
A. Project Team members (list team members here)

1. Anita Sarkar
2. Suhail Lala
3. Mansi Shah

B. Project Topic Introduction (2 points) : e.g. NLG or MT (5-10 sentences)

What are you attempting to do? Why is it worth doing?

Our goal for the final project was to use Natural Language Generation (NLG) in a real life scenario. We explored a number of different problems such as weather forecast generation and question answer generation. However, we realized that with the current pandemic, people have many questions regarding COVID-19. There was a lot of work being done to consolidate information available at different locations to answer user queries. With the ongoing pandemic, we found this to be a relevant use of NLG. We explored different models and their performance towards addressing this need.

C. Prior work (4 points) (10-30 sentences)**:** Each group member must contribute at least 1 source (citations, resources, links) based on prior work in this topic area.

Suhail Lala:

- Tutorial explaining GPT-2 https://heartbeat.fritz.ai/exploring-language-models-for-neural-machine-translation-part-two-from-transformers-to-gpt-2-7f045c95dc1e
- Paper talking about BERT in Machine Translation https://arxiv.org/pdf/2002.06823.pdf
- Project using BioBERT + GPT-2 for Medical Q&A https://github.com/re-search/DocProduct
- Fine-tuning pre-trained GPT-2 models https://www.analyticsvidhya.com/blog/2020/07/transfer-learning-for-nlp-fine-tuning-bert-for-text-classification/

Mansi Shah:

- SQUAD project implementation https://github.com/priya-dwivedi/cs224n-Squad-Project

- Blog tutorial for training GPT-2 for generating text in your own language https://towardsdatascience.com/train-gpt-2-in-your-own-language-fc6ad4d60171

Anita Sarkar:

- Open AI's paper on use of GPT-2 for different use case https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Jay Almmar's blog and github for using pretrained GPT2 model https://github.com/jalammar/jalammar.github.io/blob/master/notebooks/Simple_Transformer_Language_Model.ipynb

Initially, we studied the weather datasets and explored ways to generate natural language weather predictions. We found the data to be very intense in terms of preprocessing of weather specific numeric parameters. We also read expert opinions on latest models like GPT-3 not being useful for data to text [4]. We then decided to create a Q&A solution. We first studied the SQUAD dataset. However, we found that the emphasis there was on NLU and selecting answers within the corpus. We finally decided to implement NLG models for COVID Q&A. As a baseline, we used clustering and n-gram models and evaluated. To further improve the performance, we moved to GPT-2. We used Bleu and Rogue scores to test performance of the models.

D. Data sources (1 points) (10-20 sentences)**:** Each group member must propose at least 1 data source for the chosen topic and document it with at least 1 source.

Suhail Lala:

- DataSource 1: Sum time weather data set
https://ehudreiter.files.wordpress.com/2016/12/sumtime.zip

- DataSource 2: COVID qnA dataset
https://www.kaggle.com/xhlulu/covidqa?select=news.csv

Mansi Shah

- DataSource 1: Australian weather dataset - https://www.longpaddock.qld.gov.au/silo/point-data
- DataSource 2: SQUAD Dataset - https://github.com/priya-dwivedi/cs224n-Squad-Project

Anita Sarkar:

- Facebook weather research - https://github.com/facebookresearch/TreeNLG
- Stanford Question and Answer Dataset (SQUAD) https://rajpurkar.github.io/SQuAD-explorer/

Next: Indicate which of the data sources was selected to continue for the project.

We decided to create a question answer model for COVID-19. We selected the COVID QA dataset (https://www.kaggle.com/xhlulu/covidqa?select=news.csv). This has more than 1000 question answer pairs from a number of sources including 15 English news websites across 4 continents, CDC and WHO's official FAQs and 26 Stackexchange communities. We selected only English question answer pairs. The questions and answers are verbose, lending well to NLG model training.

E. Approach (5 points) (10-40 sentences):

Explain what is your solution and how you did it. Which techniques did you use?
Preprocessing:

Clustering:
Our earlier approach consisted of generating text based on a given sequence of words. However, since we wanted the answers to be in response to our question, we needed a way to integrate questions in our model. After considering several approaches, we felt that clustering based on questions would fit our purpose.
Through k-means clustering, we grouped similar questions together. We chose a cluster size of 150 based on the elbow method and created Laplace models for each cluster. For testing, we would run the test questions through the clustering model, find out its cluster number and use the corresponding Laplace model to generate an answer.

Clustering + Five-gram model:
Our initial approach was to use a trigram model to generate an answer. However, after testing it, we found that some of the sequences were not correct. This may be due to the fact that most of our answers were very large. Hence, we went with a five-gram model. We also chose Laplace since it would be a good starting point for our model development. Since Laplace adds 1, it considers all Words for generation.

GPT-2
Next, we chose to implement the transformer model GPT-2 to generate the answers. We used two Approaches. First, we trained a new GPT2 model using our dataset. Second, we used a pre-trained

GPT2 mode and fine tuned it to our dataset. We combined the question answer pair from the training data and used start-of-sentence tokens to distinguish between them. We used a block size of 100 to train the model. We also used Adam optimizer and Sparse Categorical Cross Entropy as the metric for fine-tuning the model. We then tested the trained model by passing a list of general questions which generated the answer based on the starting phrase which was our input question. We found that using the pretrained model, only 10 epochs were sufficient to generate good scores.

How did you measure performance?
We used a sample of random questions as well as Bleu and Rogue scores.

Also explain team member contributions here: which team member contributed to which part of the problem (make a table here to be concise)

	Suhail	Mansi	Anita
Data set identification	Studied different data sources suggested in the assignment to identify the project dataset.	Performed extensive mapping of weather dataset raw data to natural language predictions.	Studied the possibilities and challenges of using the weather dataset. Based on findings, proposed using a different dataset.
Data pre-processing/ Exploratory Data Analysis (EDA)	Combined multiple data sources, tokenized and padded sentences, and added descriptions where required. In the end, I also performed garbage collection and reclaimed memory to prepare for the next stage.	Creating word cloud of most common words, performed Text Preprocessing - by removing accented characters, expanding shortened words, removing special characters, tokenizing, and plotting freq distribution graph.	Performed analysis to understand the COVID QA dataset. Proposed plotting word frequency to observe Zipf's law on chosen dataset.
Initial Model development and Model Evaluation	-Studied and implemented clustering	-Collaborated in evaluating the model.	Studied different evaluation approaches,

	approach to integrate questions in our model, determined a good cluster size, trained models for every cluster and tested their performance on sample questions. -Contributed to the discussion about the choice of evaluation matrices and assisted in their implementation.		reviewed with team and Erfan and proposed using Bleu and Rogue
Model Enhancement	Implemented & fine-tuned pretrained GPT-2 Model. This was the final model showing improved scores.	Implemented GPT-2 Model trained on the COVID QA dataset	Studied GPT2 usage in translation use cases, which would be applicable in our QA use case. Implemented the Seq to Seq Model for the COVID QA dataset for alternative evaluation

We used Python libraries including, but not limited to Tensorflow, nltk, sklearn, pandas, numpy, matplotlib.
We used a pre-trained GPT2 model from HuggingFace Transformers .
We used Pycharm, Google Colab.

F. Results (Very important!!) (15 points, unlimited sentences)

Include relevant observations, measurements, and statistics.

Include graphs, equations, pictures, etc. as appropriate

	Question	Predicted Answer
0	What is coronavirus?	coronavirus is a type of virus that causes di...
1	What are symptoms of COVID	symptoms include fever, coughing, sore throat...
2	What is COVID?	coronaviruses are a large family of viruses w...
3	I have a cough. Do I have corona?	welfare bodies want a ban on housing eviction...
4	Does my friend have COVID?	there is no limit. even if you try to return ...
5	How many people in USA have COVID?	of course, eating too much or too little in t...
6	How many people in the world have covid19?	more than 191, 000 people worldwide have reco...

Table1. Predictions for sample custom sentences using baseline model

	Question	Predicted Answer
127	As the German chancellor said today, by the en...	at least in italy, this appears to be in part...
190	How long does the virus survive on surfaces?	diluted household bleach solutions can be use...
631	Someone I know booked a return flight on Air S...	in general, you are not freed of contractual ...
211	Apparently, a new coronavirus related paid fam...	looking on the fda website, i found this: not...
120	There's a ridiculous conspiracy theory spreadi...	scientists advising the government told them ...
2	What's the difference between physical distanc...	you might wonder whether an increased infecti...
330	UK government promises to pay 80% of wages for...	tl dr: you can, possibly, in some places, but...
586	I want to remove or replace my implant or IUD ...	all pregnant women, including those with conf...
457	Are there any projections to estimate the spre...	there are some theories that the covid- 19 ou...
350	I used to travel a lot and take photographs at...	people are laid off all the time when sales a...

Table2. Predictions for sample sentences from test set using baseline model

Before relying on any matrices, we tested using general questions which people may reasonably be expected to ask about COVID19. We also tested with random questions from our test set to find out the appropriateness of the predictions for the given questions. As we can see from the above tables, some of the responses are appropriate for the given question, while others miss the mark.

```
## Compute bleu scores
# Mean
statistics.mean(bleu_score)
```

```
0.05755938174869327
```

```
# Max
max(bleu_score)
```

```
0.7954658179355575
```

```
# Min
min(bleu_score)
```

```
5.8942399846822315e-238
```

Next, we considered numerical measures of accuracy: BLEU and ROUGE. Our BLEU score averaged around 0.058 with a range between $5.89e^{-238}$ and 0.795. This low average score may be due to the lack of overlap between the predicted and actual response.

```
{'rouge-1': {'f': 0.1850382143980261,
  'p': 0.23880430890235596,
  'r': 0.24589599278523083},
 'rouge-2': {'f': 0.03519964989094456,
  'p': 0.04415395631787685,
  'r': 0.041722112423927454},
 'rouge-l': {'f': 0.1608940067964608,
  'p': 0.19339800478750366,
  'r': 0.19546875196377592}}
```

Our ROUGE score seemed better compared to BLEU with rouge-l achieving an F-score of 0.16. Yet, it was evident that more was needed to be done to improve our model.

	Question	Predicted Answer
0	What is coronavirus?	Coronaviruses are a large family of viruses t...
1	What are symptoms of COVID	I am a PhD student in the Department of Compu...
2	What is COVID?	Coronavirus (COVID-19) is a new virus that ha...
3	I have a cough. Do I have corona?	Coronaviruses are a large family of viruses t...
4	Does my friend have COVID?	Yes, she has. \n\nHer symptoms are similar to...
5	How many people in USA have COVID?	The number of confirmed cases in the USA is c...
6	How many people in the world have covid19?	There are currently no known cases of COVID-1...

After implementing our GPT-2 model, we found an improvement in results. Although there are still incorrect responses, our answers are accurate in most cases.

```
' Coronaviruses are a large family of viruses that cause a range of respiratory diseases including colds, flu, and severe acute respiratory synd
rome (SARS-CoV-2).  \n\nThe virus that causes COVID-19 is called the "common cold", and is caused by the common cold adenovirus.  The virus caus
es about 80% of the world\'s respiratory infections, according to the World Organisation for Animal Health (OIE), and it is thought to be mainly
transmitted through respiratory droplets produced when an infected person coughs or sneezes.\n\n\n  In the United States, the most common form o
f infection is respiratory syncytial fluid (RSD), which is a fluid-like fluid that is expelled from the upper respiratory tract when a respirato
ry infection occurs. This fluid is then expelled into the air, where it can be inhaled by people who are nearby or have close contact with someo
ne who has the virus. However, there is no known way to precisely measure the amount of RSD in a person’s airways, or how much is in it, in orde
r to determine whether the person is at higher risk of developing the disease. WHO is assessing ongoing research on the ways in which people can
safely and effectively use respiratory steroids, including corticosteroids and antacids, to prevent, treat and prevent the spread of this potent
ially life-threatening illness. The most effective ways to protect yourself and others against the infection are to regularly clean your hands,
cover your cough with the bend of elbow or tissue and maintain a distance of at least 1 meter (3 feet) from people you are close to. Follow the
directions of your local health authority (WHO) on how to maintain good air quality in your home, workplace and community. If you have fever, co
ugh and difficulty breathing, seek medical attention and call in advance. You can also take\xa0proper respiratory hygiene precautions (e.g., cov
ering your mouth and nose with a tissue or sleeve when you cough, avoiding touching your eyes, nose or mouth with your bent elbow).\xa0Learn mor
```

We can also see that our GPT-2 model is able to generate exhaustive responses related to the questions.

```
# Mean
statistics.mean(bleu_score)
```

```
0.3275430683833646
```

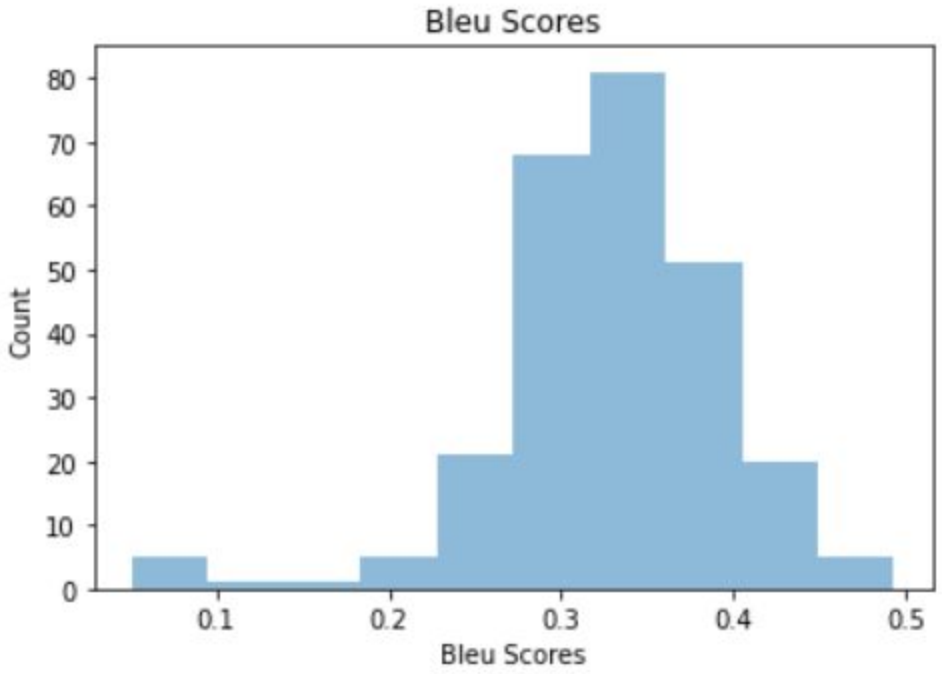
```
# Max
max(bleu_score)
```

```
0.49330550922616334
```

```
# Min
min(bleu_score)
```

```
0.050964237788241 81
```

Coming to the Bleu score, we see a stark improvement compared to our Baseline model. Our BLEU score averaged around 0.058 while our final model sits at around 0.328. We can also see the range narrow down from ($5.89e^{-238}$ - 0.795) to (0.051 to 0.493).



We also see the distribution of Bleu scores being symmetrical with the majority of the scores concentrated around the (0.3 - 0.35) range.


```
{'rouge-1': {'f': 0.18802168051490634,
  'p': 0.18812238566293757,
  'r': 0.32847448959945913},
 'rouge-2': {'f': 0.024310414606680535,
  'p': 0.02465121544290012,
  'r': 0.045447602517570136},
 'rouge-l': {'f': 0.15203930056145362,
  'p': 0.14124431247139915,
  'r': 0.25378533592157915}}
```


However, our ROUGE scores were similar to the baseline model. We suspect a lack of overlap in n-grams between the predicted and actual answers to be the culprit.

G. Summary (10-20 sentences) (3 points)

Try to draw together the Introduction, Prior Work, Approach and Results sections.

We looked at possible real world solutions that can be implemented using Natural Language Processing. We explored two use cases - weather forecasts using text generation from data and question answering. After initial analysis, we decided to use a data set comprising of question And answer pairs for COVID.
COVID QA datasets had been compiled and used to solve classification problem, categorising answers as correct and incorrect answers. We explored the SQUAD dataset as that is widely used for Question and Answer Models. However, this referenced answers in context and did not lend well to our dataset. We reviewed the use of GPT2 in language generation. We adapted this for our question and answer model on the COVID QA dataset.

For our initial model, we started with a five gram model to predict answers. We observed that the questions appear to belong to certain categories. We use k means clustering and elbow method to group the questions in clusters. We applied this combination of clustering and five gram model on Our dataset. We used rogue and bleu scores for model evaluation and found upto 80% accuracy for some questions in our test set.

As part of model enhancement, we tried a number of different techniques. We applied sequence to Sequence model with attention, but did not get good predictions. We then used two approaches using GPT2. First, we trained a new GPT2 model using our dataset and predicted results. Then we used a pre-trained GPT2 model from huggingface and fine tuned it for our dataset. We found it to be an Improvement from our baseline model.

H. References (provide any background references)

1. https://towardsdatascience.com/train-gpt-2-in-your-own-language-fc6ad4d60171
2. https://github.com/lavanyats/QuestionAnswering_From_FAQ_Tutorial/blob/master/embeddings_mlinterview_faq.ipynb
3. Language Models are Unsupervised Multitask Learners, Radford, Wu, et all (link)
4. Ehud Reiter's Blog (on GPT3 and Data to Text) https://ehudreiter.com/2020/08/10/is-gpt3-useful-for-nlg/
5. Unsupervised Text Clustering Using Natural Language Processing, Ramesh. (link)
6. How to code The Transformer in PyTorch https://blog.floydhub.com/the-transformer-in-pytorch/
7. https://blog.floydhub.com/gpt2/