Fake News Detection using ML

- Nirav R. Madhani

1. Abstract

In this paper, I've explored, from various sources about Fake News and its Detection using Machine Learning. I will begin the paper by discussing the meaning of fake News, then proceed with how to generate fake NEWS using ML models, then discuss how to detect fake News using ML with statistical analysis as well as training models. Here I'll explain in detail about 2 most successful models. Finally I'll explain some limitations of current methods and then conclude the paper.

2. Introduction

"Fake news is any false disinformation, manipulates facts and virally spreads (or maybe for a target audience). It is also accessible on daily news media or social media sites such as Facebook, Twitter and WhatsApp.."

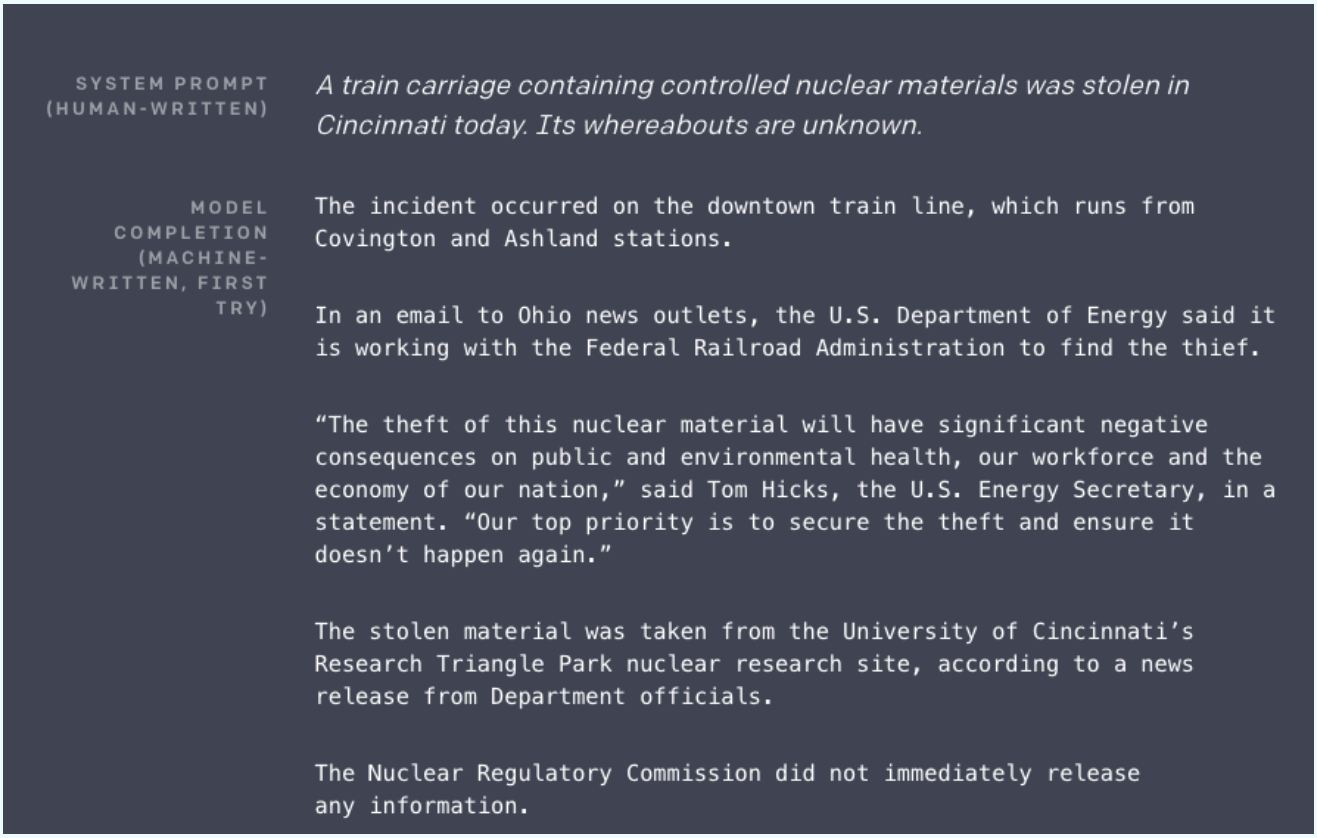
The fact that Fake News is carefully mimicking the 'style' and the 'pattern' that real news usually follows is the distinction between simple rumors such as "Moon landing was fake," and so forth. This is why an untrained human eye finds it so difficult to distinguish.

Interestingly, Fake News has been around for some time now in fact, in our entire history).

3. What is Neural Fake News?

It is any fake news developed using a neural network model. Formally speaking

"A focused propaganda that closely imitates the true style that a neural network produces"



OpenAI's GPT-2 model Fake News Example:
The GPT-2 model can be downloaded and executed freely. Note that the model is persuasive to broaden the prompt to a consistent part of the plot. That's what the research community is concerned about.

4. How Fake News is Generated? Case study of GPT-2.

GPT-2: It is an English net use statistical model. It can be used to create fake news to fool the public or discourage search engines such as Google. The performance of the production so stunned OpenAI that the whole model was too risky. How a computer like that can work and how a threat itself can be seriously discussed. GPT-2 can be consistent in many paragraphs of text. Two main aspects boost GPT-2 compared to the generation of previous text models. It is not just plausible to merge the words but to create a text in which the pronouns are equally mutually appropriate.
For two reasons it is better
First, it's far larger than many other. With an immense amount of computing power in the GPU and a large data set (1,5 billion parameters) the GPT-2 is absolutely massive. The second stable feature of GPT-2 is to provide a whole new way to model text, much better than everything on top of the transformer. We have to begin by analyzing how things function so that we can understand how transformer models work. RNNs offered us a chance to use and model languages with varying sentence lengths in every way we have learned about designing deep neural network image grading systems. Both automatic language translation systems and speech recognition systems were significantly improved in just one or two years. A well-trained RNN can make the text at least one or two sentences look pretty true.

An attention mechanism was the most important step forward. Each new word you see updates your memory of more recent previous terms, thus on long pieces of text RNNs are not working well. However, by whole paragraphs of the document the text cannot yet be generated with consistent ideas. The emphasis could weigh the

predictive value in the sentence of each earlier word. That will demonstrate the clear connection between the missing word and each word in the expression. And if we use millions and thousands of web-scrapped words, the model would understand how each English word in any possible way relates to any other word.

As with layers in a deep neural network we can place transformer modules on one top of the other: 48 of these transformer layers are stacked on top of each other in a complete GPT-2 model!
Indeed, GPT-2 is just short for "Generative Making a next word susceptible of the metadata fields in a text piece based is a great way to monitor what the model creates." The model also wrote fake dates in the British style instead of USA style (6 June).

5. How to Identify Fake News generated using Neural Network?

A. Manual Fact-Checking

We may easily google it, visit reputable news websites and review the details whether they have the same story or the like. Although this move sounds like common sense, one of the most reliable ways of making sure that a news story is real. However this move tackles only one kind of false news: one that is not viral or from a single source. What if we want to deal with the news that is viral and media-filled around us? This is generally the news that is created by a neural network, as the news so convincingly looks like "structure"and "form" real news.

B. Using Machine Learning

I. Statistical Analysis with help of HarvardNLP model (GLTR)

The Giant-Language-Test-Room or GLTR is a system developed by the great people of Harvard NLP and the MIT-IBM Watson lab.

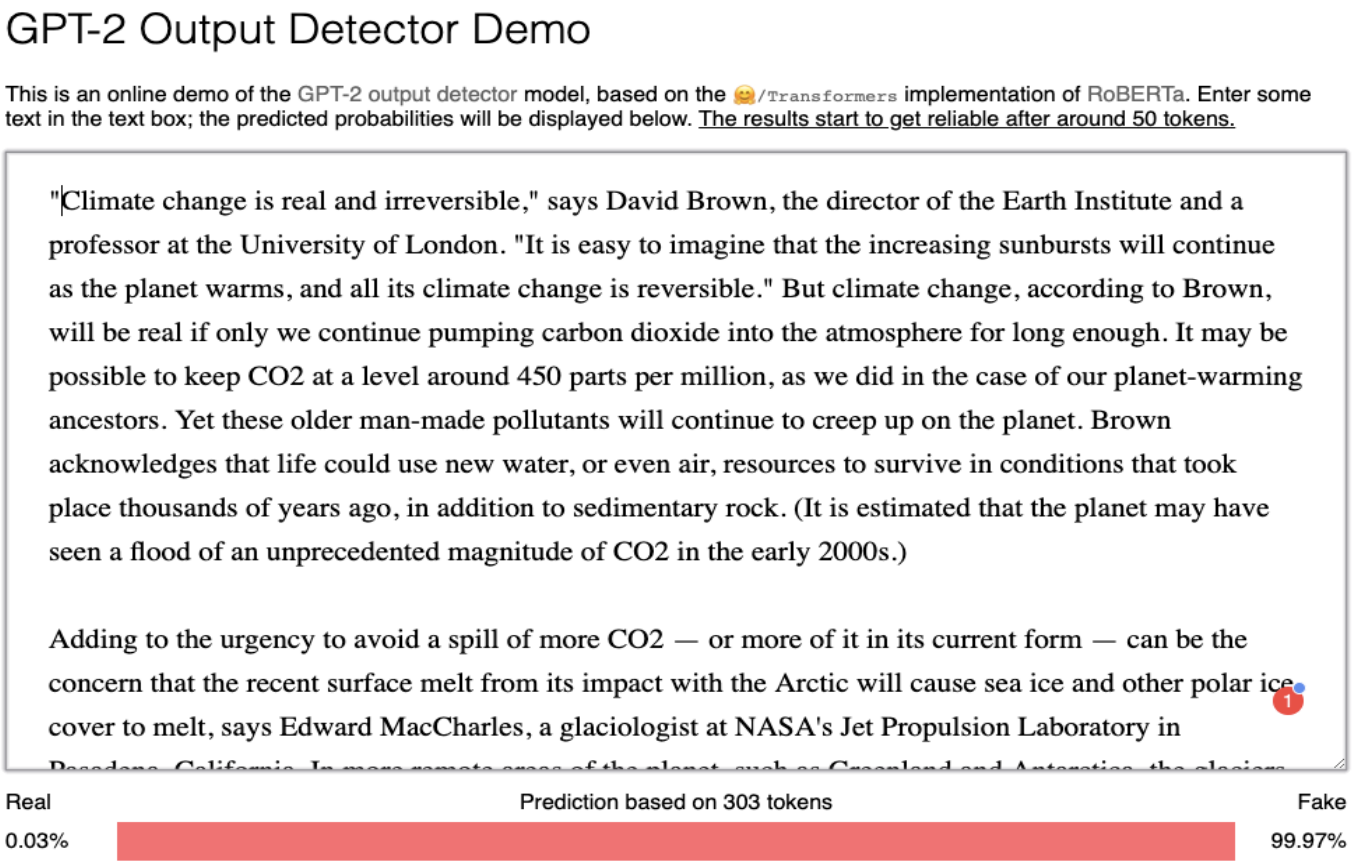
A smart combination of statistical and visual analysis of a specific piece of text for the classification of the master generated text is the main approach GLTR uses.

GLTR primarily uses the or similar model used first to produce the very same piece of text in the detection of generated text.
This is simply because the words generated by a language model are extracted from the distribution of the likelihood that the training data itself have learnt. We are aware of techniques used to test words with a certain probability, such as max sample, k-max sample search, beam search, nucleus sampling, etc. And if the text contains many such terms, then essentially this will confirm that the computer was being used to create such text.

II . Fake News Detection using a Model

- **GPT-2 Detector Model (RoBERTa)**

  RoBERTa (BERT variant) is a model detector GPT-2, that has been fine tuned to predict whether or not GPT-2 (as an easy classification problem) has been created for one particular piece of text.



RoBERTa is a broad language model that Facebook AI Research has built to improve over Google's BERT. For this reason the two frameworks are very close.

An significant point to be noted is that while the model architecture of RoBERTa is very different from the GPT-2 architecture, the former as a masked linguistic model is (such as the BERT) not of generative type, unlike the GPT-2.

Another positive thing about this model is that it is unbelievably easy to forecast compared to other techniques.

"Although the text seems very persuasive and coherent, the model classified it automatically as 'False' with an accuracy of 99.97 percent"

- **Grover (AllenNLP)**

Grover is known as an opponent game with two player models to detect neural fake news. This is the significance of this:

1. There are two models in the setup to detect text generated.
2. The goal of the adversary model is to produce false news reports, which can be either viral or persuasive for both humans and verifier models.
3. The verifier classifies whether the text is true or false.

The verifier's training data consists of infinite true reports, but only of certain false information from a certain opponent.

This is done in order to simulate the real world situation in which there are comparatively less false news than true news.

The two models have a dual purpose, which results in the attackers and defenders being 'competitive' to create and identify false news simultaneously. The adversarial model is improving with the verifier's model.

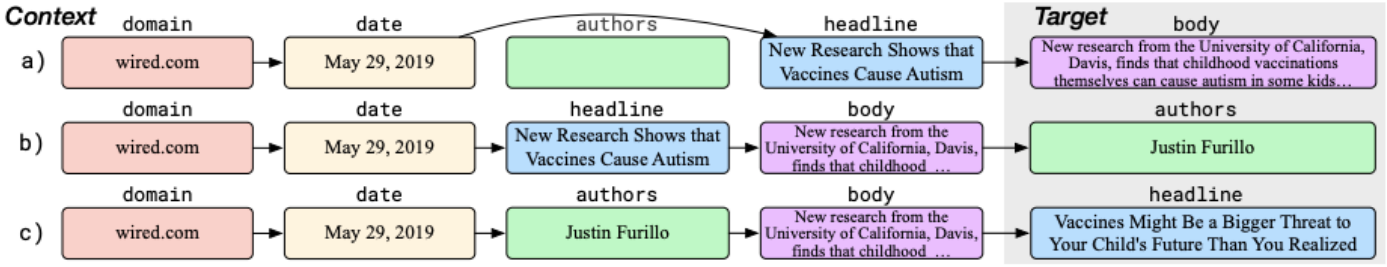
Through the joint probability distribution, we can model an article by combining all these parameters:

$$p(\text{domain}, \text{date}, \text{authors}, \text{headline}, \text{body}).$$

Fake News from Grover

Now that this is outside the reach of the paper, the underlying mathematics of how it is carried out won't go too far. But here is a graphic example to give you an insight into the whole generation process:

Here is what is taking place:



“In row a), partial context gets generated from the body.

In b), the model produces authors

And in c), the model uses subsequent generations to make the headline given more plausible.”

Dataset and Architecture

Grover uses the same GPT2 architecture:

Three model sizes exist. There are 12 layers and 124 million parameters in the smallest model, Grover-Base.

There are 24 layers with 355 million parameters in the next model, Grover-Large, on par with BERT-Large

And Large Grover, the biggest model, is amazingly 48-layered with1.5 billion parameters, on par with GPT2.

Grover's authors themselves created the dataset of RealNews which is used to train the Grover model. The data set is available and you can create it so you can download and use it or you can create your own dataset according to the requirements of Grover. The dataset is open source.

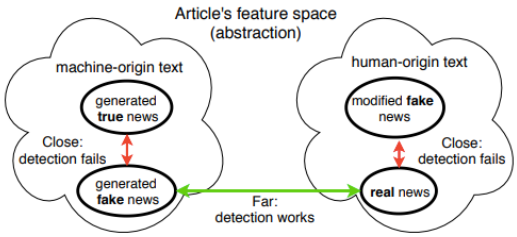
6. Limitations of Current Techniques

It is evident that current detection techniques are not ideal and have space for growth. MIT CSAIL has recently published a study of the latest methods for fake neural news identification and some of their results are eye opener.
This is because only machine-generated tools such as auto-completion, a text description etc can be used to evaluate whether a text is or is not "machine-generated" appropriate.

The popular Grammarly program, for example, uses a variant of GPT-2 to correct text grammatical errors.

The human written text, which is classified by the current methods as not neural false news, can however be slightly distorted / changed by attackers.

This is an example summarizing the identified issues of the detector model:



Since the region of neural fake news and actual news produced is very distant, a model can easily be classified as extremely easily. In the above figure.

Moreover, when the model has to classify between true generated news and false neural news, the functional region of both cannot be specified.

The same behavior would happen if the model separates real written news from those news that have been slightly modified and are now fake.

7. Conclusion

Present models employ different methods like transformers and GAN. The main focus is to verify whether the text is manually or computationally generated with the statistical parameters of existing data. However when a slightly corrupted written text is obtained, the model fails. One thing to do is to train models on factual data in line with the University Paper in Cambridge. Cambridge University and Amazon released FEVER, the Fac Extraction and VERification dataset.

Sources

1. https://towardsdatascience.com/i-trained-fake-news-detection-ai-with-95-accuracy-and-almost-went-crazy-d10589aa57c

2. https://www.analyticsvidhya.com/blog/2019/12/detect-fight-neural-fake-news-nlp/

3. https://paperswithcode.com/task/fake-news-detection/latest

4. https://medium.com/@ageitgey/deepfaking-the-news-with-nlp-and-transformer-models-5e057ebd697d

Appendix:

1. Transformers 2.0 Library
2. RoBERT code on GitHub.
3. Grover
4. RealNews
5. Installation Instructions
6. GPT-2 to help correct grammatical mistakes