Abstractive Summarization

Mandate-2

1. Problem Statement:

Given a set of tweets pertaining to a trending topic, <u>create an abstractive</u> <u>prose summary of the tweets</u>. Do not just string the tweets together to form the summary. The summary will need to paraphrase and/or say more than what is directly said in the tweets. Propose a rubric to evaluate the accuracy of your summarization.

2. Generation of Dataset:

- For corpus collection it's important to identify the relevant source and the relevance of collected content to the summary. As we have defined our problem statement over twitter tweets so we are collecting data from twitter.
- We scrape tweets of a particular hashtag.
- For that we have used an available automated platform like Apify (https://console.apify.com/) to extract the data as a CSV file of a particular hashtag or urls etc. We can also use available libraries like Python's Snscrape, BeautifulSoup, Tweepy, Scrapy etc using twitter API.
- Corpus generated -
 - 1. #rammandir 629 X 181
 - 2. #elonmusk -1000 X 251
- Some other already available datasets -

- 1. https://github.com/kavgan/opinosis-summarization (Graph algorithm based summarization framework
- 2. https://github.com/guyfe/Tweetsumm (A dataset focusing on summarization of dialogs, which represents the rich domain of Twitter customer care conversations and many more.)

3. Preprocessing of Dataset:

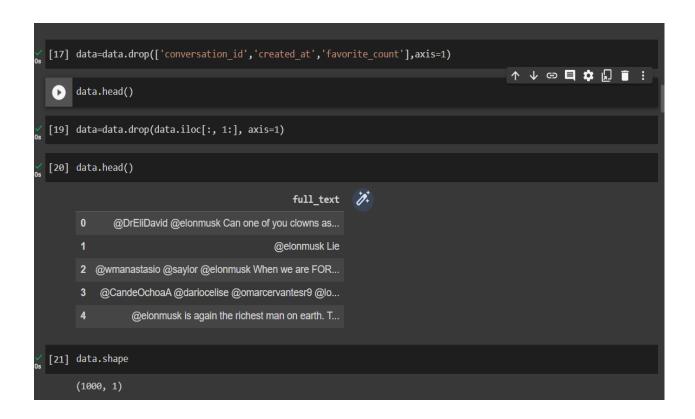
Since we have created the corpus so we need lots of data cleaning and preprocessing before applying models.

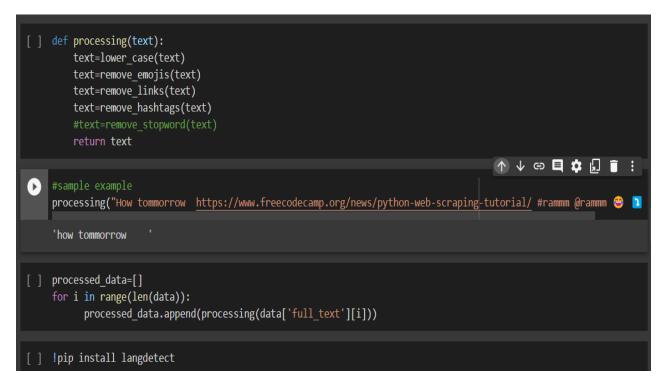
- Data Cleaning: Since our dataset contains multiple columns of fields like date, user profile, number of likes etc which are useless for our problem so we have dropped those columns and kept only one column consisting of tweets.
- Exploratory Data Analysis: To analyze the corpus.
- Preprocessing of Text : We have done following preprocessing steps
 - 1- Lower case conversion of all tweets
 - 2- Removal of all urls from tweets
 - 3- Removal of hashtags and mentions
 - 4- Removal of emojis
 - 5- Removal of tweets belonging to other language except english (unnecessary rows removal)
 - 6- Removal of punctuations

Following above steps decreases our dataset from 1000 X 251 to $686 \ \text{X}$ 1 . Only 1 column consisting of tweets and all tweets are of english language only.

Python libraries that we have used -

- Pandas
- Numpy
- Nltk
- Re
- Cleantext
- Langdetect





Lexical Processing - In NLP lexical processing refers to the process of analyzing words in a text. It is done to transform the raw , unstructured

text data into structured data which we can further analyze . For abstractive text summarization lexical preprocessing is one of the important steps .

It includes - Tokenization , Lemmatization , Stemming , Part of Speech Tagging (POS) ,Word sense disambiguation, Word Embeddings etc.

4. Tokenization-

Using Python's NLTK library we have tokenized the final cleaned and preprocessed corpus .Basically tokenization means to break sentences (tweets) into further smaller units called tokens . Here we are doing word tokenization.

Apart from NLTK we can also use Keras, Gensim to accomplish the task.

For example, the text "He is crying" can be tokenized into 'He', 'is', 'crying'.

*Part of Speech Tagging (POS) - We used the NLTK library to tag each token with the part of speech it belongs to. It is used to assign grammatical information to each word.

Ex-[('car', 'NN')] - Here NN represents a noun.

```
[63] POS_processed_data=[]
for x in f_processed_data:
    tokens = word_tokenize(x)
    pos_tags = nltk.pos_tag(tokens)
    POS_processed_data.append(pos_tags)

POS_processed_data[0]

[('can', 'MD'),
    ('one', 'CD'),
    ('of', 'IN'),
    ('you', 'PRP'),
    ('clowns', 'VBP'),
    ('ask', 'VB'),
    ('why', 'WRB'),
    ('he', 'PRP'),
    ('went', 'VBD'),
    ('to', 'TO'),
    ('epstein', 'VB'),
    ("sisland', 'NN'),
    ('sisland', 'NN'),
    ('so', 'RB'),
    ('many', 'JJ'),
    ('times', 'NNS').
```

5.Embeddings -

We are transforming words into a numerical representation as float vectors.

Word Embedding Techniques -Word2Vec, Glove Vectorization- TF-IDF, Bag of words (useful for text classification, clustering etc)

Both Word Embedding and Vectorization transforms human understandable english words into machine readable vectors but Word Embedding captures semantic relationship between the words and hence it is preferred for summarization tasks.

Python libraries that we have used -

Spacy

6. PLM -

Some famous pre trained models used for abstractive summarization are-

- 1. Transformer based models -BERT, GPT2, GPT3, PEGASUS, T5.
- 2. LSTM Based Models Seq2Seq
- 3. Attention Based Models TextRank, BERTSUM

As of now we are preferring GPT2 (Generative Pre Trained Transformer) as it has been trained on very large corpus of data and can generate natural language text. Also GPT2 is not task specific like the BERT model which is task specific.

As tweets are often written in more human friendly language instead of following proper grammatical rules hence GPT2 being capable enough to generate more human-like and coherent text.

We can also try the PEGASUS model for summarization.

7. Fine Tuning:

Fine Tuning is required to modify the weights in pre trained language to suit our context . We will tune the hyper parameters of an already trained models like GPT2 trained on large corpus etc to fit our corpus give better results.

8. Challenges faced:

- Gathering data through scraping results in lots of unnecessary data, tweets of other languages and some spam tweets too.
- For converting tweets written into languages like hindi, spanish, chinese etc in english, we need to understand the basic context of those languages and use of translation libraries is required.
- Whether we need to remove stop words or not is still in doubt as for generating a summary we need stop words too. So we can't completely remove them.
- There might be few tweets with some slang words (words more common and popular in normal speaking and has no such historical English background). We need some tools to consider them also.

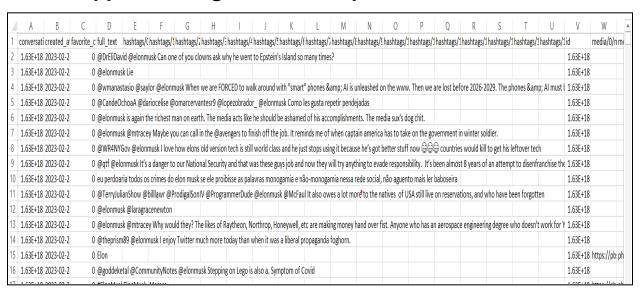
9. Conclusion:

- Understood the process of creating a dataset of tweets.
- Figured out the basic preprocessing steps.
- Used Lexical Processing as taught in class during Mandate 2.
- Explored different python libraries to solve the problem.
- Explored different pre-trained models for training our dataset.

10. References:

- https://aclanthology.org/2020.coling-main.504/
- https://pypi.org/project/twitterscraper/0.2.7/
- https://imerit.net/
- https://github.com/agarwaltanmay/text-summarizer/blob/master/Code/newqry.py
- https://www.turing.com/kb/5-powerful-text-summarization-techniques-i-
 n-python

A snippet of our generated corpus:



Before Preprocessing 1000 rows 251 columns.

| Α | В | С | D | Е | F | G | Н | 1 | J | K | L | М | N | 0 | Р | Q | R | S | T | U | V | W |
|----------------|----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|------------------------------|-----------------------------------------|---------------|---------------|----------------|-------------|---------------|---------------|---------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|--------------|-------------|---------------|---------------|
| | | 0 | | | | | | | | | | | | | | | | | | | | |
| (| can one | of you clow | ns ask why h | e went to | epstein's is | land so ma | ny times? | | | | | | | | | | | | | | | |
| | 1 when we | are forced | to walk arou | und with "s | mart" phor | nes & ai is u | ınleashed o | the www | . then we | are lost befo | ore 2026-2 | 029. the ph | ones & ai n | nust be give | n limited ca | pabilities a | nd stripped | of power of | over their o | wn being. | simply, we | nust always |
| 7 | 2 is again t | he richest n | nan on earth | . the media | a acts like h | e should b | e ashamed | of his acco | mplishmen | ts. the med | ia sux's do | g chit. | | | | | | | | | | |
| 3 | 3 maybe y | ou can call i | n the to finis | h off the jo | ob. it remin | ds me of w | hen captair | america l | nas to take | on the gove | ernment in | winter sold | ier. | | | | | | | | | |
| 4 | 4 i love ho | w elons old | version tech | is still wor | ld class and | d he just sto | ops using it l | ecause he | e's got bett | er stuff nov | v countries | would kill t | o get his le | ftover tech | | | | | | | | |
| | 5 it's a dar | ger to our n | ational secu | rity and th | at was thes | e guys job | and now the | ey will try | anything to | evade resp | onsibility. i | t's been alr | nost 8 year | s of an atte | mpt to dise | nfranchise | the americ | an people. | it's time th | e facts can | t stay their | responsibilit |
| (| it also ov | it also owes a lot more to the natives of usa still live on reservations, and who have been forgotten | | | | | | | | | | | | | | | | | | | | |
| 7 | 7 why wou | ld they? the | likes of ray | theon, nort | throp, hone | ywell, etc a | are making i | noney har | nd over fist. | . anyone wł | no has an a | erospace e | ngineering | degree who | doesn't w | ork for you | is making n | nore money | than they | ever had d | uring their i | ndoctrinati |
| 8 | 8 i enjoy twitter much more today than when it was a liberal propaganda foghorn. | | | | | | | | | | | | | | | | | | | | | |
| 9 | 9 elon | | | | | | | | | | | | | | | | | | | | | |
| 10 |) stepping | on lego is a | so a. sympto | om of covi | d | | | | | | | | | | | | | | | | | |
| 11 | 1 that's sc | ary news. | | | | | | | | | | | | | | | | | | | | |
| | there is a | o space sto | p it elon ear | th is flat | | | | | | | | | | | | | | | | | | |
| 12 | E there is i | io space sto | | | | | | | | | | | | | | | | | | | | |
| | | • | ews control | | and they w | ant black a | nd white pe | ople to no | t like each | other | | | | | | | | | | | | |
| 13 | 3 it's a rac | e thing the j | | the media | | | | | | | it. nice try | though. | | | | | | | | | | |
| 13 14 | 3 it's a rac 4 no, your | e thing the j ego made t | ews control | the media uy twitter | and laws w | ritten by a | nd for rich p | | | | it. nice try | though. | | | | | | | | | | |
| 13 14 15 | it's a rac 4 no, your 5 elon mu | e thing the j ego made t k says the u | ews control ne offer to b | the media uy twitter acist again: | and laws w | ritten by a | nd for rich p | | | | it. nice try | though. | | | | | | | | | | |

After Preprocessing 686 rows 1 column.