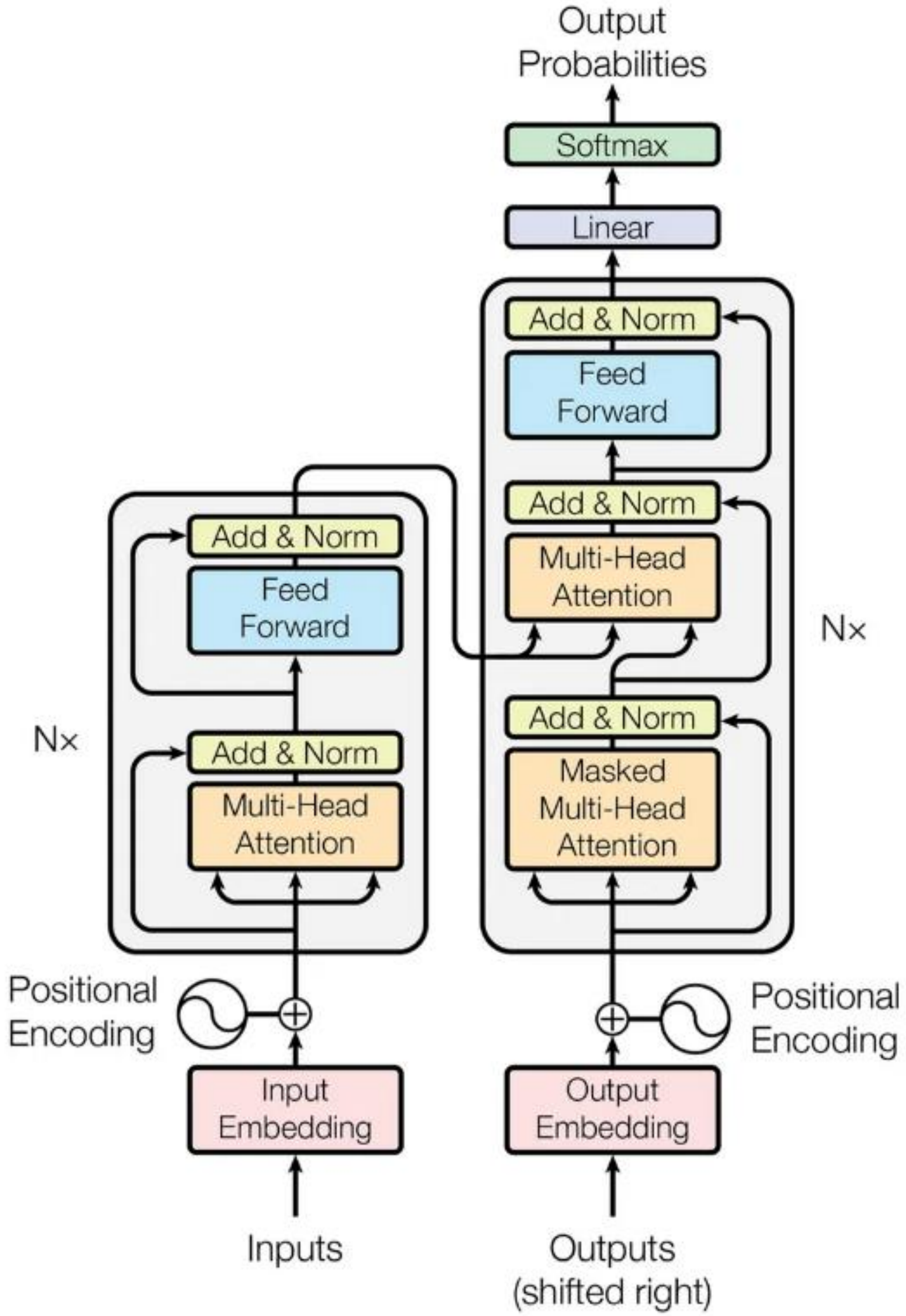
"gold mines"

Model

0729

What is in de box?

- Transformer model (2017)
- BERT (2018)
- DistillBERT (2019)
- Fine-tuned for ISIC prediction (2023)



What is in de box? *Transformer model*

- A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.
- longer-distanced context around a word (as compared to recurrent neural networks)

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

What is in de box? *(Distill)BERT model*

- Bidirectional Encoder Representations from Transformers
- Transfer learning describes an approach where a model is first pre-trained on large unlabeled text corpora using self-supervised learning:
- Pretrained on BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia
- This makes it possible to create a model that is context aware:

"Server, can I have the check"

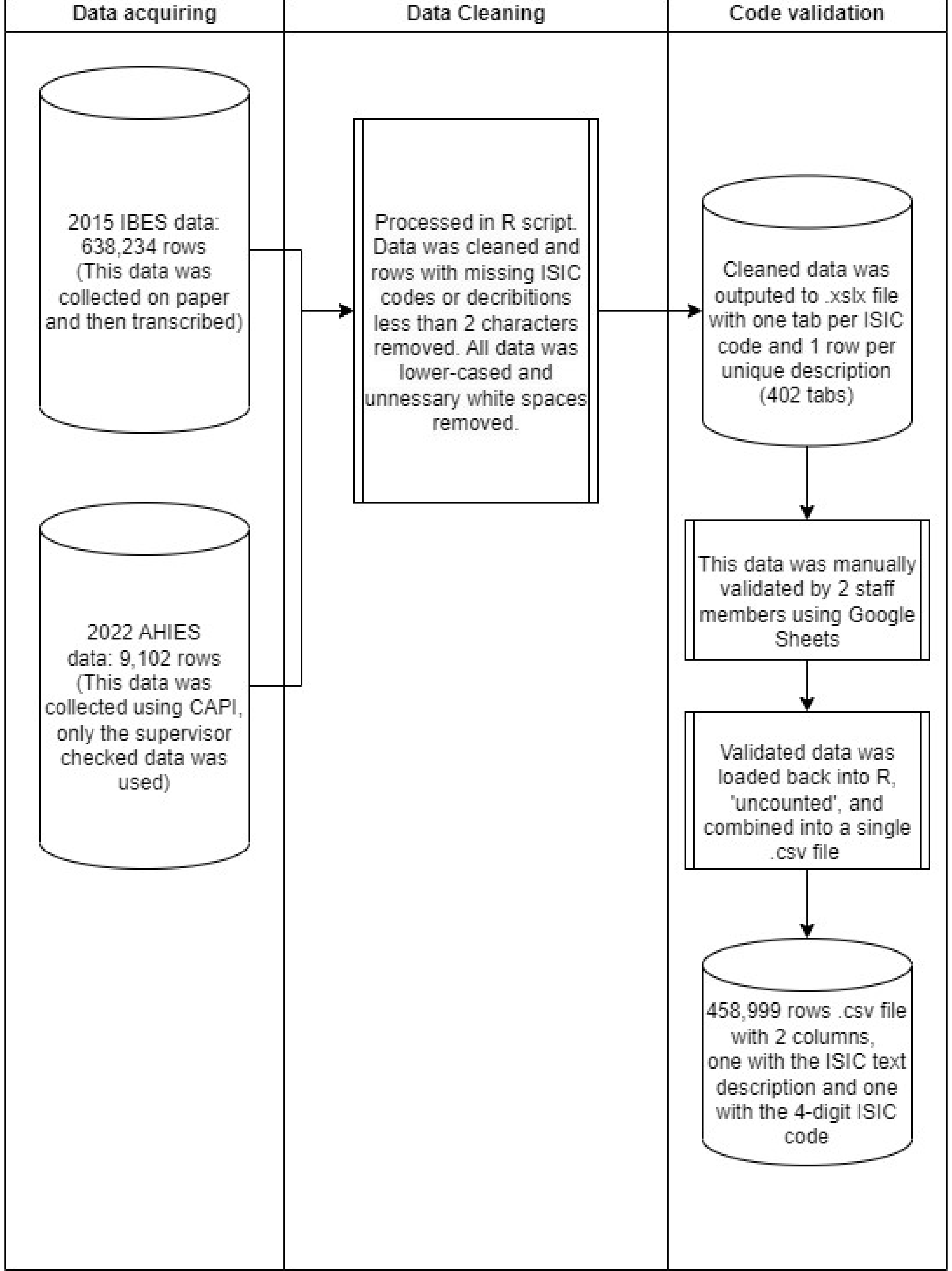
vs.

"Looks like I just crashed the server"

- Created for 'next' word prediction

What is in de box?
Fine-tuning

- Annotated (cleaned) data from previous surveys added as a final Fine-tuned classification layer
- This way the model can learn ‘Enumerator’ English
- Model was trained on Nvidia GPU cards on Runpod.io (couple of USD) in Python
- Results are hosted on Hugging Face



Data acquiring

2015 IBES data: 638,234 rows (This data was collected on paper and then transcribed)

2022 AHIES data: 9,102 rows (This data was collected using CAPI, only the supervisor checked data was used)

Data Cleaning

Processed in R script. Data was cleaned and rows with missing ISIC codes or decribitions less than 2 characters removed. All data was lower-cased and unnessary white spaces removed.

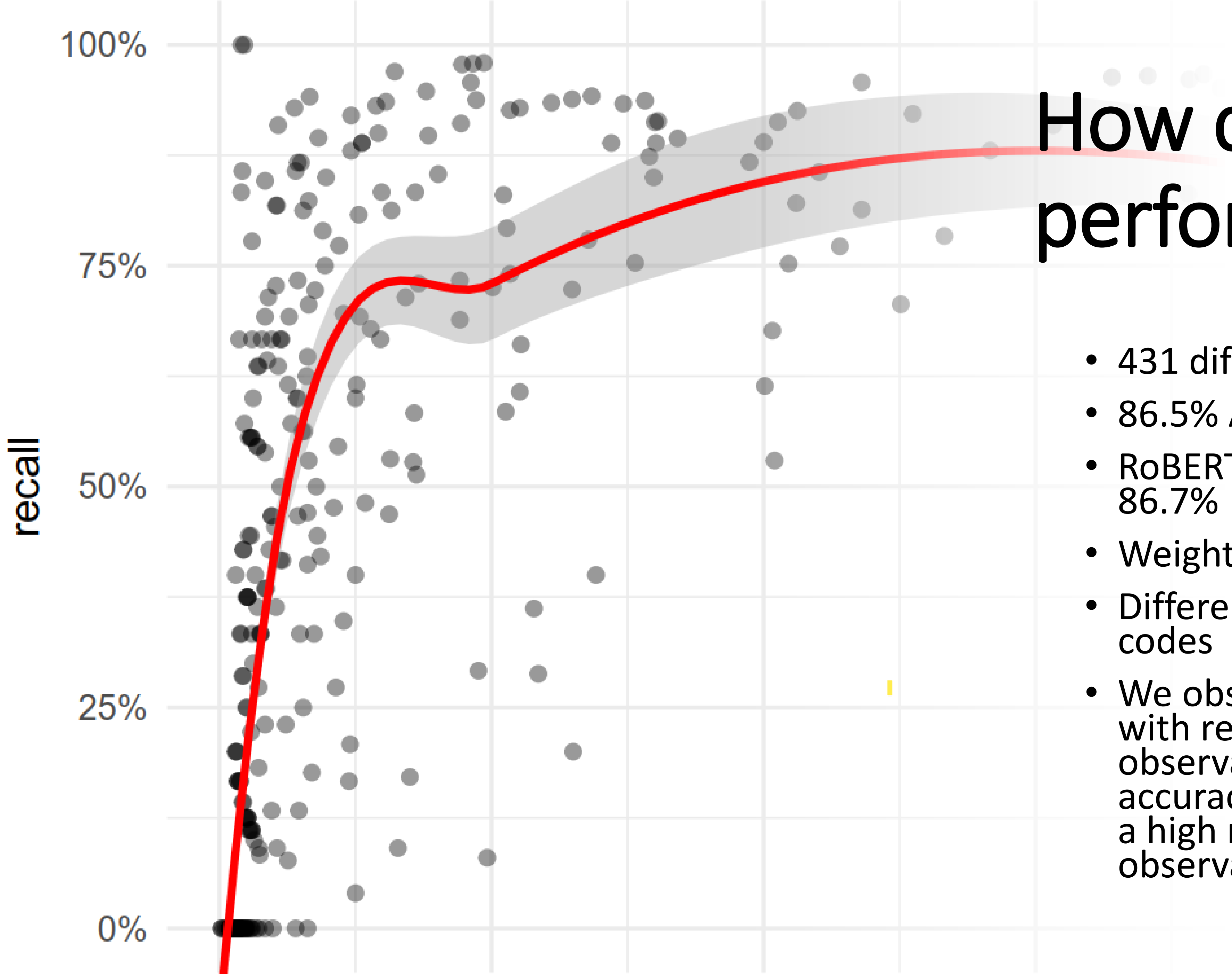
Code validation

Cleaned data was outputed to .xslx file with one tab per ISIC code and 1 row per unique description (402 tabs)

This data was manually validated by 2 staff members using Google Sheets

Validated data was loaded back into R, 'uncounted', and combined into a single .csv file

458,999 rows .csv file with 2 columns, one with the ISIC text description and one with the 4-digit ISIC code



How did the model perform?

- 431 different classes
- 86.5% Accuracy (80/20 split)
- RoBERTta: had an accuracy of 86.7%
- Weighted F1: 85.5%
- Differences between ISIC codes
- We observe that categories with relevant few observations have a lower accuracy than categories with a high number of observations.

What (didn’t work as well)

- Random forest: Model of 23GB for 83.8% accuracy
- Support Vector Machines: O(N^3)
- Creating code using gpt-3.5-turbo: Accuracy 59.1%.

```
″come up with 10 correct descriptions for establishments with
ISIC code {code} is: {description} one example would be
’{example}’ each example can be maximally 100 characters long.
try to make the example relevant for the types of businesses
you might find in Ghana, without mentioning the word Ghana.
dont mention the ISIC code in the output″
```

What?
(is next)

- Getting the code to run on Android (using Tensorflow Lite)
- At the different stages of the project, the tool is shared with thematic experts for their feedback
- Write a paper with our results
- Put it into production

Thank You

Useful links

- http://jalammar.github.io/illustrated-transformer/
- https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/
- https://mchromiak.github.io/articles/2017/Sep/12/Transformer-Attention-is-all-you-need/#.XIWlzBNKjOR
- https://www.youtube.com/watch?v=SZorAJ4I-sA&ab_channel=GoogleCloudTech
- https://www.exxactcorp.com/blog/Deep-Learning/how-do-bert-transformers-work
- https://arxiv.org/abs/1910.01108
- https://towardsdatascience.com/how-to-use-transformer-based-nlp-models-a42adbc292e5
- https://towardsdatascience.com/illustrated-guide-to-transformers-step-by-step-explanation-f74876522bc0
- https://arxiv.org/pdf/1810.04805.pdf