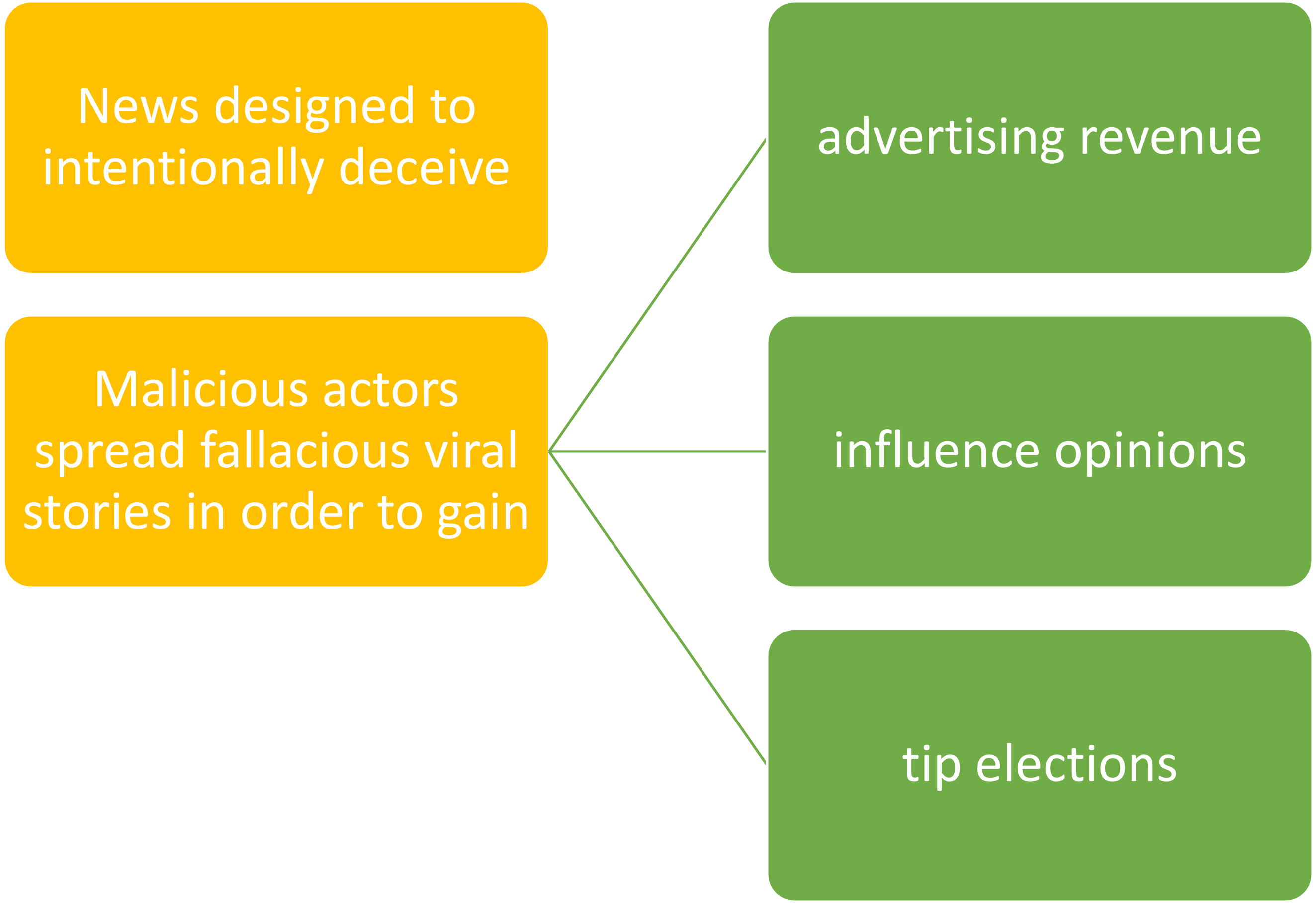
Algorithmic Defense mechanisms against Online Fake News

Presented by- Choukha Ram, Data Scientist

Fake News ?



News designed to intentionally deceive

Malicious actors spread fallacious viral stories in order to gain

advertising revenue

influence opinions

tip elections

Neural Fake News is here

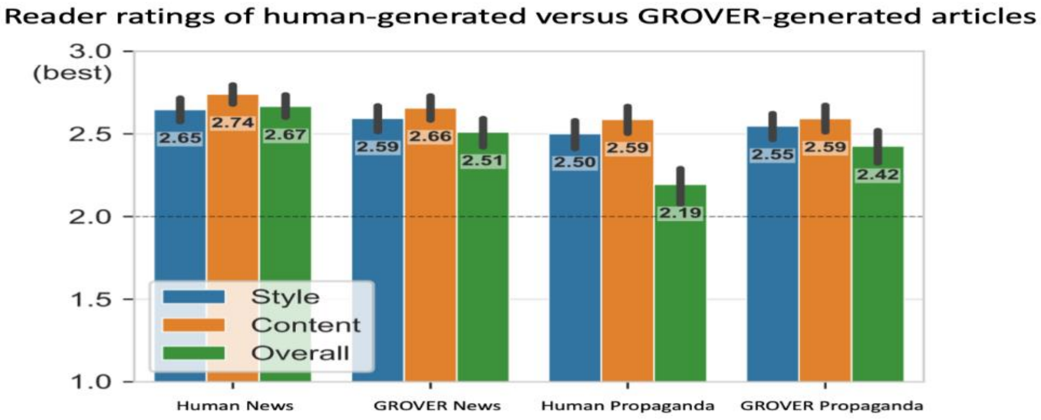
- https://grover.allenai.org/
- **GROVER** framework
- Generating aRticles by Only Viewing mEtadata Records.

Modeling Conditional Generation of Neural Fake News

- An article can be modeled by the joint distribution:

  p (domain, date, authors, headline, body)

- Trained model on a large corpus of news articles from Common Crawl. (120GB) (Training time – 2 week)
- Humans are Easily Fooled by Grover-written Propaganda.*



Automated Text Generation

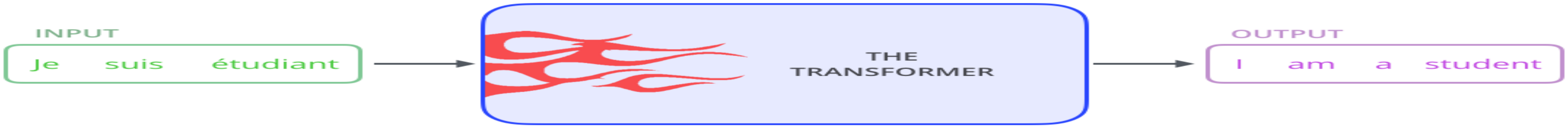
- Applied to marketing and robotics, and used to create chatbots, product reviews and write poetry.
- The ability to synthesize text, however, presents many potential risks, while access to the technology required to build generative models is becoming increasingly easy.
- State of the art language models – GPT-2, BERT & many more.

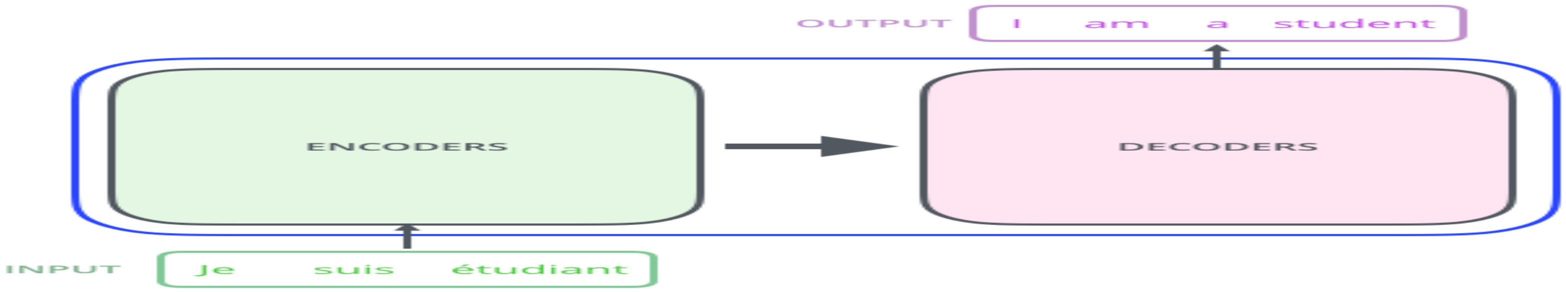
OpenAI’s GPT-2

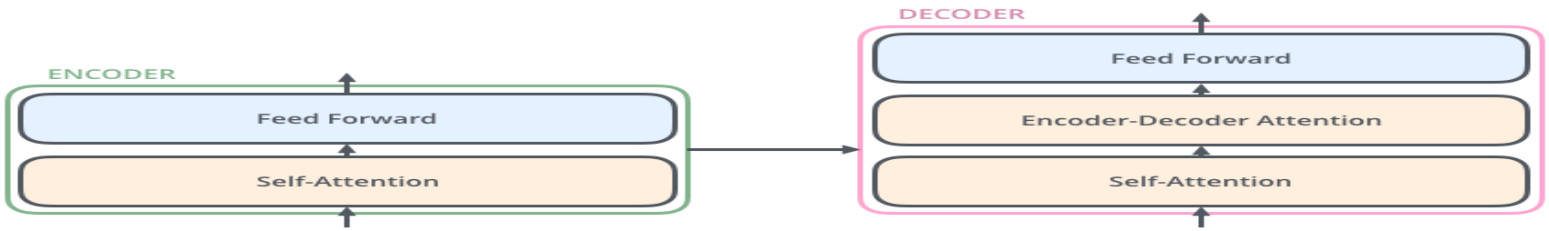
- They’ve trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.
- GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages.(40GB) GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text.
- https://openai.com/blog/better-language-models/

Transformer * (Technical Content Alert)

Transformer

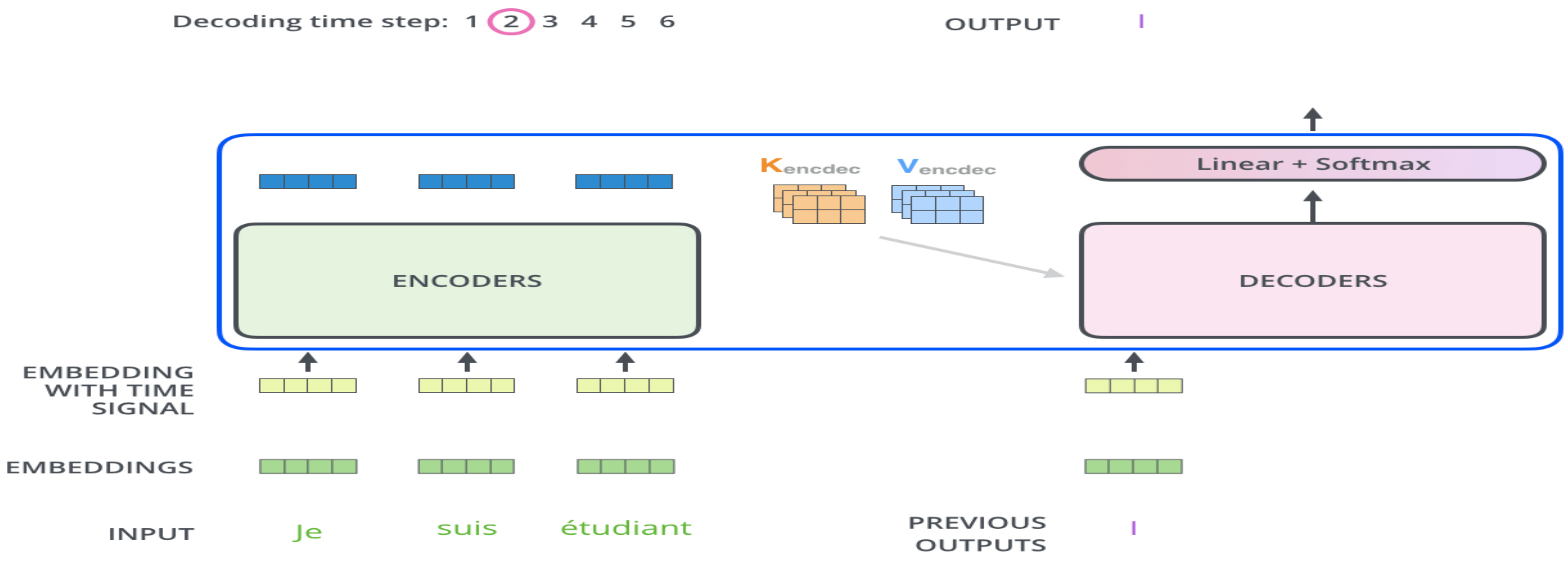






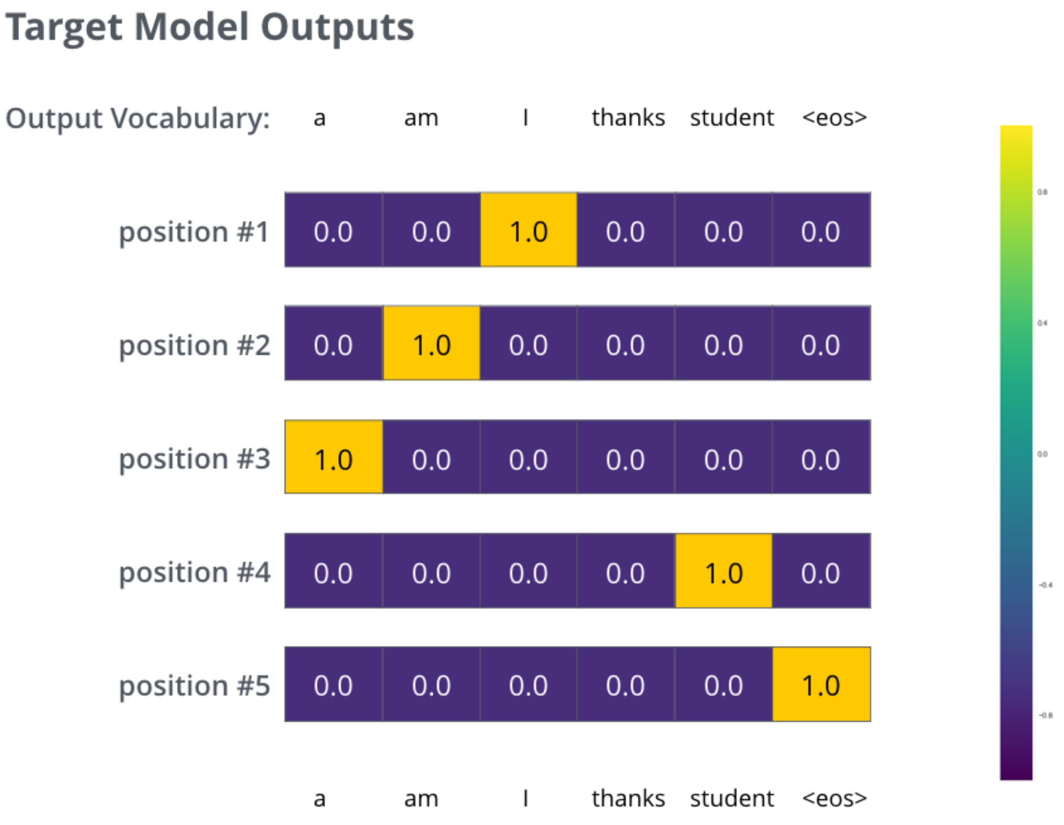
http://jalammar.github.io/illustrated-transformer/

Transformer



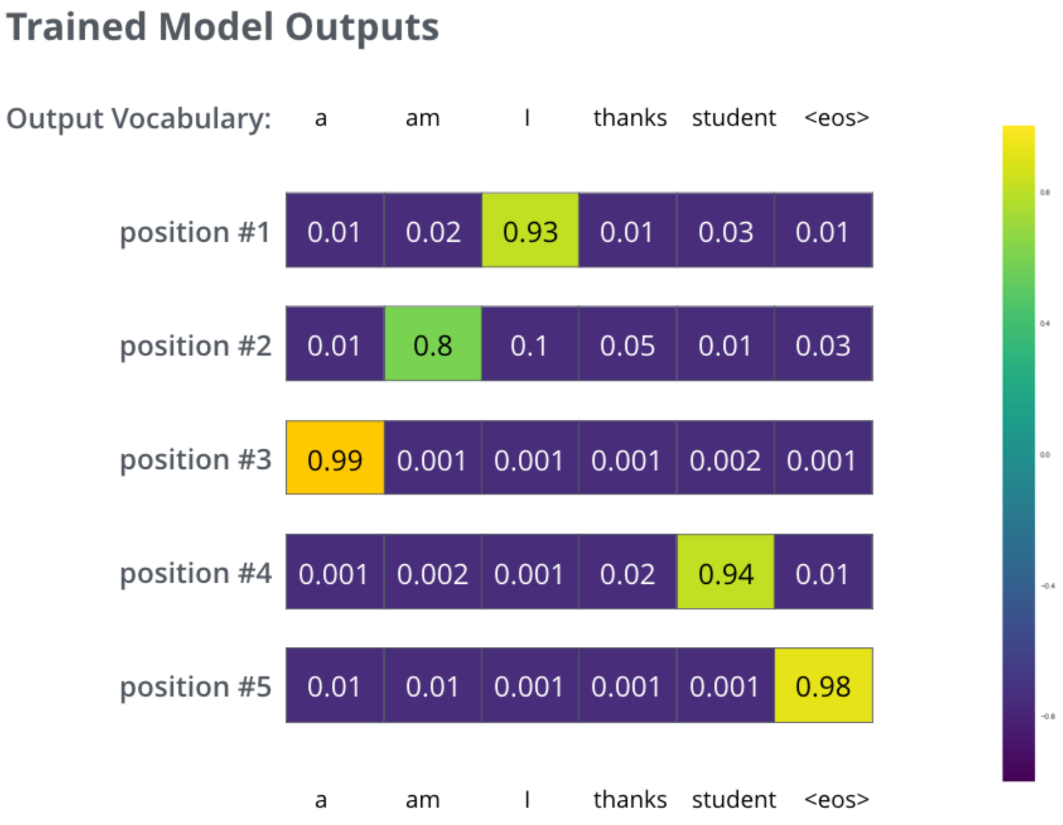
*The Transformer architecture is superior to RNN-based models in computational efficiency. RNNs are slow and their ability to learn long-term dependencies is still limited due to vanishing gradients.

Transformers output



Target Model Outputs

Output Vocabulary:	a	am	I	thanks	student	<eos>
position #1	0.0	0.0	1.0	0.0	0.0	0.0
position #2	0.0	1.0	0.0	0.0	0.0	0.0
position #3	1.0	0.0	0.0	0.0	0.0	0.0
position #4	0.0	0.0	0.0	0.0	1.0	0.0
position #5	0.0	0.0	0.0	0.0	0.0	1.0
	a	am	I	thanks	student	<eos>



Trained Model Outputs

Output Vocabulary:	a	am	I	thanks	student	<eos>
position #1	0.01	0.02	0.93	0.01	0.03	0.01
position #2	0.01	0.8	0.1	0.05	0.01	0.03
position #3	0.99	0.001	0.001	0.001	0.002	0.001
position #4	0.001	0.002	0.001	0.02	0.94	0.01
position #5	0.01	0.01	0.001	0.001	0.001	0.98
	a	am	I	thanks	student	<eos>

GPT2 vs BERT

BERT (Bidirectional Encoder Representations from Transformers)

- Uses only Encoder Transformer & is Bidirectional.

GPT2

- Uses only Decoder Transformer & allows Zero-shot learning.*

*The pre-training task for GPT-2 is solely language modeling. All the downstream language tasks are framed as predicting conditional probabilities and there is no task-specific fine-tuning.

Availability

- As easy as «git clone» --> «pip install» or «docker build»
- https://github.com/openai/gpt-2/blob/master/DEVELOPERS.md

- **Demo** - https://gpt2.apps.allenai.org/
- https://github.com/allenai/lm-explorer/blob/master/lm_explorer/lm/gpt2.py

Implications

Positives

- AI writing assistants
- More capable dialogue agents
- Unsupervised translation between languages
- Better speech recognition systems

Negatives*

- Generate misleading news articles
- Impersonate others online
- Automate the production of abusive or faked content to post on social media
- Automate the production of spam/phishing content

*These findings, combined with earlier results on synthetic imagery, audio, and video, imply that technologies are reducing the cost of generating fake content and waging disinformation campaigns.
To target the shared online commons, using things like "robotic tools, fake accounts and dedicated teams to troll individuals with hateful commentary or smears that make them afraid to speak, or difficult to be heard or believed".

Fake news Generation & Detection as an adversarial game

Adversary - Their goal is to generate fake stories that match specified attributes: generally, being viral or persuasive. The stories must read realistically to both human users as well as the verifier.

Verifier - Their goal is to classify news stories as real or fake. The verifier has access to unlimited real news stories, but few fake news stories from a specific adversary.

This setup matches the existing landscape: when a platform blocks an account or website, their disinformative stories provide training for the verifier; but it is difficult to collect fake news from newly-created accounts.

The dual objectives of these two players suggest an escalating **"arms race" between attackers and defenders.** As verification systems get better, so too will adversaries. We must therefore be prepared to deal with ever- stronger adversarial attacks.

Generator & Discriminator

The high quality of neural fake news written by Grover, as judged by humans, makes automatic neural fake news detection an important research area.

Using **discriminator** models for the role of the **Verifier** can mitigate the harm of neural fake news by classifying articles as Human or Machine written. These decisions can assist content moderators and end users in identifying likely (neural) disinformation.

Evaluation

- In the **unpaired** setting, a discriminator is provided single news articles, and must classify each independently as Human or Machine.
- In the **paired** setting, a model is given two news articles with the same metadata, one real and one machine-generated.
- The **discriminator** must assign the machine-written article a higher Machine probability than the human-written article. We evaluate both modes in terms of accuracy.
- Discrimination results: **Grover performs best at detecting Grover's fake news.**

Fact checking and verification

- Major platforms such as Facebook prioritize trustworthy sources and shut down accounts linked to disinformation.
- Some users of these platforms avoid fake news with tools such as NewsGuard and Hoaxy and websites like Snopes and PolitiFact. These services rely on manual fact-checking efforts: verifying the accuracy of claims, articles, and entire websites.

Suggested Counter-measures

- **Release of generators is critical:**
  - At first, it would seem like keeping models like Grover private would make us safer. However, Grover serves as an **effective detector** of neural fake news, even when the generator is much larger.
  - If generators are kept private, then there will be little recourse against adversarial attacks.
- **What should platforms do ?**
  - Video-sharing platforms like YouTube use deep neural networks to scan videos while they are uploaded, to filter out content like pornography. The same need to be done for news articles.
  - An ensemble of deep generative models, such as Grover, can analyze the content of text – together with more shallow models that predict human-written disinformation.
  - However, humans must still be in the loop due to dangers of flagging real news as machine-generated, and possible unwanted social biases of these models.