

The Role of AI and Neural Networks in Fake News Creation and Detection

Rogdakis F., Smpokos E., Zoumpoulakis O.

July 14, 2024

Abstract

This is the abstract of the project. It provides a brief summary of the content.

Contents

| | | |
|----------|---|----------|
| 1 | Introduction | 1 |
| 2 | Dataset | 2 |
| 3 | Using AI to detect Fake News | 2 |
| 3.1 | Using ChatGPT | 2 |
| 3.2 | Using Google's Gemini | 2 |
| 3.3 | Using Microsoft Copilot | 3 |
| 4 | Using Neural Nets to generate and classify fake news | 3 |
| 4.1 | Generator | 3 |
| 4.1.1 | GRU | 3 |
| 4.1.2 | GPT2 | 4 |
| 4.2 | Discriminator | 6 |
| 4.2.1 | Training the classifier | 6 |
| 4.2.2 | Classifying the generated data | 6 |
| 5 | Conclusions | 7 |

1 Introduction

With the rapid development in the field of Machine Learning and the vast information available on the internet, the creation and spread of Fake News has become much easier and more common. Aside from the problem that the authors of fake news and those that are willing to spread them for personal gain, there is also the fact that many people are not willing to check the authenticity of the information they read, due to lack of time or interest. Thus, our project aims to assess the effectiveness of Artificial Intelligence (A.I.) and Neural Networks (N.N.) in detecting fake news. But at the same time we would like to investigate the possibility of using NN to create fake news.

2 Dataset

We will be using a dataset of fake news and real news articles, which we got from “Kaggle”. The main reason that we chose that dataset is that it is already labeled, which means that it would be much easier to compare the result of the AI models’ prediction with the actual label of the news article. For that exact same reason, it is also easier to train the NN model that we made.

3 Using AI to detect Fake News

The main idea behind using A.I. to detect fake news is that there are various AI models out there, like ChatGPT and Gemini, which are available to every person that has access to the internet. For that reason we would like to access the effectiveness of these models, especially their free versions since that would be what the average person would use. Specifically we tried out ChatGPT, Google’s Gemini and Microsoft Copilot which are the most used A.I. when it comes to accessibility and usage.

3.1 Using ChatGPT

We tested ChatGPT’s reliability by providing it with specific fake news articles, listed in the attached Excel sheet. We asked ChatGPT to determine if the article was fake or not based on its title and text. We then saved its response and compared it with the actual result from the dataset. Out of curiosity, we also asked ChatGPT if it was confident in its answer or wanted to check again. We should mention that the test sample contained both Fake and Authentic News.

The results were quite surprising. Without taking into consideration the second time that it was asked to make sure if it was confident in its answer, ChatGPT managed to correctly predict the authenticity of 27/38 articles, which yields an accuracy of 71.05%. That means that it managed to correctly predict 7 out of 10 articles. For an A.I. that it is not a satisfactory result because by having access to the whole internet, it should be able to predict the authenticity of the articles with a much higher accuracy, which was not the case in our test.

By asking ChatGPT to check again, the accuracy of the model increased to 32/38, which yields an accuracy of 84.21%. That result can be considered as a satisfactory result, definitely not perfect but it is what we would have expected. ChatGPT managed to realise its mistakes and correct them, giving a more accurate result. But we should not have to ask it to check again, it should be able to predict the authenticity of the articles with a higher accuracy from the first time.

We were not able to use ChatGPT 4.0 due to the limitations of the free version.

3.2 Using Google’s Gemini

Unfortunately, we were not able to get any results from Google’s Gemini and that is unacceptable from a company like Google which tried to compete with OpenAI. We begun our testing by asking Gemini the same way we asked ChatGPT and till that part everything was going smoothly. But upon sending the articles, Gemini was not able to provide us with an answer, despite our efforts to force it. It kept saying that it could not

form an opinion and that we should use Google search and arrive at our own conclusion. Unacceptable.

3.3 Using Microsoft Copilot

Microsoft Copilot was slightly better than Google's Gemini but still we did not get any results. We begun our testing as we did with the previous two, but upon sending the articles, Copilot was not able to provide us with an answer, despite our efforts. It would vaguely respond by giving us a general idea of what the article was about and how we could check the article's authenticity by ourselves. If we demanded an answer, it would stop the conversation and prompt us to start a new one. Again, unacceptable.

4 Using Neural Nets to generate and classify fake news

4.1 Generator

For text generation we tried two techniques which are stated below. First, we tried the diy method and built a Gated Recurrent Unit (GRU). Then we also tried using a Generative Pre-trained Transformer (GPT)

4.1.1 GRU

Our GRU model is summarised below.

| Layer (type) | Output Shape | Param # |
|-------------------------------------|-------------------|---------|
| embedding (Embedding) | (None, None, 64) | 18752 |
| gru (GRU) | (None, None, 128) | 74496 |
| dropout (Dropout) | (None, None, 128) | 0 |
| gru_1 (GRU) | (None, None, 256) | 296448 |
| gru_2 (GRU) | (None, None, 128) | 148224 |
| dropout_1 (Dropout) | (None, None, 128) | 0 |
| gru_3 (GRU) | (None, None, 256) | 296448 |
| dense (Dense) | (None, None, 293) | 75301 |
| Total params: 909669 (3.47 MB) | | |
| Trainable params: 909669 (3.47 MB) | | |
| Non-trainable params: 0 (0.00 Byte) | | |

It consists of an embedding layer (that converts integers into dense vectors), various GRU layers (which can learn the context and structure of text) and a couple of dropout layers meant to prevent over-fitting.

Before the data is fed into the NN ,preprocessing is handled by keras, including text tokenization and vectorization.This step converts the raw text into a numerical format suitable for model training.

After the model is trained (eventually on a much larger dataset which consists of around 45000 real articles) , a customised function is used for text generation. This function takes a small piece of text as input (a.k.a a prompt) and is able to generate the subsequent text based on the learned patterns in the training data.

Some of the samples created are presented below.

| Prompt | Generated Text |
|--------------------------|--|
| The elections are | The elections are leavel the well his donald trump was can't be the fbi did the finally when the alternate current radio network trump last we are change anyone of the final press deserior posted |
| The future of AI is | The future of AI is a president democrat donald trump a member of mark with donald trump doesn't want donald trump donald trump is the democrats democratic democratic democratic was more and ... |
| Climate change impacts | Climate change impacts to press new york trump is a stranded that white held and with donald trump was federal democrats again the white house we all known the fbi didn't be democratic president |
| Health and wellness tips | Health and wellness tips and clinton department the media didn't speak the white was his media democrat donald trump is on facebook with defense the world despite email the media department defense the demo |

It is easy to notice that the text generated above lacks coherency and is very repetitive. This means that the NN did not go through the whole training dataset and only 'learned' a small portion of the given text.

Comment: The NN could have been retrained on the same or larger dataset but the volume of the train dataset was already too big and we had serious problems while trying to train our model.

4.1.2 GPT2

Not being satisfied by the text generated by the GRU model we were driven to try another approach.

GPT-2 is pre-trained on a massive amount of text data. Also it's based on a Transformer architecture, which is great at handling long-range dependencies in text. This means it can understand relationships between words and phrases across longer spans of text, leading to more coherent outputs.

Although this model needed no training, we struggled to make it produce text of average article length. In order to do so, we used a grid search which checked various parameter values, different prompts and was able to determine the optimal combination of parameters in order to get text closer to our expectations.

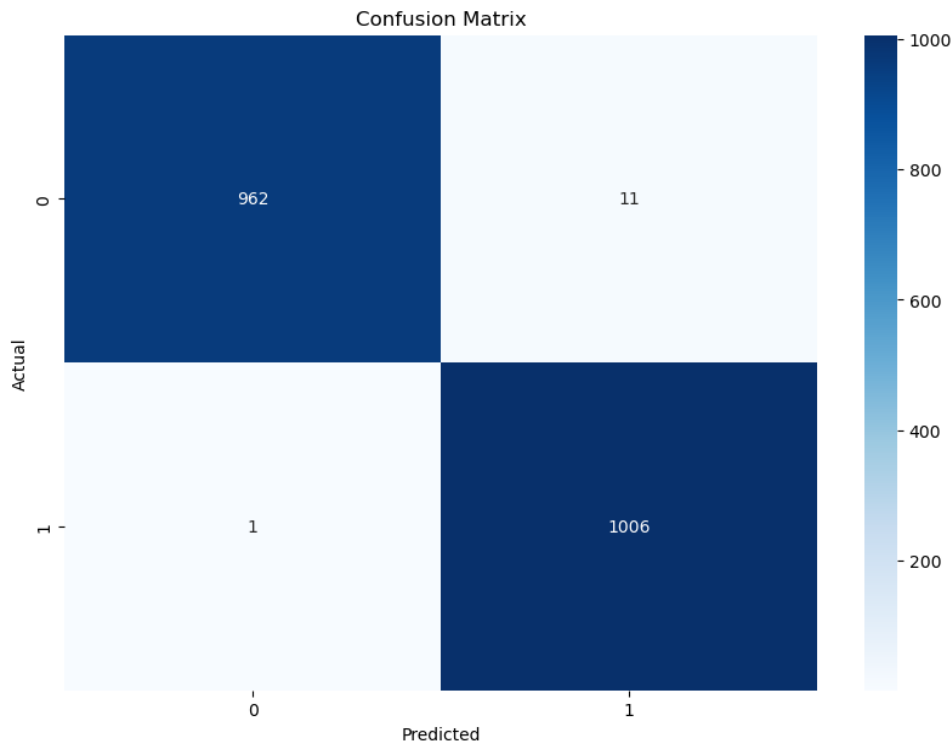
Some of the text generated with gpt2 is shown below.

| Prompt | Generated Text |
|--|--|
| BREAKING NEWS! There has been an attack | BREAKING NEWS! There has been an attack on a mosque in the UK. We are working with the police to establish the facts. — Donald J. Trump (@realDonaldTrump) March 22, 2017. The White House did not immediately respond to a request for comment. |
| In his latest public speech Trump said | In his latest public speech Trump said, "I am the least racist person that you've ever seen in your entire life. Ever." "I don't care what color you are, where you come from, what your religion is, how much money you make, whether you're rich or poor, black, white, Hispanic, Asian, Native American, whatever," he continued. "We all bleed the same red blood of patriots and we all love this country." |
| An expected increase in the prices of fruits | An expected increase in the prices of fruits and vegetables, as well as a rise in fuel prices, will make it more difficult for many families to afford food. "It's going to be very hard for people to get by," said Naveen Jain, executive director of the Indian American Legal Defense and Educational Fund. "We're talking about a lot of people who have been struggling for a long time." |
| The british museum has announced | The British Museum has announced that it will be opening its doors to the public for the first time in over a century. The British Museum, which was founded in 1865, is one of the oldest museums in the world and is renowned for its collection of artefacts from across the globe. It is also home to some of Britain's most famous artworks, including works by Leonardo da Vinci, Rembrandt van Rijn, Vincent van Gogh, and Andy Warhol. |

4.2 Discriminator

4.2.1 Training the classifier

In order to be able to train the discriminator we first need to preprocess and tokenize the dataset. After doing that we are ready to build and train our classifier. It is a sequential model consisted of three layers. It starts with an embedding layer to learn dense representations of words, followed by an LSTM layer to capture sequential dependencies, and ends with a dense layer and sigmoid activation for binary classification. We were able to achieve an accuracy score of 99.6% on the validation set. The confusion matrix and the classification report are presented below.



| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Fake | 1.00 | 0.99 | 0.99 | 973 |
| Real | 0.99 | 1.00 | 0.99 | 1007 |
| accuracy | | | 0.99 | 1980 |
| macro avg | 0.99 | 0.99 | 0.99 | 1980 |
| weighted avg | 0.99 | 0.99 | 0.99 | 1980 |

4.2.2 Classifying the generated data

We used our discriminator only on the data generated by GPT-2 because we were satisfied with it, opposed to the one generated by GRU. It was able to classify correct the majority of the dataset. As we can see it got an accuracy score of 66.67%. Keep in mind that all the true labels are fake.

| | Generated Text | predicted_label |
|---|---|-----------------|
| 0 | breaking news attack mosque uk working police ... | Fake |
| 1 | latest public speech trump said least racist p... | Fake |
| 2 | expected increase prices fruits vegetables wel... | Real |
| 3 | british museum announced opening doors public ... | Real |
| 4 | new movie openheimer released openheimer chara... | Fake |
| 5 | met gala making long time coming excited final... | Fake |
| 6 | white house news lot lately wrong reasons pres... | Real |
| 7 | exciting discovery made team researchers unive... | Fake |
| 8 | agatha kristy published new book subject calle... | Fake |

5 Conclusions

All in all, we are pretty satisfied with the results we got from both the generated data(GPT-2) and the discriminator. We were hoping for better results from our GRU model. In order to achieve better results we need very large datasets for the training which requires lots of time running the code and lots of memory that sadly our equipment didn't have enough.