Understanding Machine vs Human Generated Text in News

TOPIC: CLIMATE CHANGE

GROUP 10

Raakesh Sureshkumar
Computer Informatics and Decision Systems Engineering
Arizona State University
rsures14@asu.edu

Dunchuan Wu
Computer Informatics and Decision Systems Engineering
Arizona State University

***Abstract*—With the advent of AI adversaries, there is mass generation of good quality fake news around different topics of interest. These differences are quite subtle and can be identified only by the models that generated it which is quite counter-intuitive. So, in this project focuses on extracting human written news articles by journalists and comparing them with Artificial Intelligence model generated news articles with the prompt as the news article headings.**

Keywords—adversary, artificial neural network, Recurrent Neural Networks, tokenization, Rectified linear unit, Long short-term memory, Summarization,

I. Introduction

The objective of this project to crawl social media (mainly news content webpages) that contains authentic human generated text. After collecting this text, the heading of this text is to be used in NLP AI models to generate fake news that is non-existential in the real-world scenario. There are many text generation models that can be used to create quite convincing non-existential text starting from the sequence-to-sequence model to the advanced transformer-based model called the GPT-2 having stacks of encoders and decoders.

Machine Learning has been playing a big role in our daily lives. It has many different aspects of application. It has been said that AI can draw pictures, play RTS games against top gamers and write a poem. Now, when developers enrich the functionalities of AI, their product can be beneficial and harmful at the same time.

One problem that comes into my view is that those AI who can generate text have their own impact in a bad way, which is that they can produce disinformation. If machine-generated articles fool humans when we deal with financial cases, scientific facts or public media, it is expected that there will be more social problems. To better understand the text generation models, we need to study their results and find out the difference between human written and machine written text.

II. Literature Review

A. *Ashish Vaswani, Llion Jones, Noam Shazeer, Niki Parmar, Aidan N. Gomez, Illia Polosukhin, Jakob Uszkoreit, Łukasz Kaiser.* ***Attention Is All You Need****.* *arXiv:1706.03762v5 [cs.CL] 6 Dec 2017*

Overview: The major focus is on empowering the transformer architecture consisting of multiple stacks of encoders and decoders (about 6 for each in the GPT-2 model released in the year 2019 by OpenAI). This addresses the previously existing problem of RNNs (Recurrent Neural Networks). Even with advancements like the bi-directional, multi-layer, memory based gates (LSTMs [Long Short Term Memory] / GRUs [Gated Recurrent Units]), there were problems like the vanishing gradient which occurred due to the sequential processing of a single word at a time.

$$P(w_{1:T}|\,W_0) = \prod_{t=1} P(w_t|w_{1:t-1}, W_0)\ \ with\ 1:0 = \phi$$

The concept of attention was brought into the picture to solve this problem. Instead of a single token vector to represent a sentence, this consisted of a context vector that takes a global approach on the inputs and gives the most significant tokens. Mathematically, this can be achieved using cosine similarity.

Two types of attention are discussed in this paper. Scaled dot-product which computes the dot product of 2 queries Q and keys K and scales it to 1 / d, where d represents the dimensions of the keys after tokenization. Then it is multiplied with the values V matrix. The next type would be the Multi-Head Attention which involve the projection of the queries, keys and values. These values are concatenated and then projected to get the output vector of d dimension. This also ensures parallelization in the process of obtaining the final vector.

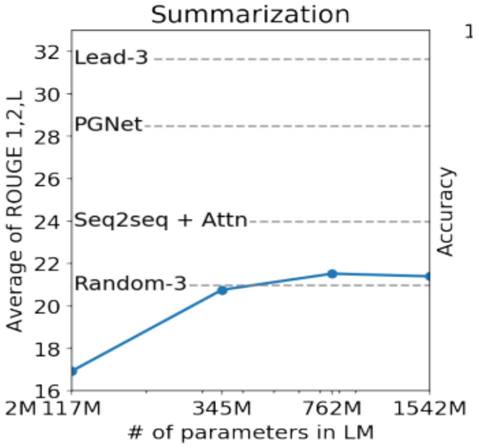
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

The model architecture consists of 2 linear transformations with 2 weights and 2 biases. ReLU (Rectified Linear Unit) is the activation function used. The input and output layers consist of 512 nodes with the hidden layers consisting of about 2048 layers

Further the significance of self-attention is explained. One is the total computational complexity per layer. Another is the amount of computation that can be parallelized, as measured by the minimum number of sequential operations required. The third is the path length between long-range dependencies in the network. The major advantage of using a self-attention model compared to a recurrent model is that a self-attention layer connects all positions with a constant number of sequentially executed operations, whereas a recurrent layer requires O(n) sequential operations.

The task could either be taken as a question answering task or a summarization task. In both ways, the key terms are obtained from the prompt and the upcoming token probabilities are set based on this.

FIGURE 1 SUMMARIZATION TASK ACCURACY FOR GPT-2 USING ROUGE METRIC



B. *Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, Yejin Choi. **The curious case of neural text degeneration**. arXiv:1904.09751v2 [cs.CL] 14 Feb 2020*

Overview: The major problems in the existing / previous language generation models is the repetition of text or the loss of context as the flow goes by. In the modern language models, most of the attention is mainly on the most recent few tokens for the generation of the upcoming word. So, sampling methods are being used to mitigate this effect of loss of context. This considers the whole text for the generation and samples it to improve the efficiency of the process. But, a straight-forward sampling could run into a problem too. It could give a chance of 1 in 3 for completely derailing our train of thought as the upcoming word could be out of the bottom 30% of the samples. To manipulate this issue, 2 major processes were brought into effect: temperature sampling and top k sampling.

According to statistical thermodynamics, temperature is inversely proportional to energy. So, here the logits play the role of energy. This analogy specifies that we divide the logits by temperature before feeding them into the SoftMax and obtaining the sampling probabilities. The confidence factor plays a key role in statistical analysis to define the accuracy of the result. This relative factor is found to be inversely proportional to the temperature. If the temperature is greater than 1, then the confidence factor that the model exhibits decreases significantly. 0 temperature corresponds to argmax/max likelihood, while infinite temperature corresponds to a uniform sampling. Note that this value has been set as 0.7 in our automated news article generation for this project.

In the top-k sampling, a kth token is defined, which forms the threshold for filtering out the sorted probabilities that are below the kth token. This ensures an improvement in quality and accuracy of text generation which sticks to the context. By quality, here we refer the coherence in the text. The problem here is that we are not sure at what optimum value for the kth token, there are words with reasonable sampling that form a broad distribution or in some others there are none with a narrow distribution. In our example, we have taken this value as 200.

The most recent and advanced sampling technique is the top p sampling, aka the nucleus sampling. The significance of this sampling is that we use the cumulative distribution function (CDF) instead of the number of suitable samples in order to threshold the required sampled that could best fit the sentence formation. The major advantage of this type of generation is that on an abstract level, if it gets one token wrong by generating a bad distribution, the next token uses the "correct" human generated context independent of the last prediction. Also, it preserves variety in a low confidence state. The typical value for this parameter is usually set as 0.95 where 0.95 represents the CDF and 5% of the most suitable values are considered for the final text generation.

C. *Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi.* ***Defending Against Neural Fake News*** *[arXiv:1905.12616v2 [cs.CL] 29 Oct 2019]*

Overview: This paper is mainly about the new technology called Grover which identifies fake news identifier based on an advanced robust classifier model. It was initially developed as an adversary for the task of summarization that is given a headline, it could generate the rest of the article. The task at hand here is to classify a news article as fake if generated by an AI instead of a human. So, it entirely depended on the dataset available. With huge proportions of available data, the neural net could learn to classify better than humans.

Some facts about Grover include, with more and more AI generated news articles available, Grover can display an accuracy of upto 97%. Secondly, it could also classify about 96% of GPT-2's generation as machine written. Another aspect is that it could even detect human written fake news with an accuracy of 98%.

Certain challenging cases for Grover come up when adversaries use rejection sampling and financial news articles for its classification. Further there are discussions about the classification of the Grover models that are available based upon the amount of data they were trained upon: Grover Base, Grover Large, Grover MEGA. Out of these, the first 2 are available to the public. There is a counter-intuitive study that as the text generator gets better and better, the classifier gets even better. For example, in the paper, an example of 5000 samples in the dataset has shown an accuracy of 92% and this percentage is expected to increase with increase in the available dataset size from Grover. So, when given 10k news article, it produced an accuracy of 94% and 97.5% when trained on 80k articles.

Now comes the interesting part, with just the available 500 sets of sample data from a different source that is currently the world's best text generator by OpenAI a zero-shot learning was conducted. Despite, a different adversary being used and limited dataset, Grover was able to produce an accuracy of 96.1% in the classification process.

Considering the limitations of Grover, rejection-sampling is one major feature used by an adversary that Grover fails to recognize. This follows the concept that an adversary can be trained until a discriminator's feedback is fed into it and the discriminator completely fails to classify the adversary's generated text as a fake one. The testing for this approach was done for 5000 contexts. It consisted of the parameters like headlines, dates, authors and domain. 64 such samples were generated for different nucleus sampling thresholds ranging from 0.9 to 1.0. Note that at this point in time, the verifier doesn't receive any feedback from the adversary, but the adversary does. So, the following three cases were tested upon for classification.

Another argument is that iterative rejection sampling or adversarial filtering is quite efficient only with short sentences (3 or fewer) with the current available NLP models whereas any typical news article may contain hundreds of sentences.

There have also been findings on content based classification that is on what news channel it had been hosted which brought into light the financial news accuracy to go down when compared with the others especially the ones based on stock market.

III. Methodologies

In this section a brief describe is given on how the data is collected and stored in a certain file format provided in the project description. We have chosen to go ahead with the csv file format as the output for the required dataset and to use automated solutions to scrape out the required news articles along with their headings.

A. *Content Scraping*

Belongs to step 1 as per the project description. This step can either be automated or can be done manually. It involves the collection of news articles in text documents which is considered to be the human generated text. The objective is to collect human-generated news articles pertaining to one or more topics from news sources. The requested number of articles is 80-100. So, it has been decided to automate the process of scraping through the news websites to gather the required data. In this part of the project we use Python as the main compiler-based programming language. Selenium and Requests are used as web automation tools. Beautiful Soup is used as a tool for web scraping.

Selenium

Selenium is a free (open source) automated testing framework used to validate web applications across different browsers and platforms. It supports multiple programming languages like Java, C#, Python. Selenium Software is not just a single tool but a suite of software, each piece catering to different testing needs like the Selenium IDE, Remote Control, Web driver and Grid.

In this case, we use the Selenium web driver to automate the process of getting into the required URL and controlling the website navigation in order to get the required HTML for processing. In our case, we have used the Chrome web driver as it is most widely used apart from Firefox. This HTML file can later be used by the Beautiful Soup to extract the required news articles.

The request for the CNN or Fox news URL is posted by Selenium via the web driver. This fetches the HTML5 webpages as objects in Python which can be parsed by Beautiful Soup.

Beautiful Soup

Beautiful Soup is a Python library for getting data out of HTML, XML, and other markup languages. Beautiful Soup helps you pull particular content from a webpage, remove the HTML markup, and save the information. It is a tool for web scraping that helps you clean up and parse the documents you have pulled down from the web. It provides idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

In our case, we create a list of items to store in memory while scraping out the data.

The technique used for automated text generation is quite important in determining the difference between human generated text and machine generated text. So, here we use 2 different techniques to be on the safer side of knowing the actual difference between the human-generated and the machine generated text.

Requests

Requests is a web tool that allows you to send HTTP/1.1 requests extremely easily. There would be no requirement but to add query strings to your URLs, or to form-encode the POST data. The base package used by Requests is urllib3 which is a primitive package to POST requests to websites to get to a particular webpage from the World Wide Web.

B. *Machine Text Generation*

Belongs to step 1 as per the project description. This step can either be automated or can be done manually. It involves the collection of news articles in text documents which is considered to be the human generated text.

GPT-2 model

There are usually 4 modes or models in GPT-2 for automated text generation. These correspond to reading comprehension, translation, summarization, question answering. In this particular task, we have used the summarization task. This takes in a prompt - basically a text (in our case, the new heading) as an input and generates the continuing text from it.

The models are trained on various parameters as per the neural net framework. The number of parameters represent the number of input nodes in the neural net that form the connections to pass on the values with the updated values of weights and biases. So, the below represents a plot between the number of parameters used and the ROUGE 1,2, L metric that is used to detect its accuracy relative to the human text.

GPT-2 (a successor to GPT), was trained simply to predict the next word in 40GB of Internet text.

Using GPT-2 Transformers

Most competitive neural sequence transduction models have an encoder-decoder structure. Here, the encoder maps an input sequence of symbol representations (x1, ..., xn) to a sequence of continuous representations z = (z1, ..., zn). Given z, the decoder then generates an output sequence (y1, ..., ym) of symbols one element at a time. At each step the model is auto regressive, consuming the previously generated symbols as additional input when generating the next. The Transformer follows this overall architecture using stacked self-attention and pointwise, fully connected layers for both the encoder and decoder respectively [2].

The following are the decoding methods that are most widely used: Greedy search, Beam search, Top-K sampling and Top-p sampling.

a. Working of greedy search:

Greedy search simply selects the word with the highest probability as its next word:

$$w_t = argma_w P(w|w_{1:t-1})$$

Disadvantages:

Starts repeating itself after a certain number of word sequence. It also misses high probability words as the tokens with high conditional probability are at times “hidden” behind low conditional probability words as it considers the probability of the next words by multiplying the conditional probabilities of the current and upcoming word.

b. Working of beam search:

Here, by specifying the number of beams, we can let the system decide the top 2-word sequences with their cumulative probabilities to form the next word in the sequence. Here, each beam is a hypothesis of a sequence of words that can be used to append to the current word to form a new sequence.

The keyword *early_stopping* is set to True so that generation is finished when all beam hypotheses reached the EOS token. The EOS token is the End of Sequence token which is used to mark the end of a sentence.

Disadvantages:

The output includes repetitions though the choice of words is better now with improved probability.

c. Using n-grams penalty:

n-grams refers to a sequence of n words. A penalty on this is set when repetition occurs for these words. This makes sure that no n-word sequences appear twice. This is set manually by setting the probability of upcoming words that has the capability of creating an already existing n-gram sequence to zero.

The corresponding parameter in the gpt-2 model is *no_repeat_ngram_size*. This will ensure that n-grams doesn’t repeat the number of times provided as input. For example, by setting the value to 2, a particular word sequence doesn’t occur twice. But this again has a disadvantage.

Disadvantages:

The randomness in the word generation becomes so high that the upcoming text loses the context of the general summary.

d. Choosing the best among the top beams of text sequences:

This particular feature gives an n number of outputs from which we could analyze and choose the one that best fits our context and purpose. This is provided as the *num_return_sequences* parameter. For example, if there are a total of 10 beams that are possible from a given prompt, then, *num_return_sequences* number of beams will be returned out of the 10 beams. So, we need to ensure that this number is less than the total number of possible beams. The number of possible beams depends upon the training data used to create the model.

Disadvantages:

In particular to beam search, this technique does not return much different sequences that is it showcases repetitive generation even with the top 10 to 15 of its various text combination / sequence outputs. In particular, this is disadvantageous during story generation and dialog creation.
High quality human language does not follow a distribution of high probability next words. In other words, as humans, we want generated text to surprise us and not to be boring/predictable.

e. The Sampling Approach:

This is just the opposite of what we had been trying to accomplish in the previous approaches. Here, the repetition might almost be nil. But there will surely be problem of the text going out of the context.

The next word in the sequence, represents the conditional probability of choosing the word w based on the occurrence of any of the words in the previous sequence.

This can be adjusted by using the *do_sample* variable which is a Boolean to turn on sampling at the output sequence.

Disadvantages:

Coherence in the text is lost as the generation is made by bringing in high randomness.

f. Varying the temperature of the SoftMax:

The SoftMax is generally the final layer in a neural network, in this case an RNN to generate a sequence model. So, in our case, the SoftMax spits out the required text based upon the weights that specify the significance for a particular word based upon its frequency of occurrence in the trained model input.

Our objective here is to decrease the randomness and hence cool down the *temperature* of the SoftMax function. The range of temperature for the SoftMax is 0 to 1 typically. The optimum value for this is set to be 0.7 in our example.

Disadvantages:

An optimum value is to be set for this parameter as a value 0 estimates the highest confidence that a model can give but, the randomness increases proportionally. Though this keeps the text interesting and coherent, the context is lost. Again, a value greater than 1 causes the confidence to go significantly low.

f. Top-k sampling:

Just as explained in section [B] of the literature review corresponding to the scientific paper "The curious case of neural text degeneration", it considers the top k samples that give the highest confidence for predicting the upcoming tokens based on the existing text in hand. The typical value chosen for this is 200.

Disadvantages:

But this may still have the disadvantage of having a good distribution of suitable samples below the broad distribution and low confidence in case of narrow distribution.

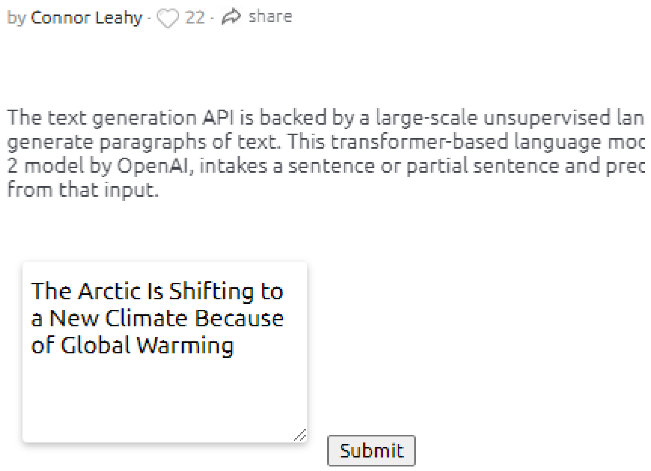
f. Top-p sampling:

Again, this concept is explained in section [B] of the literature review. This corresponds to the scientific paper "The curious case of neural text degeneration". This considers the CDF (cumulative distribution function with an optimum value of 0.95 chosen f or our project that only the samples that are estimated to give a high score for the confidence for the upcoming tokens.

DeepAI model and API

DeepAI model, according to the description of their official website, is a transformer-based language model based on GPT-2. Its text generation API takes a sentence(headline) as input and produces some paragraphs as the output.

FIGURE 2 DEEPAI ONLINE API INPUT

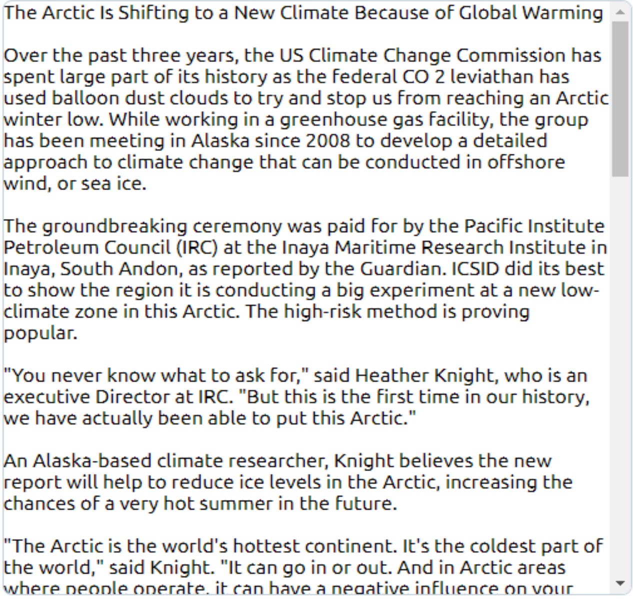


by Connor Leahy · ♡ 22 · share

The text generation API is backed by a large-scale unsupervised lan
generate paragraphs of text. This transformer-based language moc
2 model by OpenAI, intakes a sentence or partial sentence and prec
from that input.

The Arctic Is Shifting to a New Climate Because of Global Warming

FIGURE 3 DEEPAI ONLINE API OUTPUT



The Arctic Is Shifting to a New Climate Because of Global Warming

Over the past three years, the US Climate Change Commission has spent large part of its history as the federal CO 2 leviathan has used balloon dust clouds to try and stop us from reaching an Arctic winter low. While working in a greenhouse gas facility, the group has been meeting in Alaska since 2008 to develop a detailed approach to climate change that can be conducted in offshore wind, or sea ice.

The groundbreaking ceremony was paid for by the Pacific Institute Petroleum Council (IRC) at the Inaya Maritime Research Institute in Inaya, South Andon, as reported by the Guardian. ICSID did its best to show the region it is conducting a big experiment at a new low-climate zone in this Arctic. The high-risk method is proving popular.

"You never know what to ask for," said Heather Knight, who is an executive Director at IRC. "But this is the first time in our history, we have actually been able to put this Arctic."

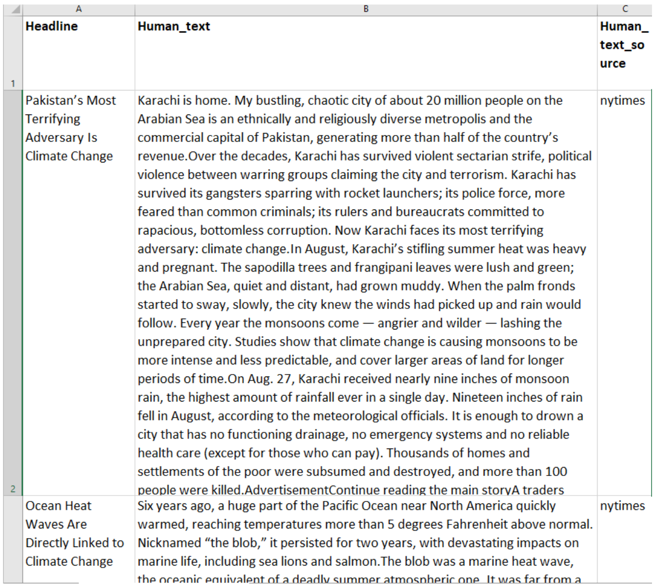
An Alaska-based climate researcher, Knight believes the new report will help to reduce ice levels in the Arctic, increasing the chances of a very hot summer in the future.

"The Arctic is the world's hottest continent. It's the coldest part of the world," said Knight. "It can go in or out. And in Arctic areas where people operate, it can have a negative influence on your

IV. Dataset

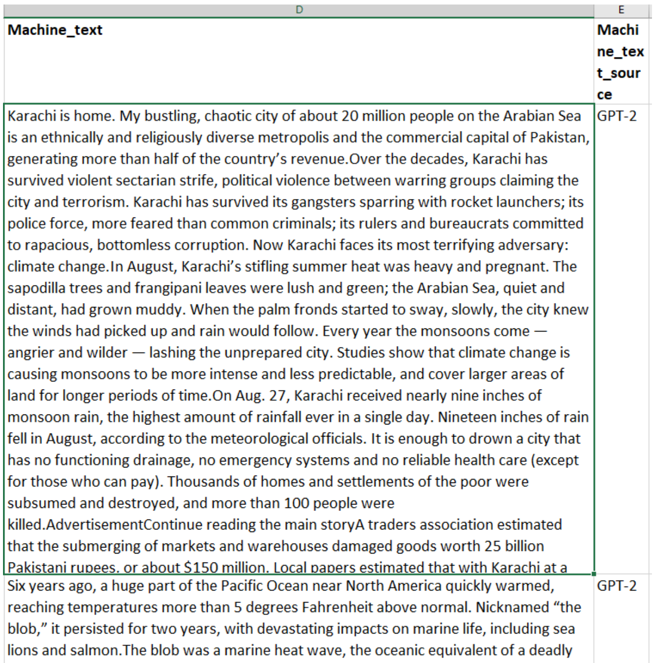
As per the requirement, the dataset consists of the following headings: headline (generic), human_text (the news article obtained from an internet source), human_text_source (the news agency from which the article has been taken), machine_text (news article created using the summarization approach with the headline as the prompt for generation), machine_text_source (the model used to generate the text)

FIGURE 4 CSV DATASET SAMPLE HUMAN TEXT COLUMNS



Headline	Human_text	Human_text_source
Pakistan's Most Terrifying Adversary Is Climate Change	Karachi is home. My bustling, chaotic city of about 20 million people on the Arabian Sea is an ethnically and religiously diverse metropolis and the commercial capital of Pakistan, generating more than half of the country's revenue.Over the decades, Karachi has survived violent sectarian strife, political violence between warring groups claiming the city and terrorism. Karachi has survived its gangsters sparring with rocket launchers; its police force, more feared than common criminals; its rulers and bureaucrats committed to rapacious, bottomless corruption. Now Karachi faces its most terrifying adversary: climate change.In August, Karachi's stifling summer heat was heavy and pregnant. The sapodilla trees and frangipani leaves were lush and green; the Arabian Sea, quiet and distant, had grown muddy. When the palm fronds started to sway, slowly, the city knew the winds had picked up and rain would follow. Every year the monsoons come — angrier and wilder — lashing the unprepared city. Studies show that climate change is causing monsoons to be more intense and less predictable, and cover larger areas of land for longer periods of time.On Aug. 27, Karachi received nearly nine inches of monsoon rain, the highest amount of rainfall ever in a single day. Nineteen inches of rain fell in August, according to the meteorological officials. It is enough to drown a city that has no functioning drainage, no emergency systems and no reliable health care (except for those who can pay). Thousands of homes and settlements of the poor were subsumed and destroyed, and more than 100 people were killed.AdvertisementContinue reading the main storyA traders	nytimes
Ocean Heat Waves Are Directly Linked to Climate Change	Six years ago, a huge part of the Pacific Ocean near North America quickly warmed, reaching temperatures more than 5 degrees Fahrenheit above normal. Nicknamed "the blob," it persisted for two years, with devastating impacts on marine life, including sea lions and salmon.The blob was a marine heat wave, the oceanic equivalent of a deadly summer atmospheric one. It was far from a	nytimes

FIGURE 5 CSV DATASET SAMPLES MACHINE TEXT COLUMNS



Machine_text	Machine_text_source
Karachi is home. My bustling, chaotic city of about 20 million people on the Arabian Sea is an ethnically and religiously diverse metropolis and the commercial capital of Pakistan, generating more than half of the country's revenue.Over the decades, Karachi has survived violent sectarian strife, political violence between warring groups claiming the city and terrorism. Karachi has survived its gangsters sparring with rocket launchers; its police force, more feared than common criminals; its rulers and bureaucrats committed to rapacious, bottomless corruption. Now Karachi faces its most terrifying adversary: climate change.In August, Karachi's stifling summer heat was heavy and pregnant. The sapodilla trees and frangipani leaves were lush and green; the Arabian Sea, quiet and distant, had grown muddy. When the palm fronds started to sway, slowly, the city knew the winds had picked up and rain would follow. Every year the monsoons come — angrier and wilder — lashing the unprepared city. Studies show that climate change is causing monsoons to be more intense and less predictable, and cover larger areas of land for longer periods of time.On Aug. 27, Karachi received nearly nine inches of monsoon rain, the highest amount of rainfall ever in a single day. Nineteen inches of rain fell in August, according to the meteorological officials. It is enough to drown a city that has no functioning drainage, no emergency systems and no reliable health care (except for those who can pay). Thousands of homes and settlements of the poor were subsumed and destroyed, and more than 100 people were killed.AdvertisementContinue reading the main storyA traders association estimated that the submerging of markets and warehouses damaged goods worth 25 billion Pakistani rupees, or about $150 million. Local papers estimated that with Karachi at a	GPT-2
Six years ago, a huge part of the Pacific Ocean near North America quickly warmed, reaching temperatures more than 5 degrees Fahrenheit above normal. Nicknamed "the blob," it persisted for two years, with devastating impacts on marine life, including sea lions and salmon.The blob was a marine heat wave, the oceanic equivalent of a deadly	GPT-2

Pandas

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal. The two primary data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering

Why choose Pandas / CSV

Our other preferred choices were JSON and SQL. Pandas can handle huge database in the form of series or dataframes that act as tables with columnar data. As the number of articles is quite less (80 to 100) and there is no need for faster querying or complex querying, SQL was avoided. Also, it involves certain amount of data processing before storage and after extraction, so JSON was avoided. Pandas thus would be the best choice among the 3 with the required number of tools or functionalities. Also, data is stored as CSV (comma separated values) for easy conversion from dataframes and good visualization as a local file using Microsoft Excel or Libre Office.

V. OBSERVATIONS AND DISCUSSIONS

OBSERVATION SET 1

1. Interesting differences between human and machine-generated texts:

There are three main aspects to this:

a. Randomness in the text: In mathematical terms this is denoted by the sampling theory and entropy.

   a. Human: Given any context, the human mind tries to collect variety of experiences from its past and tries to put it together in a language that everyone can understand. This involves a certain level of randomness and repetition is avoided.
   b. Machine: This is an entity that we specify as an input in terms of temperature, a random seed or n_grams. Currently, no adaptive technology is available that can choose these values dynamically to induce randomness the way we humans do [1].

b. Contextual integrity: In mathematical terms this is denoted by probability and confidence of an AI model.

   a. Human: Journalists usually try and stick to the context. By context, here, the topic or subject of interest is referred. For example, we are given the topic of "Climate change" which can have further subjects within themselves like the "California fires". All the text written in a news article by a journalist would always revolve around this context and never go off topic for that particular topic.
   b. Machine: Again, this is an aspect that has to consider the previously available texts for generating its upcoming tokens in order to keep itself in the context. In our case, the GPT-2 model uses the top-k and top-p sampling techniques that provide a global view on the prompt and the generated tokens to finally produce the desired resultant token with a high confidence [0].

c. Grammatical correctness: In mathematical terms this is denoted by probability and confidence of an AI model.

   a. Human: Editors have a role of verifying the text before a news article has been released.
   b. Machine: This is entirely based upon the model used for training and the probability distribution token that is fed as a threshold by the user.

2. If you did not know which is which, would you be able to distinguish between the two:

This can be considered context based and the dataset with which the adversaries were trained. So, considering two examples from the dataset:

Example 1: Headline: U.S. and European Oil Giants Go Different Ways on Climate Change

Extract 1: The global fossil fuel industry is at high risk of becoming at risk due to climate change, particularly since global coal consumption has increased over the past few decades, said Sean Fitzgerald of the Canadian Oil Sands Coalition at Argonne National Laboratory. Fitzgerald said there were questions that will be asked from experts as to how climate change would affect how much a newly developed country could produce

Extract 2: The plants also absorb carbon through photosynthesis, which Exxon scientists are trying to speed up while producing more oil. "Step 1, you have to do the science, and it is impossible to put a deadline on discovery," said Vijay Swarup, Exxon's vice president for research and development. Research into fusion, algae and carbon capture has been going on for decades

Inference: Here, the fit might seem really good as there is negligible difference that we could make out between extract 1 and 2 and both seem real. However, here extract 1 is the machine text by GPT-2 Large model and extract 2 is the human generated text.

Example 2: Headline: Hurricane Sally Is a Slow-Moving Threat. Climate Change Might Be Why.

Extract 1: A hurricane warning remained in effect for an area stretching eastward from Bay St. Louis, Miss., near the Louisiana border, to Navarre, near the tip of the Florida panhandle — a distance that includes most of Mississippi's and Alabama's coastlines. A tropical storm warning covered the area west of the Pearl River to Grand Isle, La. — including metropolitan New Orleans — and east of Navarre to Indian Pass

Extract 2: The Antarctic Ice Caps Melt From Large Volumes Of Antarctic Ice High Altitude Arctic Global Temperatures Rise Over The Past 15 Years, Here's What Scientists Say Arctic Streams galaxy-GLOW FOR JULY 2014In the Deep South Pacific Sea Equestrian Activity Is on High Over Time Pilot (Astroon Dickinson Space Sciences Laboratory, University of Bristol)

Inference: Here, the difference is quite clear between the human and machine generated text due to the loss of context and certain grammatical errors. By now, we could make out that Extract 1 is the human generated text and Extract 2 is machine generated text.

3. Observation on STYLE:

This seems to be a play with the randomness and the sampling factor. With the most recent advent in transformer model, choosing optimum value of temperature, top-k and top-p are quite important to maintain the consistency. The articles seem to be fairly consistent with the current input parameters for temperature, top-k and top-p as 0.7, 300 and 0.95 respectively.

This could be made more reliable by the adaptive change of the aforementioned values by taking feedback during the backpropagation in the neural nets.

4. Observation on CONTENT:

As discussed in section 1 of the observations, the content of the article does not make sense at times based upon the context and the input prompt. The major factor for this the training data and the search technique used for the upcoming tokens. For example, search techniques like the beam and the greedy search do not necessarily produce content-rich text as they consider just the last word before the generated token

Fact checking is one way we could identify if an article is fake or not. Since journalists usually obliged to submit fact-checked resources, we could differentiate between the machine generated text in terms of the content if the classifier (human / bot) is trained in that subject.

5. Observation OVERALL:

The technology used in this project is the latest as of now that is freely available as an adversary (GPT-2 Large) with about 117 million parameters. The original GPT-2 consists of 1.5 billion parameters trained on over 8 million websites. A release of gpt-3 model is anticipated this year by OpenAI (thanks to Elon Musk and Microsoft) which is much more powerful than gpt-2 model.

At the end of the day, it may be difficult for humans to understand the hidden pattern in a machine generated text until and unless they are expert in the subject and are good at fact checking.

So, it is understood that separate counter-intuitive models have been generated that act as both adversaries and classifiers between human machine generated text. The latest technology on this is the Grover which measures the accuracy of classification on various scales like BLUE, ROUGE, Cosine similarity, Jaccard similarity etc. Out of this, the ROUGE metric is used for detecting the flaws in the summarization process which has been used in this project.

OBSERVATION SET 2

Generally:

It takes only a few seconds to finish crawling all human articles from a website.

A lot of the results retrieved from CNN.com with a keyword “climate change” are irrelevant to our topic, rather, it shows so much political content. NY times can return articles that are closely related to climate change.

DeepAI model sometimes cannot produce a result. Instead of returning an article, it returned an error message stating that there is something wrong with the input, try it again. After one trail of execution of our program, 80 are supposed to be stored in the database, leaving 19 cells containing the error message.

DeepAI takes too much time to generate 80 articles due to the response speed of the internet and their website’s limited ability to process requests.

To answer specific questions:

1. Is there any interesting difference between human and machine-generated texts?
For machine-generated texts:

Grammarly: Some basic error on nouns like (These tool/ These tools, singular or plural format is not carefully handled)

Double spelled a word: For example, we saw “by targeting audience audiences”, “these tools can be used to attack the infrastructure, infrastructure, or infrastructure that have been compromised...” as the text literally prints out.

Content seems not closely related to the title: keywords from the title like president, fire, pandemic is rarely mentioned. Machine-generated articles are talking about something not related to the topic. (To be fair, the unrelated content problem of DeepAI model is caused by the wrongdoing of CNN.com search engine. It returns unrelated human’s articles at first)

For human-generated articles:

There are no such problems. Every facet of results from human-generated articles is opposite to what is mentioned in the last paragraph.

2. If you did not know which is which, would you be able to distinguish between the two?

I cannot distinguish between them. Both can state facts and tell me what happened recently as I heard from public media. Even though there are some stories that I did not know, they both make the statement look real.

• Style - Is the style of the generated article consistent?
Yes, the style of the generated article is consistent. Each one has a formal and professional style with a title and main content divided into separate paragraphs.

• Content - Does the content of this article make sense?
Yes, the content of this article makes sense.

• Overall- Does the article read like it comes from a trustworthy source?
Yes, the article reads like it comes from a trusted source.

VI. Deliverables

A. *Code Requirements*

Here, the various toolkits along with their versions required will be specified to run and test in the local system. Also, it is important to note that the “nytimes.com” website is to be subscribed before scraping out the data. For this project, the nytimes.com was subscribed for a week in order to scrape the data out.

Please note that the code was developed in the python v3.8.1 compiled language. The PyCharm IDE with windows as the OS was used for developing the code base. Please ensure that the latest version of pip v20.2.3 is installed before proceeding with the installation. The following packages are to be included in the virtual environment before running the code:

TABLE I. REQUIRED PACKAGES AND VERSIONS

Sl. No	Package name	Version
1	transformers	3.2.0
2	requests	2.24.0
3	selenium	3.141.0
4	beautifulsoup4	4.9.2
5	chromedriver	2.24.1
6	tensorflow-gpu	2.3.1
7	tensorflow-cpu	2.3.1
8	pandas	1.1.2
9	html5lib	1.1

B. *Code Settings*

These are the settings / variables that are included in the code base for convenient running of the code and toggling between output displays:

1. DISPLAY_OUTPUT - Choice to display the human generated text output or not

2. DISPLAY_MACHINE_TEXT - Choice to display the machine generated text output or not

3. GPT2 - Choice to generate the GPT-2 based machine generated text output or not

4. DEEP_AI - Choice to generate the Deep AI based machine generated text output or not

5. CNN - Give True for CNN scraping and False for NYTIMES scraping

6. DISPLAY_NYTIMES - Choice to display the machine generated text output or not for Nytimes

7. DISPLAY_CHROME_HEAD - Displays the chrome browser when Selenium runs

Please note that only one of the machine generation text models can be chosen at a time

VII. REFERENCES

[0] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. ***Attention Is All You Need***. *[*arXiv:1706.03762v5 [cs.CL] 6 Dec 2017]

[[1] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, Yejin Choi. ***The curious case of neural text degeneration***.*[*arXiv:1904.09751v2 [cs.CL] 14 Feb 2020]

[2] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, Yejin Choi. ***Defending Against Neural Fake News*** *[*arXiv:1905.12616v2 [cs.CL] 29 Oct 2019*]*

[3] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova. ***BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding***. arXiv:1810.04805v2 [cs.CL] 24 May 2019

[4] Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. ***Language Models are Unsupervised Multitask Learners***.

[5] Sydney Maples, Angela Fan, Mike Lewis, Yann Dauphin. ***Hierarchical Neural Story Generation.***
arXiv:1805.04833v1 [cs.CL] 13 May 2018

ONLINE REFERENCES

[6] https://huggingface.co/blog/how-to-generate

[7]https://medium.com/analytics-vidhya/understanding-the-gpt-2-source-code-part-1-4481328ee10b

[8]https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277

[9]https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[10]https://pandas.pydata.org/pandas-docs/stable/getting_started/overview.html

VIII. APPENDIX

INDIVIDUAL CONTRIBUTIONS

TABLE II. REPORT CONTRIBUTIONS

Sl. No	Package name	Author
1	1. Introduction	Raakesh, Dunchuan
3	2a. Literature Review (Attention is all you need)	Raakesh
4	2b. Literature Review (The curious case of neural text degeneration)	Raakesh
5	2c. Literature Review (Defending Against Neural Fake News)	Raakesh
6	3a. Methodologies (Content Scraping) Selenium; BeautifulSoup; Requests	Raakesh
7	3b. Methodologies (Machine Text Generation) GPT-2 model; Using GPT-2 Transformers	Raakesh
8	DeepAI API and model	Dunchuan
9	Dataset description	Raakesh, Dunchuan
10	Pandas; Why choose pandas	Raakesh
11	Observation Set 1	Raakesh
12	Observation Set 2	Dunchuan
13	Deliverables	Raakesh

TABLE III. CODE BASE CONTRIBUTIONS

Sl. No	Code module	Author
1	Scraping for CNN articles	Raakesh
2	Scraping for NY Times articles	Raakesh
3	Text generation using GPT-2 model	Raakesh
4	Text generation using DeepAI API	Dunchuan
5	Write to CSV	Dunchuan