BIG DATA PROJECT

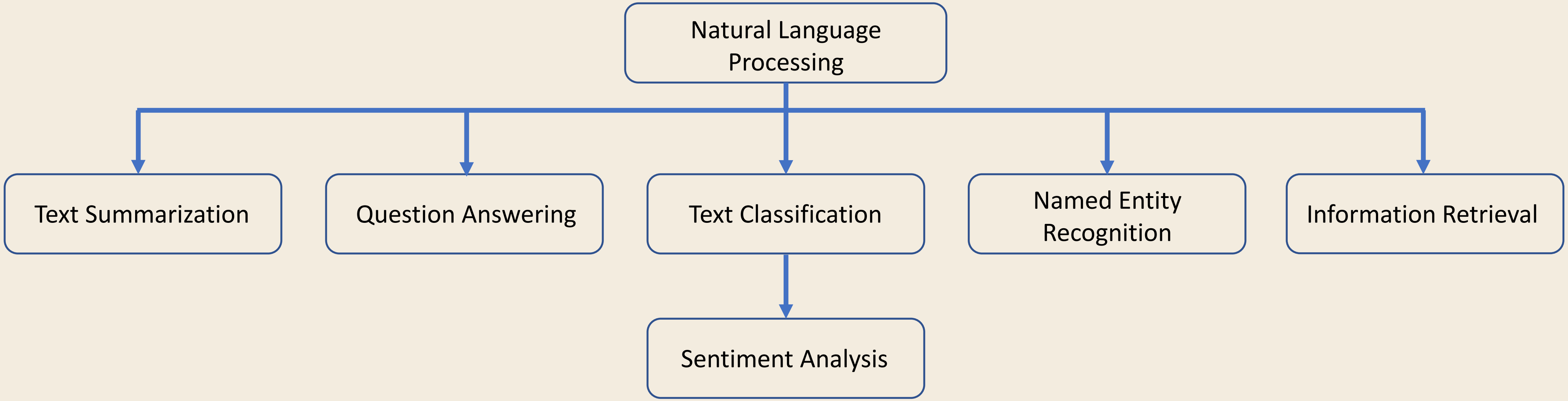
Text Summarization & Text Classification

Submitted By
Bhavay Pant (21MB0017)
Kushagra Sharma (21MB0029)
Hemant Agrawal (21MB0025)

OBJECTIVE OF THE STUDY

The main objective of the thesis is to explore various language summarization & classification approaches that are used in Natural Language Processing (NLP).

Different Domains to NLP


Natural Language Processing

Text Summarization

Question Answering

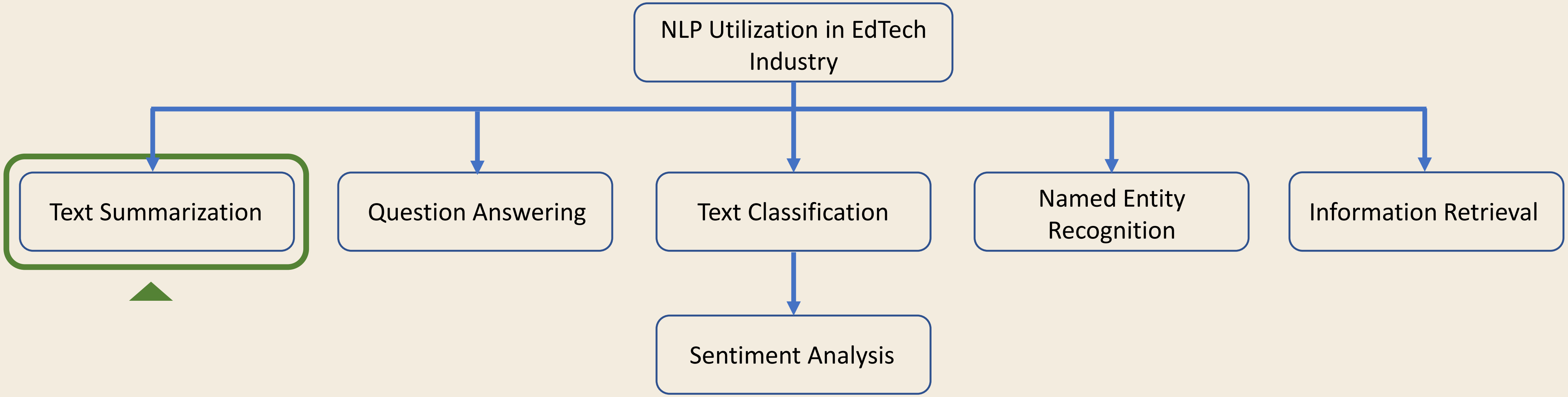
Text Classification

Named Entity Recognition

Information Retrieval

Sentiment Analysis

Different Domains to NLP



NLP Utilization in EdTech Industry

Text Summarization

Question Answering

Text Classification

Named Entity Recognition

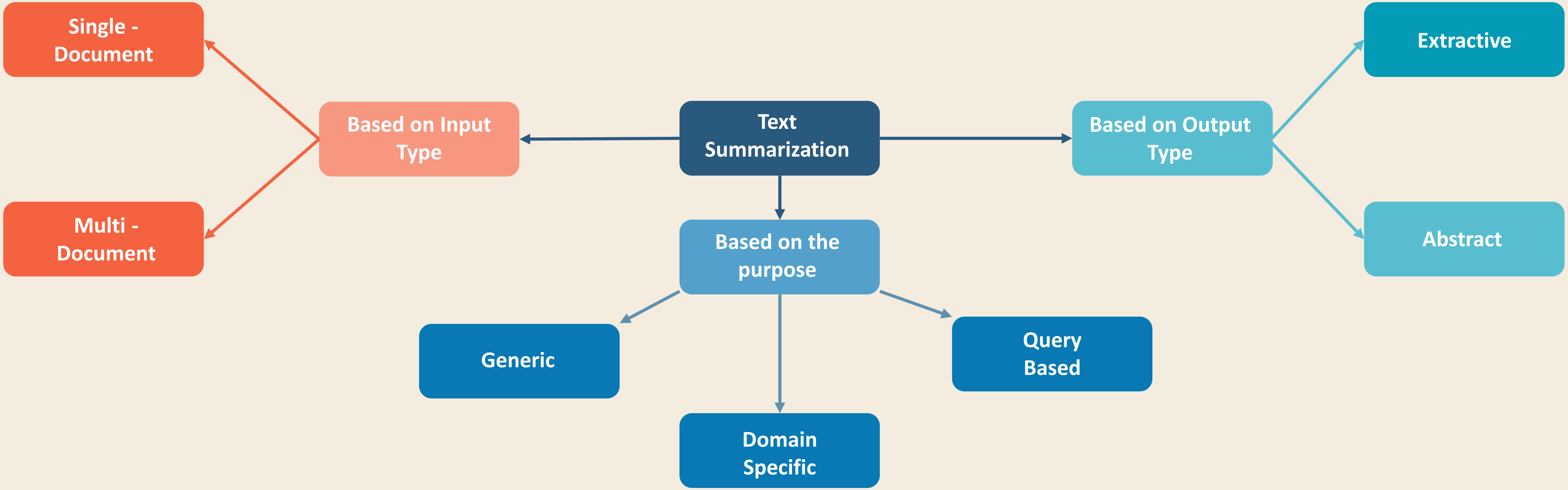
Information Retrieval

Sentiment Analysis

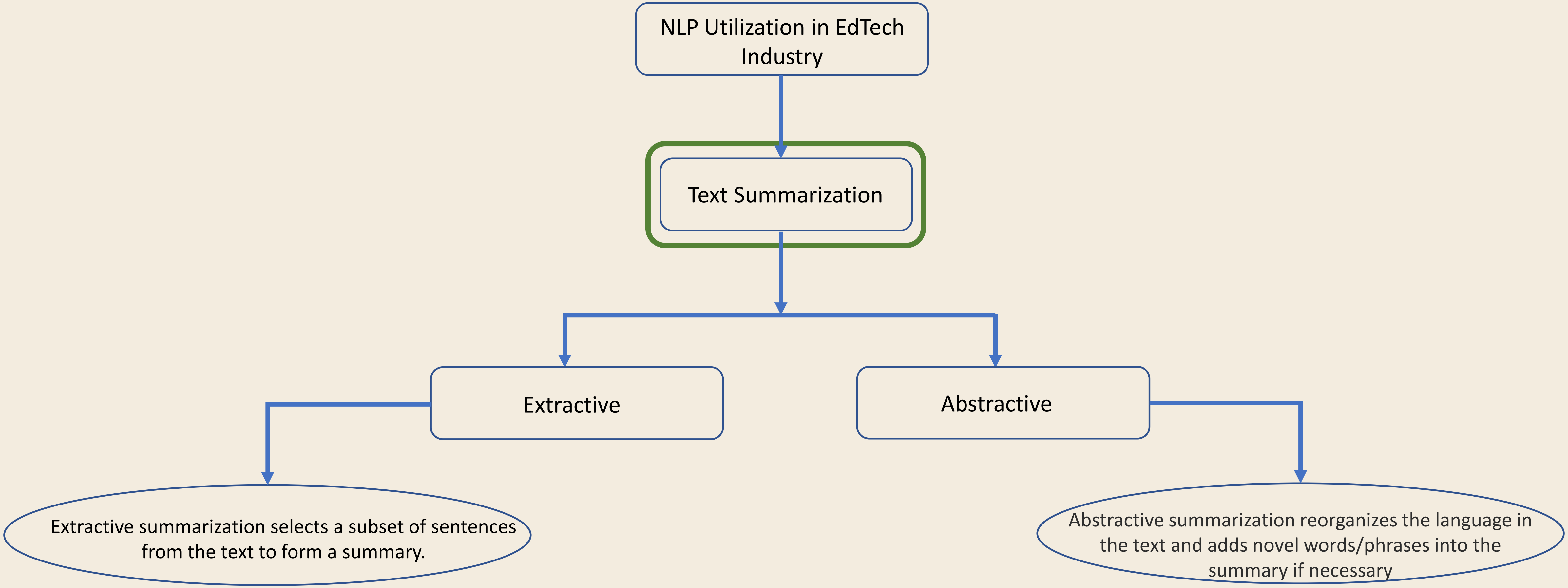
INTRODUCTION

Text summarization is the machine learning practice of breaking down long publications into manageable paragraphs or sentences. The procedure extracts important information while also ensuring that the paragraph's sense is preserved. This shortens the time it takes to comprehend long materials like research articles while without omitting critical information.

Type of summarization



Text Summarization Techniques



NLP Utilization in EdTech Industry

Text Summarization

Extractive

Abstractive

Extractive summarization selects a subset of sentences from the text to form a summary.

Abstractive summarization reorganizes the language in the text and adds novel words/phrases into the summary if necessary

LITERATURE REVIEW

Sno.	Research Paper	Data Description	Methodology
1.	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, & Comprehension	Pre-training datasets - BookCorpus - CCNeWs - WebText - Stories	**Break Throughs:** - Bidirectional in nature & autoencoder in nature. - Combines BERT as encoder & GPT as decoder. **Approach :** - Transformer-based encoder-decoder model. - Corruption of Data: Token Deletion, Token Masking, Sentence Permutation, Text Infilling - Reconstruction of corrupted (original) sentence in decode.
2.	Transformers: Attention is all you need	Pre-training datasets - WMT 2014 English-to-German translation task. - SQuAD 1.1 - SQuAD 2.0	**Break Throughs:** - Self attention mechanism for encoding long-range dependencies. - Self-supervision for leveraging large unlabelled datasets. **Approach :** - Pre-training on large unlabelled datasets. - Training for downstream-tasks on labelled data (supervised learning). a) fine-tuning approach. b) feature-based approach.
3.	BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers): Pre-training of Deep Bidirectional Transformers for Language Understanding	Pre-training datasets - BookCorpus (800M words) - Wikipedia (2500M words)	**Break Throughs:** - Masked Language Model (MLM). - Next Sentence Prediction as additional pre-training task. (Binary Classification Task) **Approach :** - Input word embeddings are the sum of the token embeddings, the segmentation embeddings & the position embeddings.

RESEARCH FORMULATION & METHODOLOGY

STEP 1 Understanding various NLP models from research papers & learn the underlying processes within the 3 target models

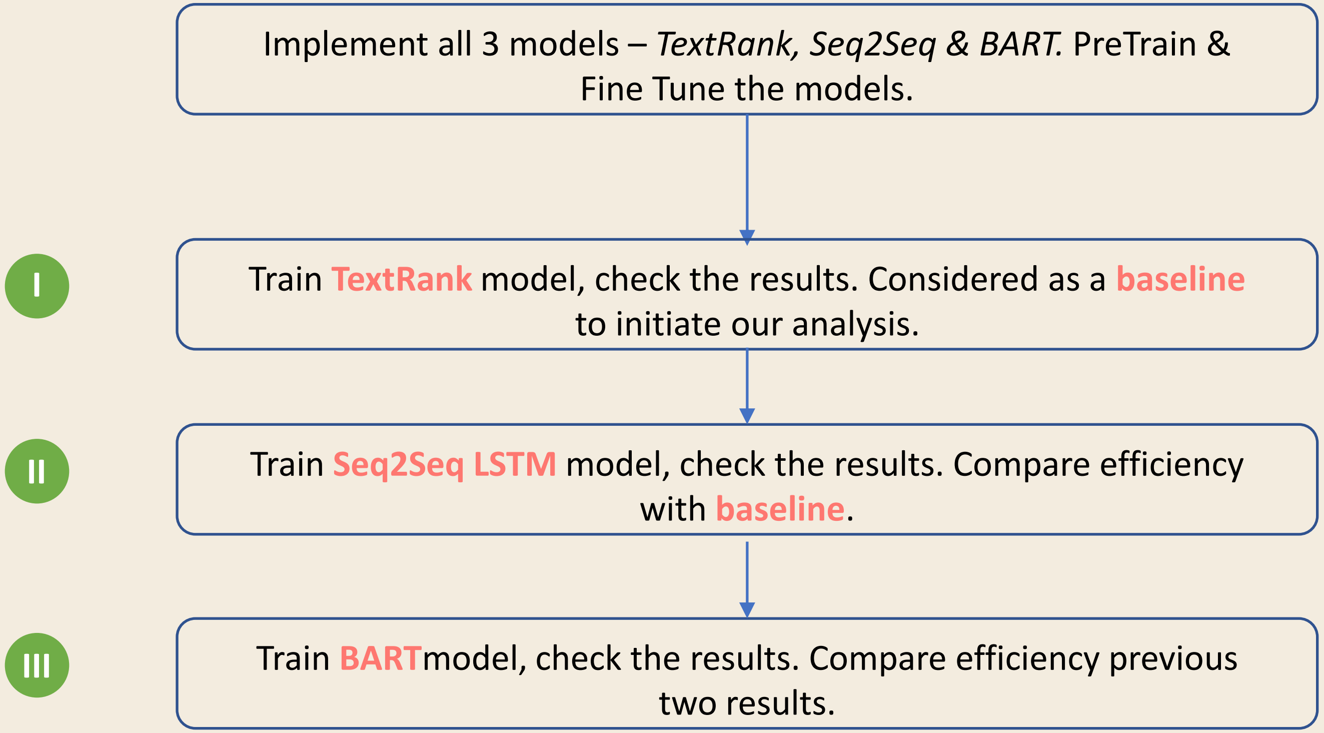
STEP 2 Identify the dataset which can capture the relative results of all three models. Also check on the pre-trained word embeddings.

STEP 3 Execute Exploratory data analysis on CNN & Dailymail dataset.

STEP 4 Implement all 3 models – *TextRank, Seq2Seq & BART.* PreTrain & Fine Tune the models.

STEP 5 Performance Analysis: Identify the right accuracy/loss metric. Recommendation on best suited model.

RESEARCH FORMULATION & METHODOLOGY



Implement all 3 models – *TextRank, Seq2Seq & BART.* PreTrain & Fine Tune the models.

I

Train **TextRank** model, check the results. Considered as a **baseline** to initiate our analysis.

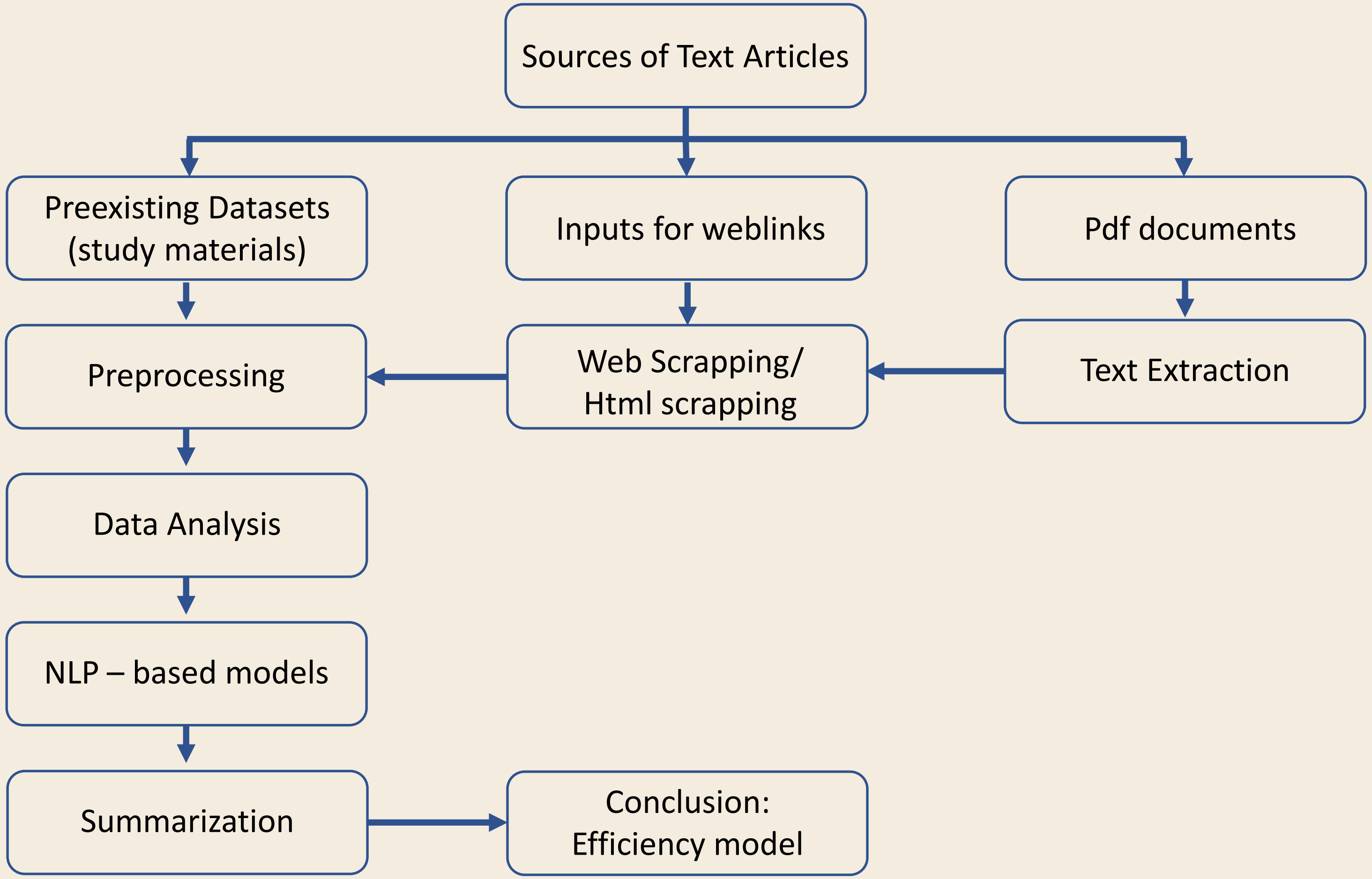
II

Train **Seq2Seq LSTM** model, check the results. Compare efficiency with **baseline**.

III

Train **BART**model, check the results. Compare efficiency previous two results.

ARCHITECTURE / FRAMEWORK



Sources of Text Articles

Preexisting Datasets (study materials)

Inputs for weblinks

Pdf documents

Preprocessing

Web Scrapping/ Html scrapping

Text Extraction

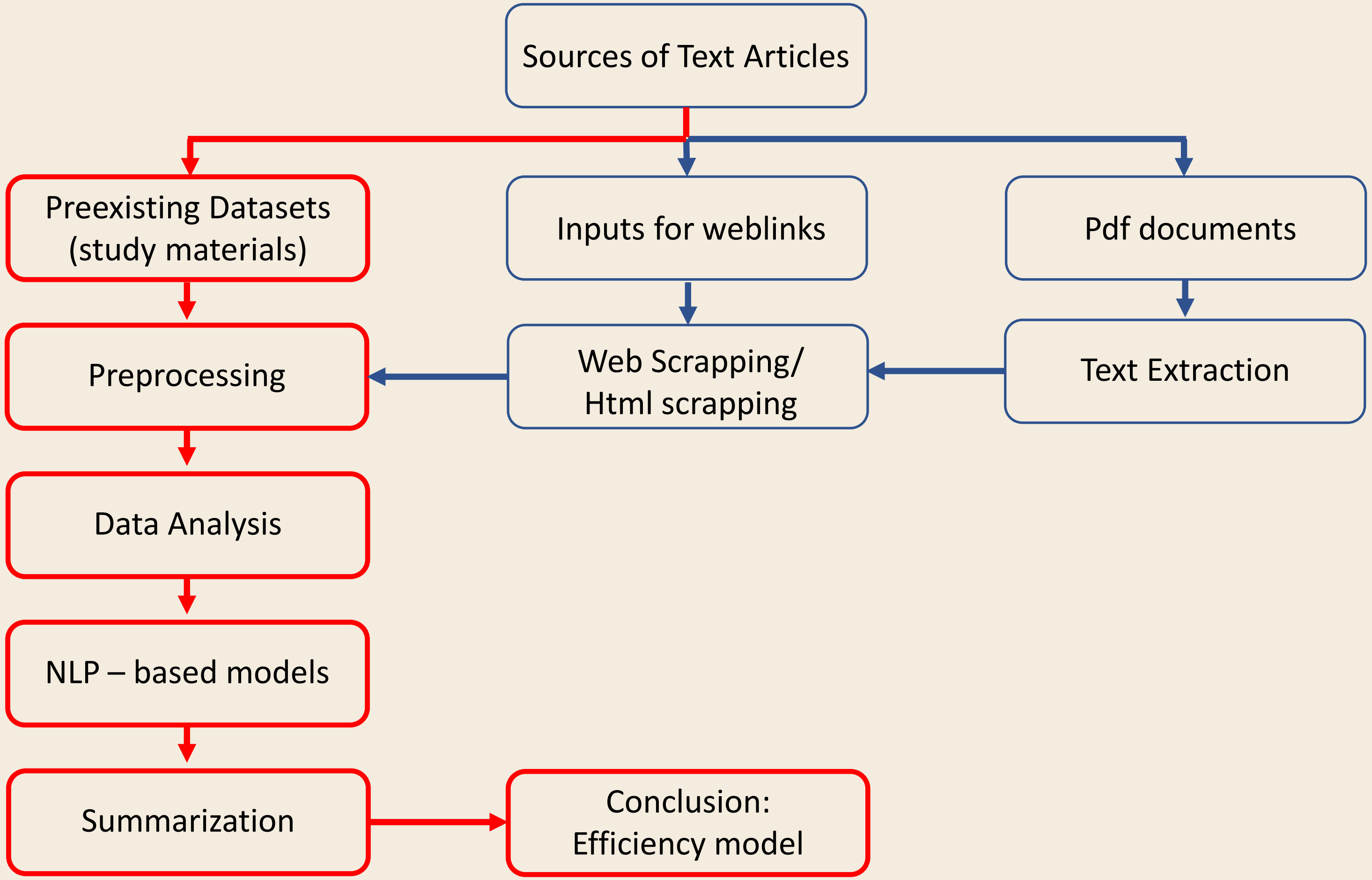
Data Analysis

NLP – based models

Summarization

Conclusion: Efficiency model

ARCHITECTURE / FRAMEWORK



Sources of Text Articles

Preexisting Datasets (study materials)

Inputs for weblinks

Pdf documents

Preprocessing

Web Scrapping/ Html scrapping

Text Extraction

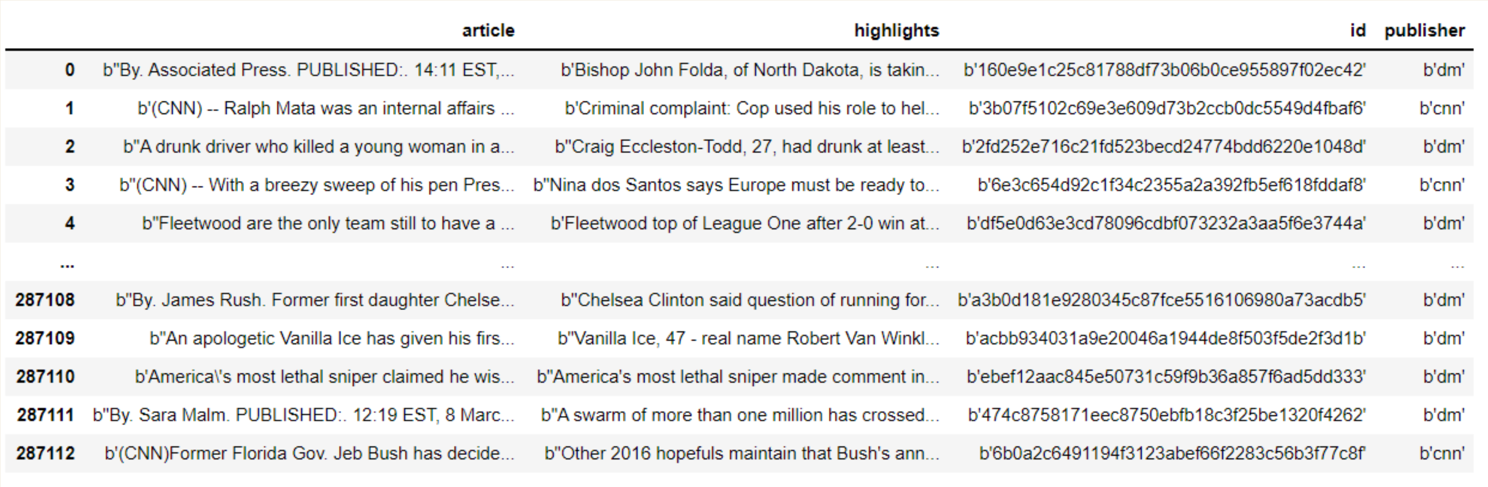
Data Analysis

NLP – based models

Summarization

Conclusion: Efficiency model

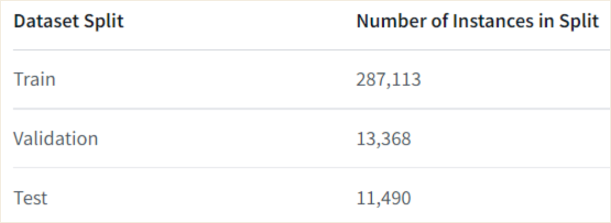
DATASET ANALYSIS



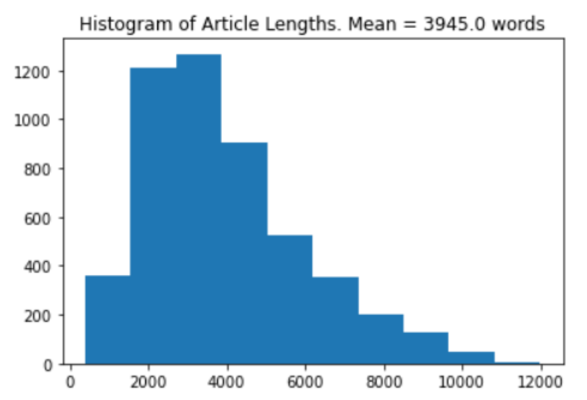
	article	highlights	id	publisher
0	b"By. Associated Press. PUBLISHED:. 14:11 EST,...	b'Bishop John Folda, of North Dakota, is takin...	b'160e9e1c25c81788df73b06b0ce955897f02ec42'	b'dm'
1	b'(CNN) -- Ralph Mata was an internal affairs ...	b'Criminal complaint: Cop used his role to hel...	b'3b07f5102c69e3e609d73b2ccb0dc5549d4fbaf6'	b'cnn'
2	b"A drunk driver who killed a young woman in a...	b"Craig Eccleston-Todd, 27, had drunk at least...	b'2fd252e716c21fd523becd24774bdd6220e1048d'	b'dm'
3	b"(CNN) -- With a breezy sweep of his pen Pres...	b"Nina dos Santos says Europe must be ready to...	b'6e3c654d92c1f34c2355a2a392fb5ef618fddaf8'	b'cnn'
4	b"Fleetwood are the only team still to have a ...	b'Fleetwood top of League One after 2-0 win at...	b'df5e0d63e3cd78096cdbf073232a3aa5f6e3744a'	b'dm'
...	...	...	...	...
287108	b"By. James Rush. Former first daughter Chelse...	b"Chelsea Clinton said question of running for...	b'a3b0d181e9280345c87fce5516106980a73acdb5'	b'dm'
287109	b"An apologetic Vanilla Ice has given his firs...	b"Vanilla Ice, 47 - real name Robert Van Winkl...	b'acbb934031a9e20046a1944de8f503f5de2f3d1b'	b'dm'
287110	b'America\'s most lethal sniper claimed he wis...	b"America's most lethal sniper made comment in...	b'ebef12aac845e50731c59f9b36a857f6ad5dd333'	b'dm'
287111	b"By. Sara Malm. PUBLISHED:. 12:19 EST, 8 Marc...	b"A swarm of more than one million has crossed...	b'474c8758171eec8750ebfb18c3f25be1320f4262'	b'dm'
287112	b'(CNN)Former Florida Gov. Jeb Bush has decide...	b"Other 2016 hopefuls maintain that Bush's ann...	b'6b0a2c6491194f3123abef66f2283c56b3f77c8f'	b'cnn'

Source link : huggingface website - https://huggingface.co/datasets/cnn_dailymail

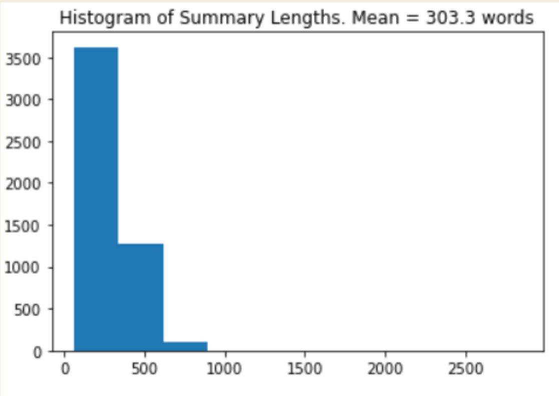
The data consists of news articles and highlight sentences. In the summarization setting, the highlight sentences are concatenated to form a summary of the article. The CNN articles were written between April 2007 and April 2015. The Daily Mail articles were written between June 2010 and April 2015.



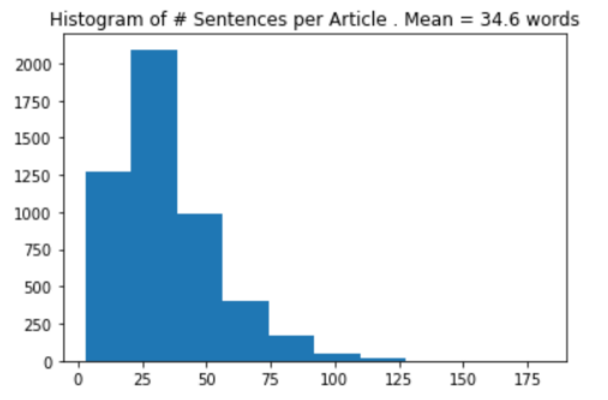
Dataset Split	Number of Instances in Split
Train	287,113
Validation	13,368
Test	11,490



Histogram of Article Lengths. Mean = 3945.0 words



Histogram of Summary Lengths. Mean = 303.3 words

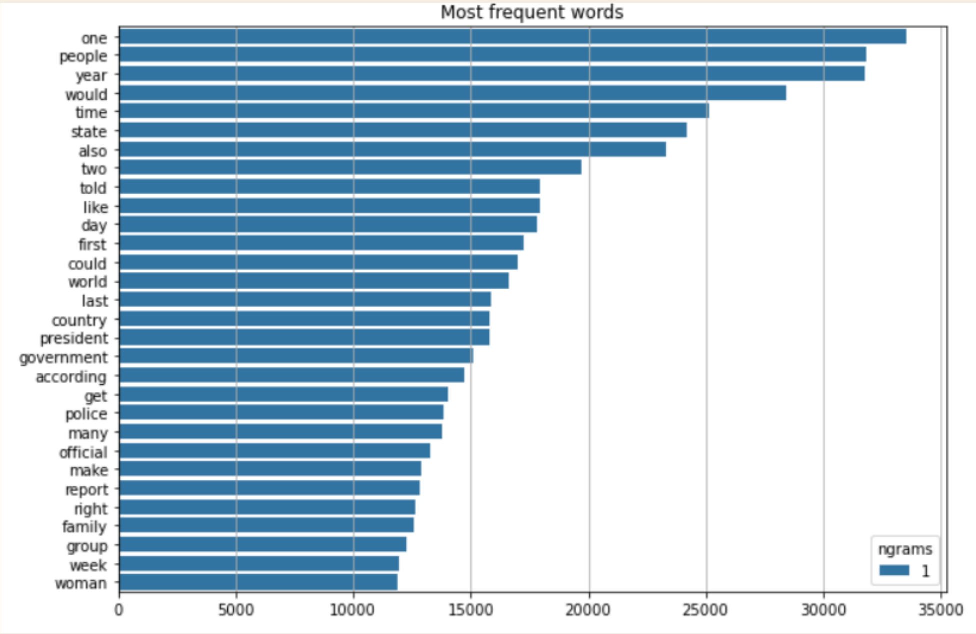


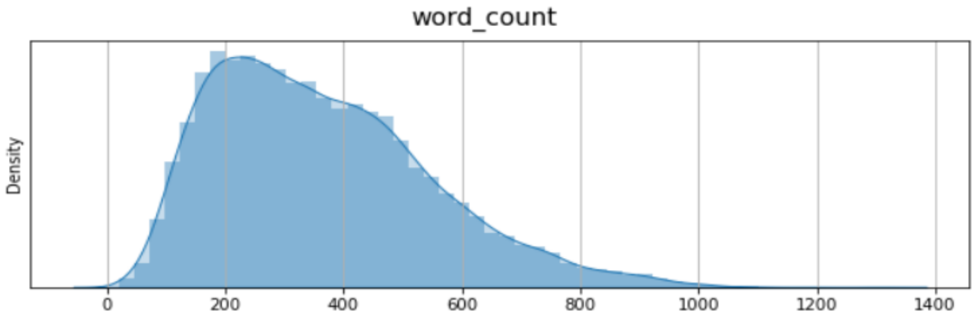
Histogram of # Sentences per Article . Mean = 34.6 words

Parameters:
Articles – text input
Highlights/headings – summary (predicted output)
Dataid – authentication token
Publisher- source.

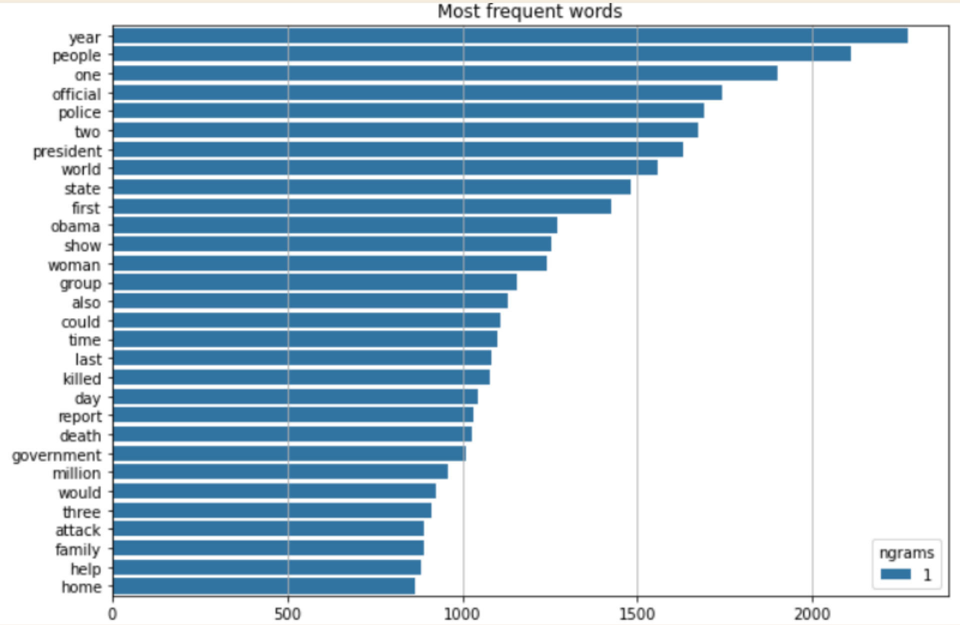
DATASET ANALYSIS

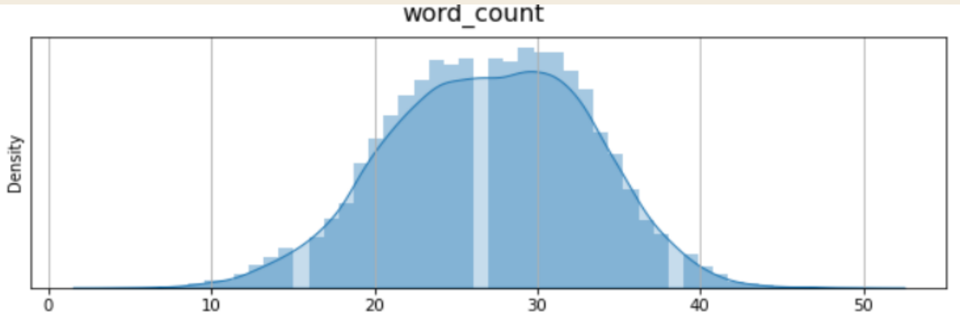
Word frequency count (articles)





Word frequency count (summaries)





Checked on word frequencies inside both the word corpus- articles (input) & summaries (output). Went on to remove those words which are repeating less than 5 times to increase computational power.

METHODS

Gensim is an open source library in python used in unsupervised topic modelling and natural language processing. It is designed to extract semantic topics from documents. Summarizing is based on ranks of text sentences using a variation of the TextRank algorithm.

NLTK is an essential library supports tasks such as classification, stemming, tagging, parsing, semantic reasoning, and tokenization in Python. It’s basically your main tool for natural language processing and machine learning.

spaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. It is designed specifically for production use. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning.

Sumy is an open-sourced Python library to extract summaries from HTML pages and text files. The package also contains an evaluation framework for text summaries. Sumy offers several algorithms and methods for text summarization such as LexRank and TextRank.

PRE PROCESSING

Deep learning models cannot use raw text directly, in order to achieve high performance from the language model we perform pre-processing of the input text to obtain text representation. We accomplish this task of processing by employing Natural Language Tool kit (NLTK).

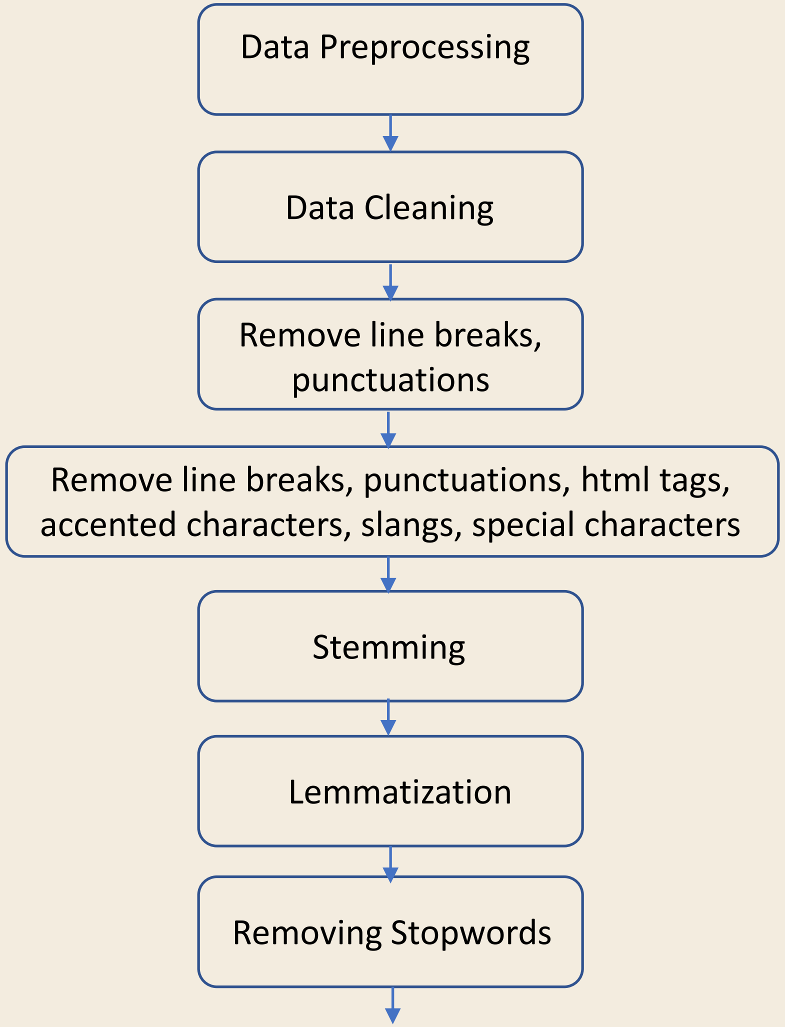
Sentence Separation It is a process of recognizing the individual sentences in a document which is used as a separation unit in summarization.

Stop words removal The process of stop-words removal eliminates the most frequent words occurring in a document like articles, prepositions, conjunctions, interrogations, help verbs, etc. The stop words are removed due to their insignificant contribution in sentence extraction process.

Part-of-Speech Tagging It is a process of identifying the part-of-speech words such as noun, adverb, verb, etc., in a sentence. However, the computational applications generally use more fine-grained POS tags like ‘nounplural’. Here, we have used the Stanford Log-linear POS tagger.

Keywords extraction In this step, we extract the keywords from a document. Here, all the words other than stop words are considered as keywords.

MODEL WORKING



Data Preprocessing

Data Cleaning

Remove line breaks, punctuations

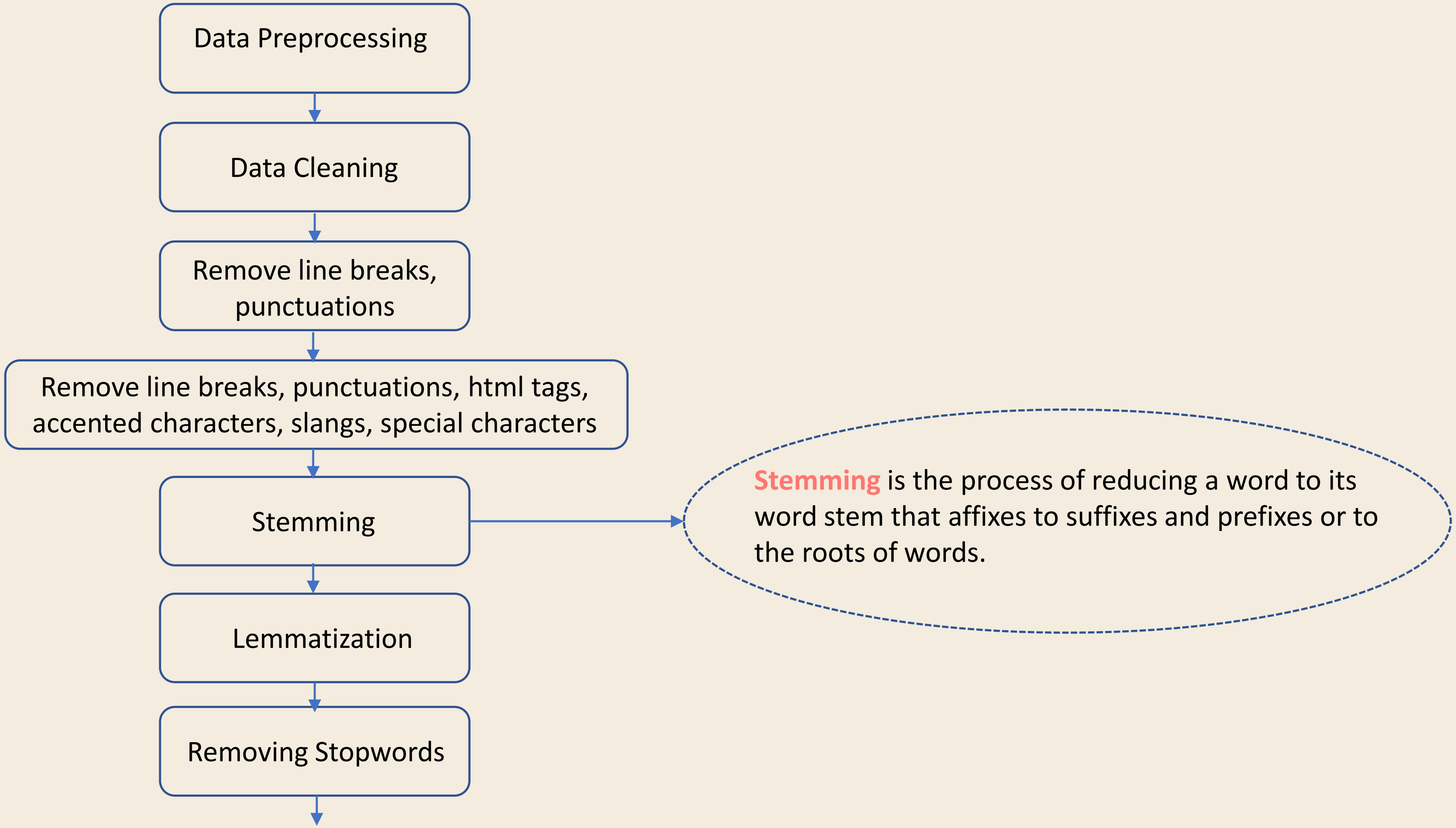
Remove line breaks, punctuations, html tags, accented characters, slangs, special characters

Stemming

Lemmatization

Removing Stopwords

MODEL WORKING (contd..)



Data Preprocessing

Data Cleaning

Remove line breaks, punctuations

Remove line breaks, punctuations, html tags, accented characters, slangs, special characters

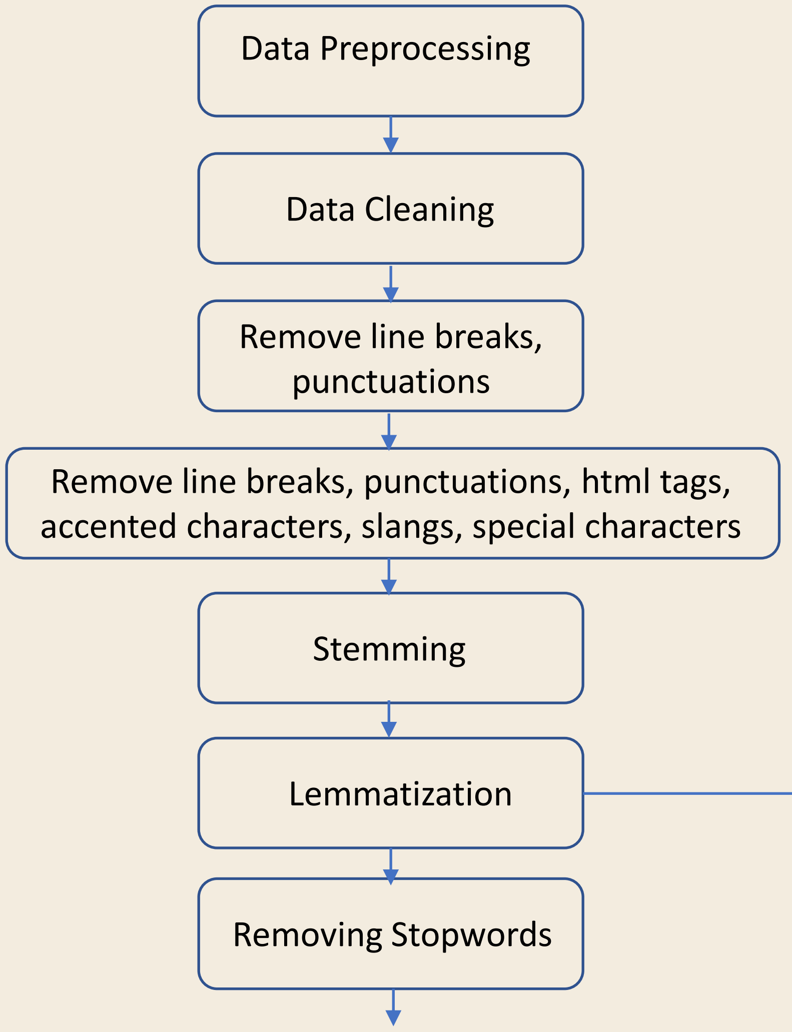
Stemming

Lemmatization

Removing Stopwords

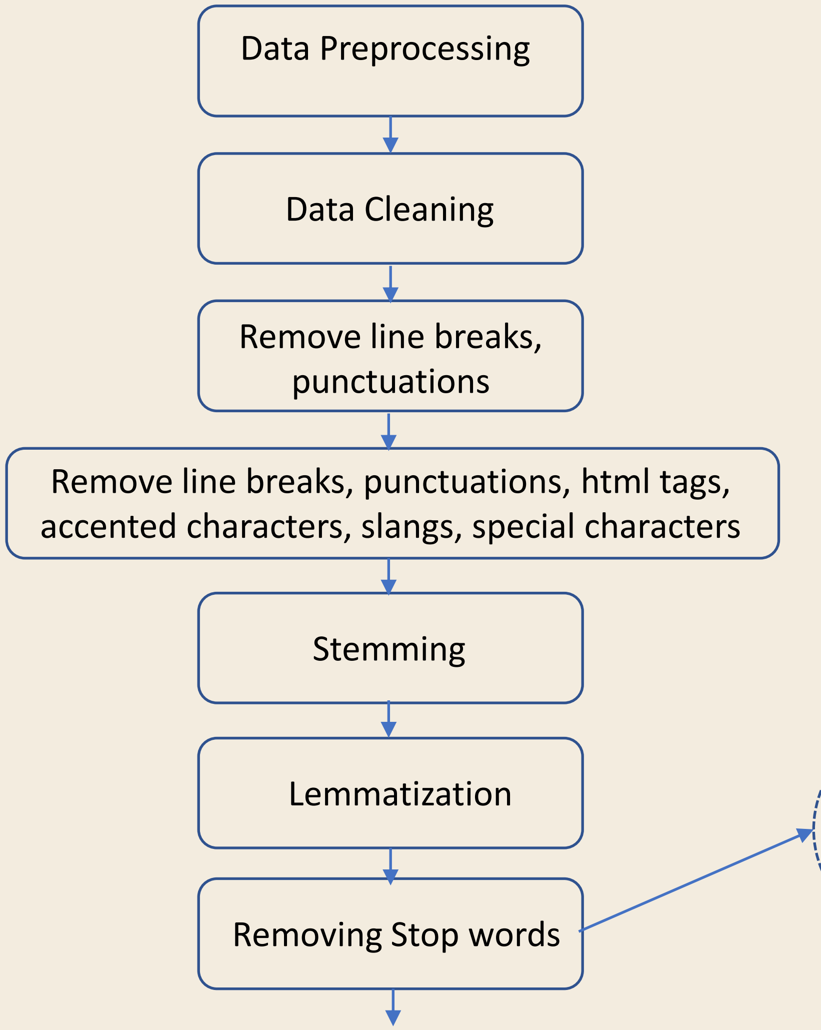
Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words.

MODEL WORKING (contd..)


Data Preprocessing
Data Cleaning
Remove line breaks, punctuations
Remove line breaks, punctuations, html tags, accented characters, slangs, special characters
Stemming
Lemmatization
Removing Stopwords

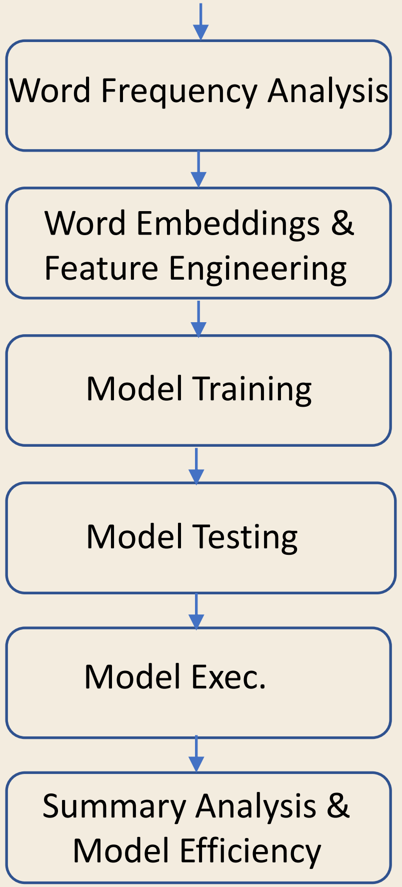
Lemmatization is a text normalization technique used in Natural Language Processing (NLP), that switches any kind of a word to its base root mode.

MODEL WORKING (contd..)


Data Preprocessing
Data Cleaning
Remove line breaks, punctuations
Remove line breaks, punctuations, html tags, accented characters, slangs, special characters
Stemming
Lemmatization
Removing Stop words

Stop words are a set of commonly used words in any language. For example, in English, “the”, “is” and “and”, would easily qualify as stop words. They are used to **eliminate unimportant words**, allowing applications to **focus on the important words instead**.

MODEL WORKING (contd..)


Word Frequency Analysis
Word Embeddings & Feature Engineering
Model Training
Model Testing
Model Exec.
Summary Analysis & Model Efficiency

MODEL TRAINING (contd..)

TextRank

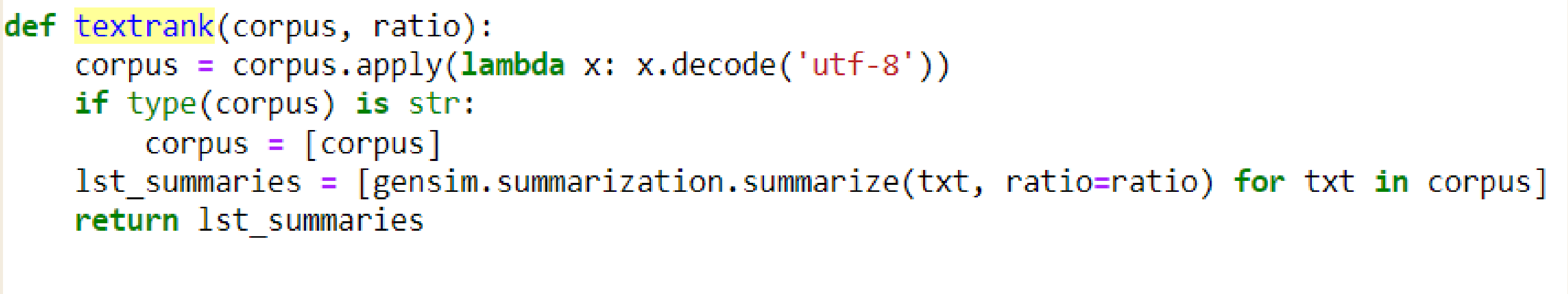
BASELINE MODEL – TEXTRANK

We used Textrank as a baseline model & to initiate our study. Textrank is a basic graph-based ranking model use in text processing. The model is based on Google's PageRank algorithm, which finds the most relevant sentences in a text. In short if Page A is taking reference from Page B, while Page in turn is linked to Page C, then the order of sorting would be Page C, Page B & finally Page A.

TextRank is unsupervised and follows below steps to build the model:

1. Split the whole text into sentences.
2. Build graphical model where sentences are the nodes and overlapping words are the links.
3. The model finally identifies the most important nodes of the network of sentences.

Used **genism** library to apply the TextRank algorithm.

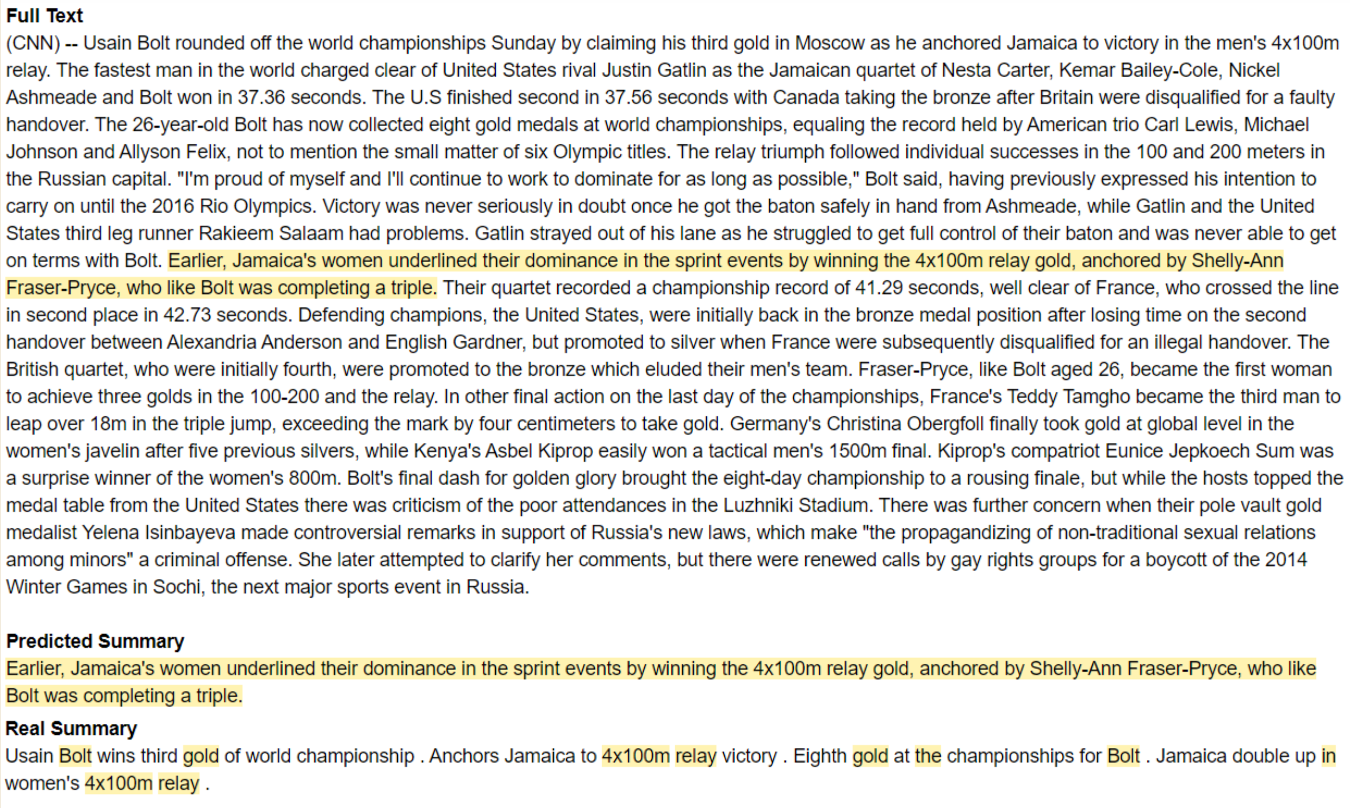


```python
def textrank(corpus, ratio):
    corpus = corpus.apply(lambda x: x.decode('utf-8'))
    if type(corpus) is str:
        corpus = [corpus]
    lst_summaries = [gensim.summarization.summarize(txt, ratio=ratio) for txt in corpus]
    return lst_summaries
```

MODEL TRAINING (contd..)

TextRank

Output :



Full Text

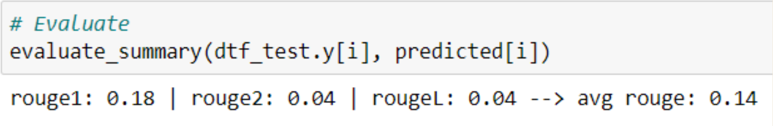
(CNN) -- Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m relay. The fastest man in the world charged clear of United States rival Justin Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel Ashmeade and Bolt won in 37.36 seconds. The U.S finished second in 37.56 seconds with Canada taking the bronze after Britain were disqualified for a faulty handover. The 26-year-old Bolt has now collected eight gold medals at world championships, equaling the record held by American trio Carl Lewis, Michael Johnson and Allyson Felix, not to mention the small matter of six Olympic titles. The relay triumph followed individual successes in the 100 and 200 meters in the Russian capital. "I'm proud of myself and I'll continue to work to dominate for as long as possible," Bolt said, having previously expressed his intention to carry on until the 2016 Rio Olympics. Victory was never seriously in doubt once he got the baton safely in hand from Ashmeade, while Gatlin and the United States third leg runner Rakieem Salaam had problems. Gatlin strayed out of his lane as he struggled to get full control of their baton and was never able to get on terms with Bolt. Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple. Their quartet recorded a championship record of 41.29 seconds, well clear of France, who crossed the line in second place in 42.73 seconds. Defending champions, the United States, were initially back in the bronze medal position after losing time on the second handover between Alexandria Anderson and English Gardner, but promoted to silver when France were subsequently disqualified for an illegal handover. The British quartet, who were initially fourth, were promoted to the bronze which eluded their men's team. Fraser-Pryce, like Bolt aged 26, became the first woman to achieve three golds in the 100-200 and the relay. In other final action on the last day of the championships, France's Teddy Tamgho became the third man to leap over 18m in the triple jump, exceeding the mark by four centimeters to take gold. Germany's Christina Obergfoll finally took gold at global level in the women's javelin after five previous silvers, while Kenya's Asbel Kiprop easily won a tactical men's 1500m final. Kiprop's compatriot Eunice Jepkoech Sum was a surprise winner of the women's 800m. Bolt's final dash for golden glory brought the eight-day championship to a rousing finale, but while the hosts topped the medal table from the United States there was criticism of the poor attendances in the Luzhniki Stadium. There was further concern when their pole vault gold medalist Yelena Isinbayeva made controversial remarks in support of Russia's new laws, which make "the propagandizing of non-traditional sexual relations among minors" a criminal offense. She later attempted to clarify her comments, but there were renewed calls by gay rights groups for a boycott of the 2014 Winter Games in Sochi, the next major sports event in Russia.

Predicted Summary

Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple.

Real Summary

Usain Bolt wins third gold of world championship . Anchors Jamaica to 4x100m relay victory . Eighth gold at the championships for Bolt . Jamaica double up in women's 4x100m relay .



```
# Evaluate
evaluate_summary(dtf_test.y[i], predicted[i])
```

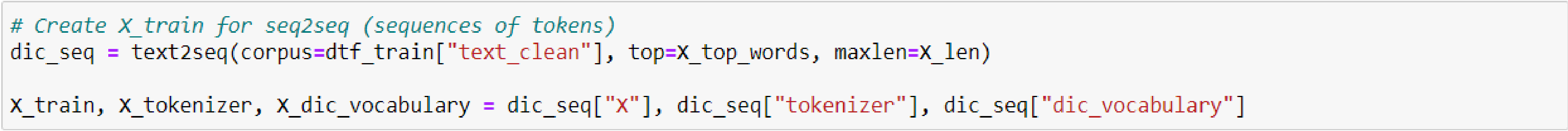
```
rouge1: 0.18 | rouge2: 0.04 | rougeL: 0.04 --> avg rouge: 0.14
```

MODEL TRAINING (contd..)

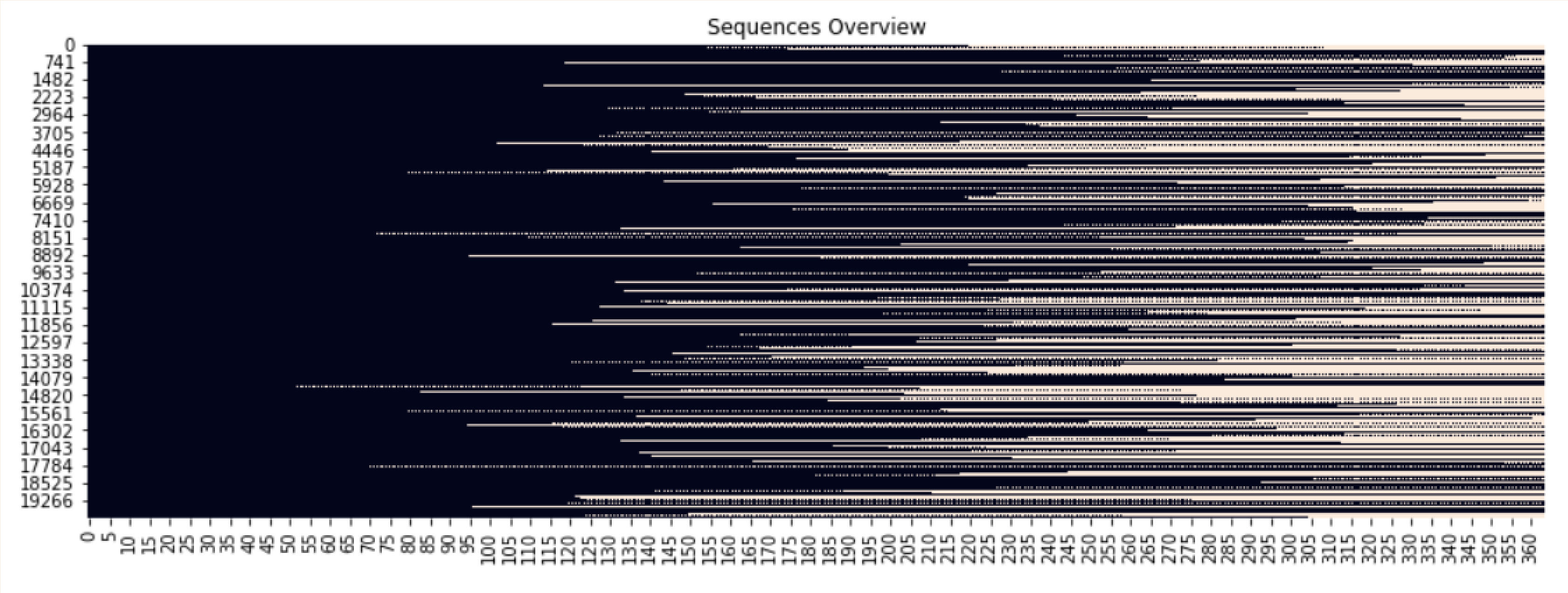
Seq2Seq – LSTM

Feature Engineering :

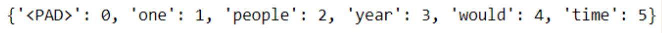
The feature matrix is created by converting the preprocessed corpus into a list of sequences.



```
# Create X_train for seq2seq (sequences of tokens)
dic_seq = text2seq(corpus=dtf_train["text_clean"], top=X_top_words, maxlen=X_len)

X_train, X_tokenizer, X_dic_vocabulary = dic_seq["X"], dic_seq["tokenizer"], dic_seq["dic_vocabulary"]
```



Input sequence created out of word corpus in articles (input) dataset.

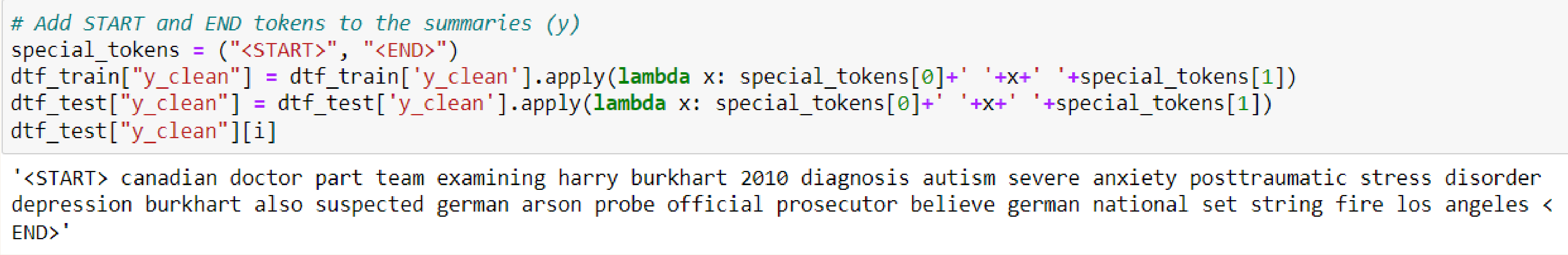


```
{'<PAD>': 0, 'one': 1, 'people': 2, 'year': 3, 'would': 4, 'time': 5}
```

MODEL TRAINING (contd..)

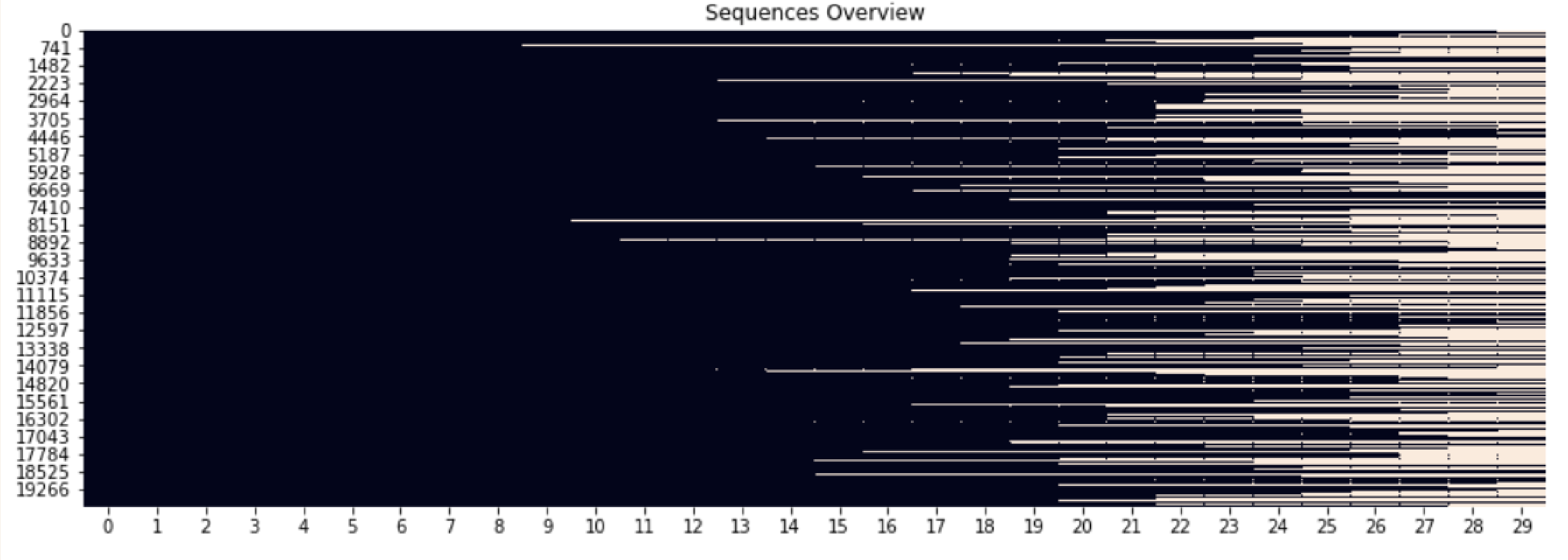
Seq2Seq – LSTM

Feature Engineering :

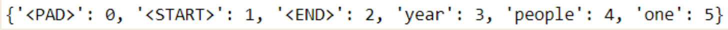


```
# Add START and END tokens to the summaries (y)
special_tokens = ("<START>", "<END>")
dtf_train["y_clean"] = dtf_train['y_clean'].apply(lambda x: special_tokens[0]+' '+x+' '+special_tokens[1])
dtf_test["y_clean"] = dtf_test['y_clean'].apply(lambda x: special_tokens[0]+' '+x+' '+special_tokens[1])
dtf_test["y_clean"][i]
```

```
'<START> canadian doctor part team examining harry burkhart 2010 diagnosis autism severe anxiety posttraumatic stress disorder
depression burkhart also suspected german arson probe official prosecutor believe german national set string fire los angeles <
END>'
```



prediction sequence created out of word corpus in headlines (target) dataset.

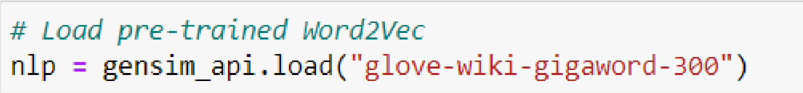


```
{'<PAD>': 0, '<START>': 1, '<END>': 2, 'year': 3, 'people': 4, 'one': 5}
```

MODEL TRAINING (contd..)

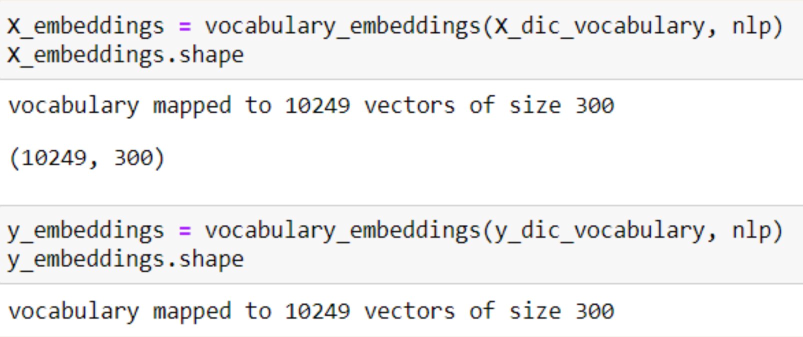
Seq2Seq – LSTM

Pre trained Word Embeddings :



```
# Load pre-trained Word2Vec
nlp = gensim_api.load("glove-wiki-gigaword-300")
```

Word Embeddings :



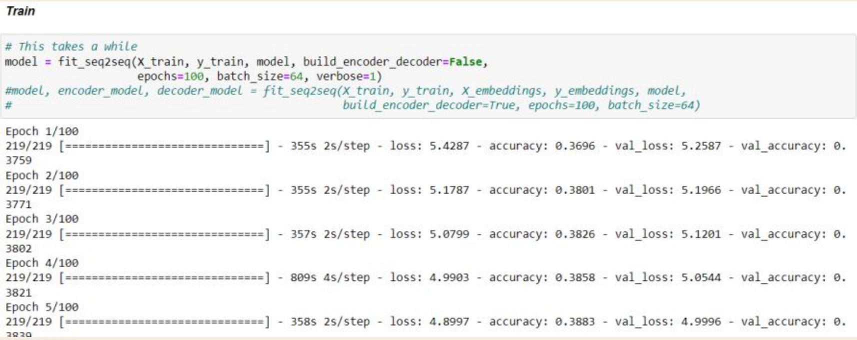
```
X_embeddings = vocabulary_embeddings(X_dic_vocabulary, nlp)
X_embeddings.shape
```

```
vocabulary mapped to 10249 vectors of size 300

(10249, 300)
```

```
y_embeddings = vocabulary_embeddings(y_dic_vocabulary, nlp)
y_embeddings.shape
```

```
vocabulary mapped to 10249 vectors of size 300
```

Model Training :

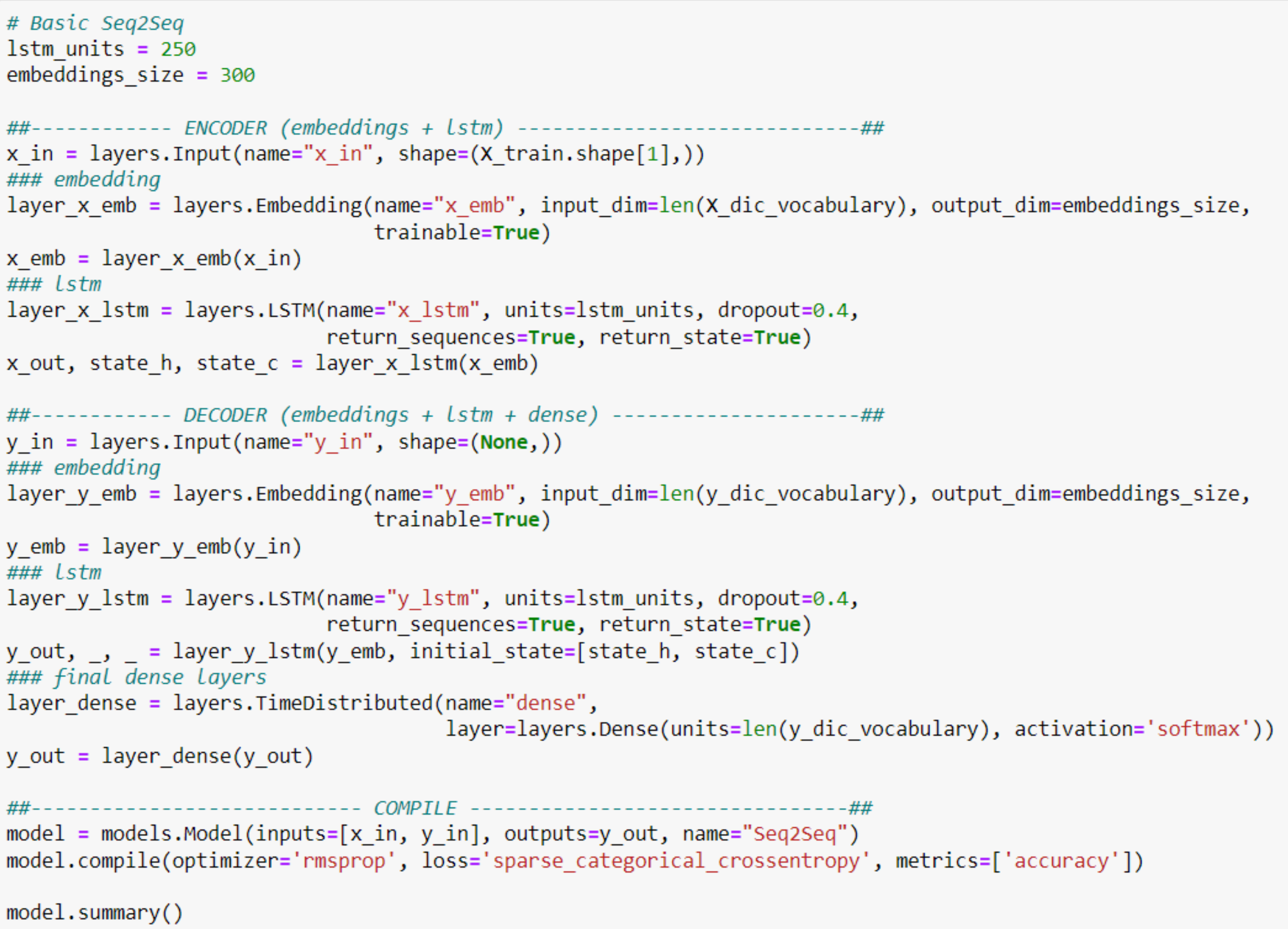


Train

```
# This takes a while
model = fit_seq2seq(X_train, y_train, model, build_encoder_decoder=False,
                    epochs=100, batch_size=64, verbose=1)
#model, encoder_model, decoder_model = fit_seq2seq(X_train, y_train, X_embeddings, y_embeddings, model,
#                                                  build_encoder_decoder=True, epochs=100, batch_size=64)
```

```
Epoch 1/100
219/219 [==============================] - 355s 2s/step - loss: 5.4287 - accuracy: 0.3696 - val_loss: 5.2587 - val_accuracy: 0.
3759
Epoch 2/100
219/219 [==============================] - 355s 2s/step - loss: 5.1787 - accuracy: 0.3801 - val_loss: 5.1966 - val_accuracy: 0.
3771
Epoch 3/100
219/219 [==============================] - 357s 2s/step - loss: 5.0799 - accuracy: 0.3826 - val_loss: 5.1201 - val_accuracy: 0.
3802
Epoch 4/100
219/219 [==============================] - 809s 4s/step - loss: 4.9903 - accuracy: 0.3858 - val_loss: 5.0544 - val_accuracy: 0.
3821
Epoch 5/100
219/219 [==============================] - 358s 2s/step - loss: 4.8997 - accuracy: 0.3883 - val_loss: 4.9996 - val_accuracy: 0.
3839
```

Model Design (Encoder) :

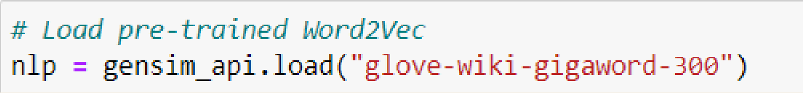


```
# Basic Seq2Seq
lstm_units = 250
embeddings_size = 300

##------------ ENCODER (embeddings + lstm) ------------------------------##
x_in = layers.Input(name="x_in", shape=(X_train.shape[1],))
### embedding
layer_x_emb = layers.Embedding(name="x_emb", input_dim=len(X_dic_vocabulary), output_dim=embeddings_size,
                               trainable=True)
x_emb = layer_x_emb(x_in)
### lstm
layer_x_lstm = layers.LSTM(name="x_lstm", units=lstm_units, dropout=0.4,
                           return_sequences=True, return_state=True)
x_out, state_h, state_c = layer_x_lstm(x_emb)

##------------ DECODER (embeddings + lstm + dense) ---------------------##
y_in = layers.Input(name="y_in", shape=(None,))
### embedding
layer_y_emb = layers.Embedding(name="y_emb", input_dim=len(y_dic_vocabulary), output_dim=embeddings_size,
                               trainable=True)
y_emb = layer_y_emb(y_in)
### lstm
layer_y_lstm = layers.LSTM(name="y_lstm", units=lstm_units, dropout=0.4,
                           return_sequences=True, return_state=True)
y_out, _, _ = layer_y_lstm(y_emb, initial_state=[state_h, state_c])
### final dense layers
layer_dense = layers.TimeDistributed(name="dense",
                                     layer=layers.Dense(units=len(y_dic_vocabulary), activation='softmax'))
y_out = layer_dense(y_out)

##--------------------------- COMPILE --------------------------------##
model = models.Model(inputs=[x_in, y_in], outputs=y_out, name="Seq2Seq")
model.compile(optimizer='rmsprop', loss='sparse_categorical_crossentropy', metrics=['accuracy'])

model.summary()
```

MODEL TRAINING (contd..)

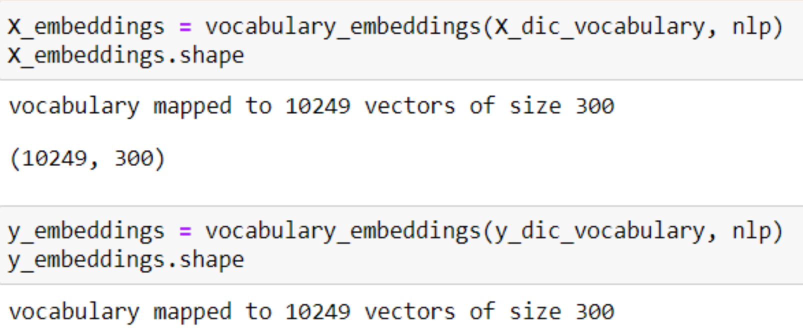
Seq2Seq – LSTM

Pre trained Word Embeddings :



```
# Load pre-trained Word2Vec
nlp = gensim_api.load("glove-wiki-gigaword-300")
```

Word Embeddings :



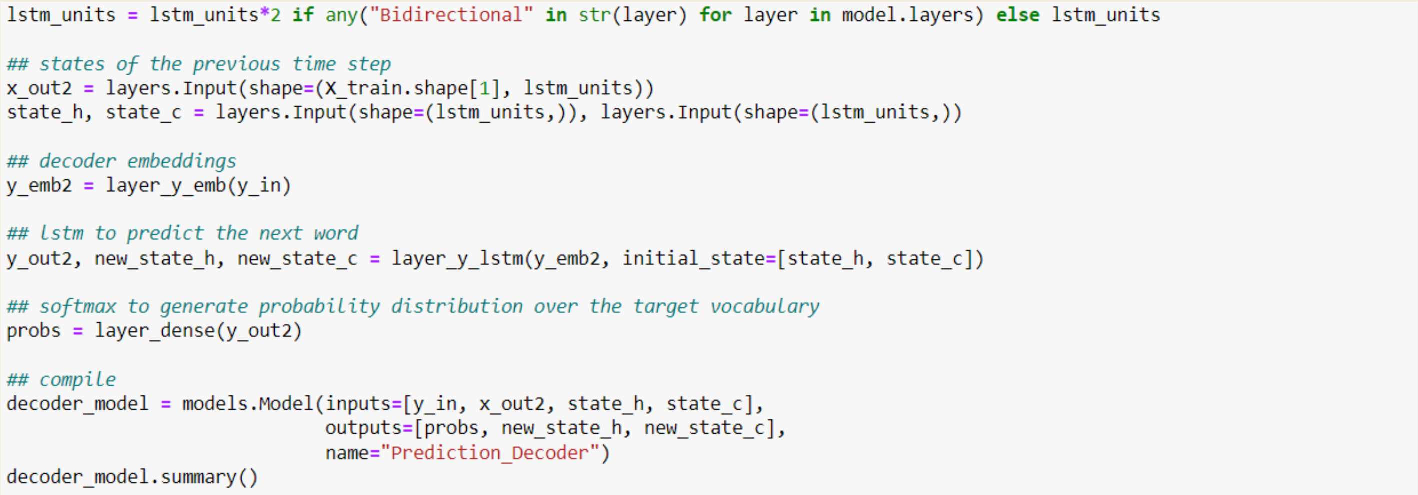
```
X_embeddings = vocabulary_embeddings(X_dic_vocabulary, nlp)
X_embeddings.shape
```

```
vocabulary mapped to 10249 vectors of size 300

(10249, 300)
```

```
y_embeddings = vocabulary_embeddings(y_dic_vocabulary, nlp)
y_embeddings.shape
```

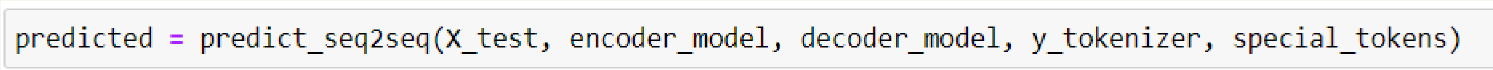
```
vocabulary mapped to 10249 vectors of size 300
```

Model Design (Decoder) :



```
lstm_units = lstm_units*2 if any("Bidirectional" in str(layer) for layer in model.layers) else lstm_units

## states of the previous time step
x_out2 = layers.Input(shape=(X_train.shape[1], lstm_units))
state_h, state_c = layers.Input(shape=(lstm_units,)), layers.Input(shape=(lstm_units,))

## decoder embeddings
y_emb2 = layer_y_emb(y_in)

## lstm to predict the next word
y_out2, new_state_h, new_state_c = layer_y_lstm(y_emb2, initial_state=[state_h, state_c])

## softmax to generate probability distribution over the target vocabulary
probs = layer_dense(y_out2)

## compile
decoder_model = models.Model(inputs=[y_in, x_out2, state_h, state_c],
                             outputs=[probs, new_state_h, new_state_c],
                             name="Prediction_Decoder")
decoder_model.summary()
```

Model Execution :

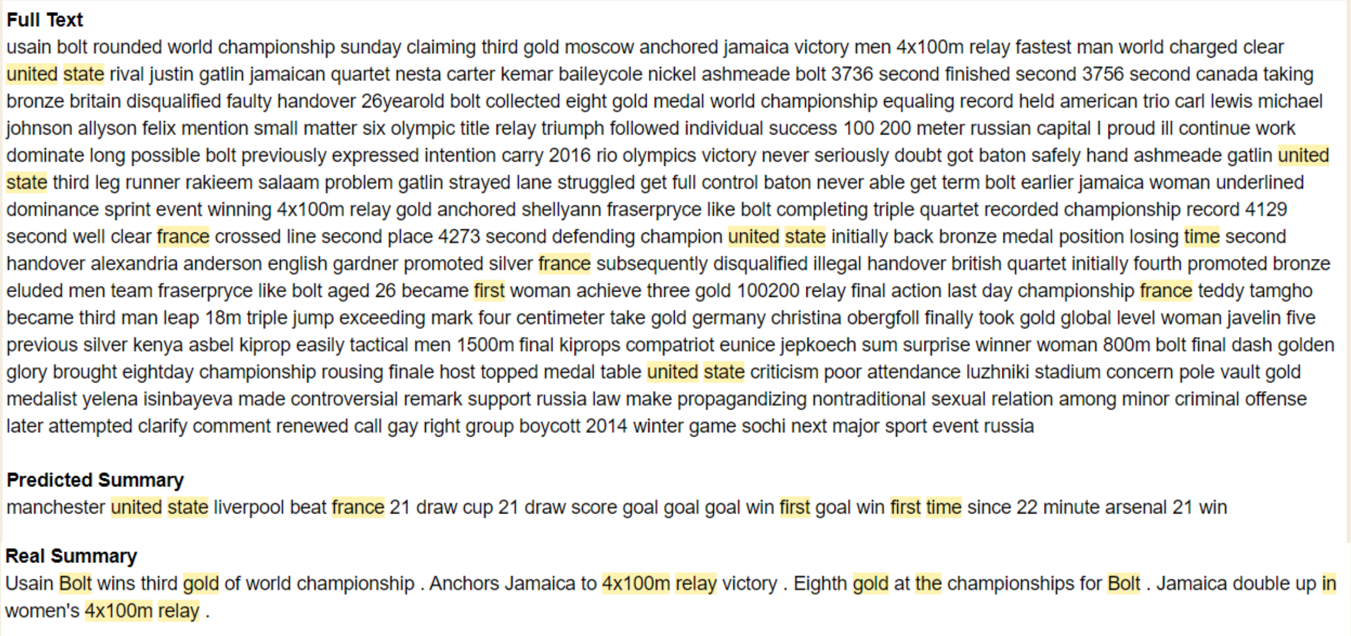


```
predicted = predict_seq2seq(X_test, encoder_model, decoder_model, y_tokenizer, special_tokens)
```

MODEL TRAINING (contd..)

Seq2Seq – LSTM

Output :



Full Text

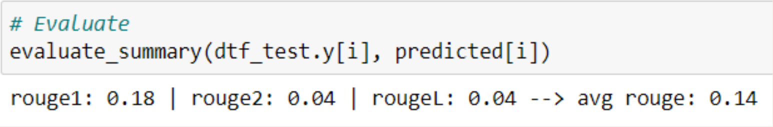
usain bolt rounded world championship sunday claiming third gold moscow anchored jamaica victory men 4x100m relay fastest man world charged clear united state rival justin gatlin jamaican quartet nesta carter kemar baileycole nickel ashmeade bolt 3736 second finished second 3756 second canada taking bronze britain disqualified faulty handover 26yearold bolt collected eight gold medal world championship equaling record held american trio carl lewis michael johnson allyson felix mention small matter six olympic title relay triumph followed individual success 100 200 meter russian capital I proud ill continue work dominate long possible bolt previously expressed intention carry 2016 rio olympics victory never seriously doubt got baton safely hand ashmeade gatlin united state third leg runner rakieem salaam problem gatlin strayed lane struggled get full control baton never able get term bolt earlier jamaica woman underlined dominance sprint event winning 4x100m relay gold anchored shellyann fraserpryce like bolt completing triple quartet recorded championship record 4129 second well clear france crossed line second place 4273 second defending champion united state initially back bronze medal position losing time second handover alexandria anderson english gardner promoted silver france subsequently disqualified illegal handover british quartet initially fourth promoted bronze eluded men team fraserpryce like bolt aged 26 became first woman achieve three gold 100200 relay final action last day championship france teddy tamgho became third man leap 18m triple jump exceeding mark four centimeter take gold germany christina obergfoll finally took gold global level woman javelin five previous silver kenya asbel kiprop easily tactical men 1500m final kiprops compatriot eunice jepkoech sum surprise winner woman 800m bolt final dash golden glory brought eightday championship rousing finale host topped medal table united state criticism poor attendance luzhniki stadium concern pole vault gold medalist yelena isinbayeva made controversial remark support russia law make propagandizing nontraditional sexual relation among minor criminal offense later attempted clarify comment renewed call gay right group boycott 2014 winter game sochi next major sport event russia

Predicted Summary

manchester united state liverpool beat france 21 draw cup 21 draw score goal goal goal win first goal win first time since 22 minute arsenal 21 win

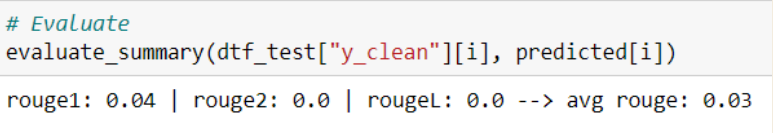
Real Summary

Usain Bolt wins third gold of world championship . Anchors Jamaica to 4x100m relay victory . Eighth gold at the championships for Bolt . Jamaica double up in women's 4x100m relay .



```
# Evaluate
evaluate_summary(dtf_test.y[i], predicted[i])
```

```
rouge1: 0.18 | rouge2: 0.04 | rougeL: 0.04 --> avg rouge: 0.14
```



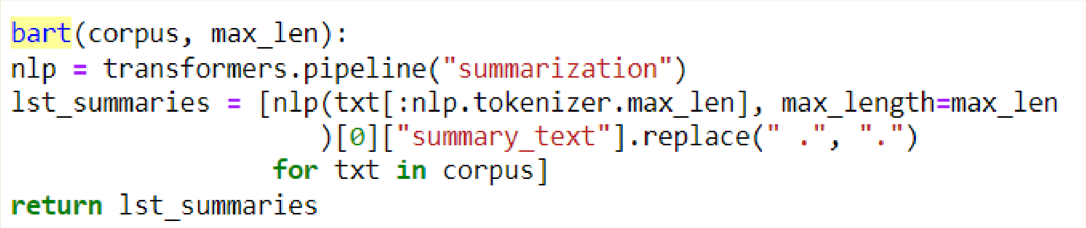
```
# Evaluate
evaluate_summary(dtf_test["y_clean"][i], predicted[i])
```

```
rouge1: 0.04 | rouge2: 0.0 | rougeL: 0.0 --> avg rouge: 0.03
```

MODEL TRAINING (contd..)

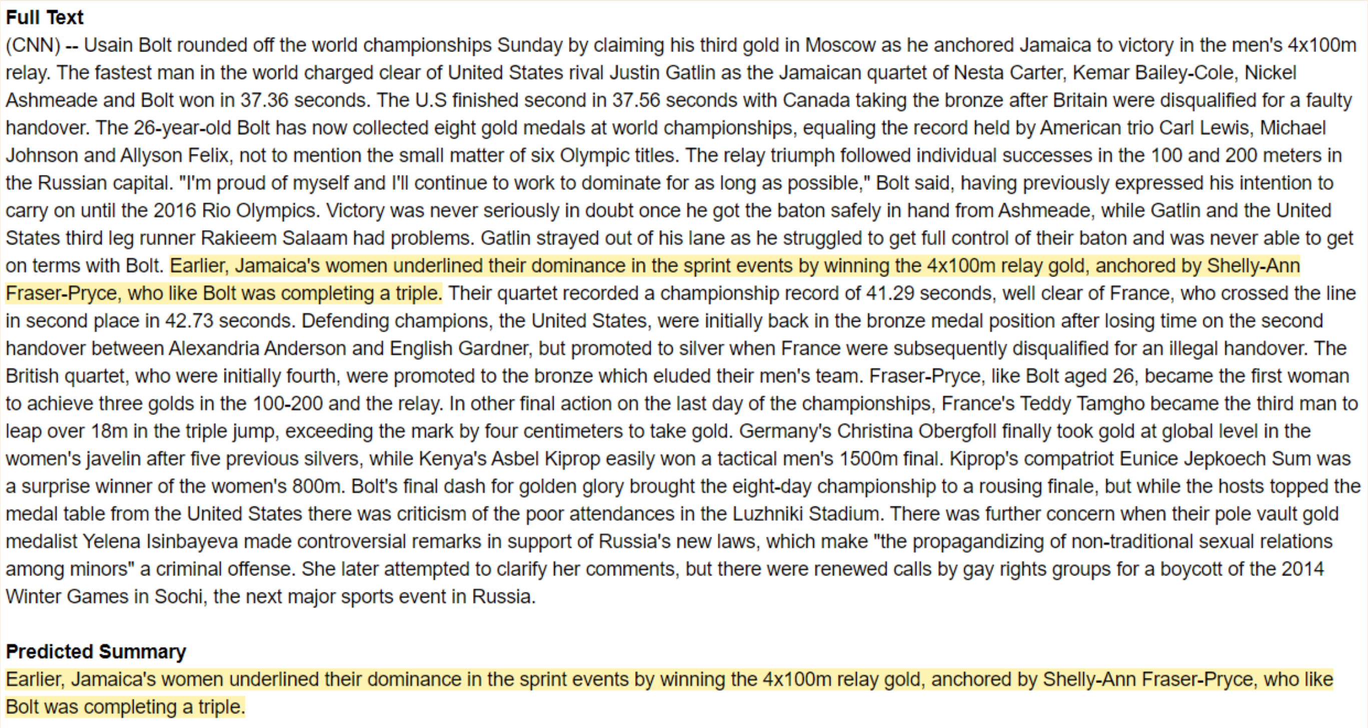
BART- Bidirectional Autoregressive Transformer

Model Execution :



```
bart(corpus, max_len):
nlp = transformers.pipeline("summarization")
lst_summaries = [nlp(txt[:nlp.tokenizer.max_len], max_length=max_len
                    )[0]["summary_text"].replace(" .", ".")
                 for txt in corpus]
return lst_summaries
```

Output :



Full Text

(CNN) -- Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m relay. The fastest man in the world charged clear of United States rival Justin Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel Ashmeade and Bolt won in 37.36 seconds. The U.S finished second in 37.56 seconds with Canada taking the bronze after Britain were disqualified for a faulty handover. The 26-year-old Bolt has now collected eight gold medals at world championships, equaling the record held by American trio Carl Lewis, Michael Johnson and Allyson Felix, not to mention the small matter of six Olympic titles. The relay triumph followed individual successes in the 100 and 200 meters in the Russian capital. "I'm proud of myself and I'll continue to work to dominate for as long as possible," Bolt said, having previously expressed his intention to carry on until the 2016 Rio Olympics. Victory was never seriously in doubt once he got the baton safely in hand from Ashmeade, while Gatlin and the United States third leg runner Rakieem Salaam had problems. Gatlin strayed out of his lane as he struggled to get full control of their baton and was never able to get on terms with Bolt. Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple. Their quartet recorded a championship record of 41.29 seconds, well clear of France, who crossed the line in second place in 42.73 seconds. Defending champions, the United States, were initially back in the bronze medal position after losing time on the second handover between Alexandria Anderson and English Gardner, but promoted to silver when France were subsequently disqualified for an illegal handover. The British quartet, who were initially fourth, were promoted to the bronze which eluded their men's team. Fraser-Pryce, like Bolt aged 26, became the first woman to achieve three golds in the 100-200 and the relay. In other final action on the last day of the championships, France's Teddy Tamgho became the third man to leap over 18m in the triple jump, exceeding the mark by four centimeters to take gold. Germany's Christina Obergfoll finally took gold at global level in the women's javelin after five previous silvers, while Kenya's Asbel Kiprop easily won a tactical men's 1500m final. Kiprop's compatriot Eunice Jepkoech Sum was a surprise winner of the women's 800m. Bolt's final dash for golden glory brought the eight-day championship to a rousing finale, but while the hosts topped the medal table from the United States there was criticism of the poor attendances in the Luzhniki Stadium. There was further concern when their pole vault gold medalist Yelena Isinbayeva made controversial remarks in support of Russia's new laws, which make "the propagandizing of non-traditional sexual relations among minors" a criminal offense. She later attempted to clarify her comments, but there were renewed calls by gay rights groups for a boycott of the 2014 Winter Games in Sochi, the next major sports event in Russia.

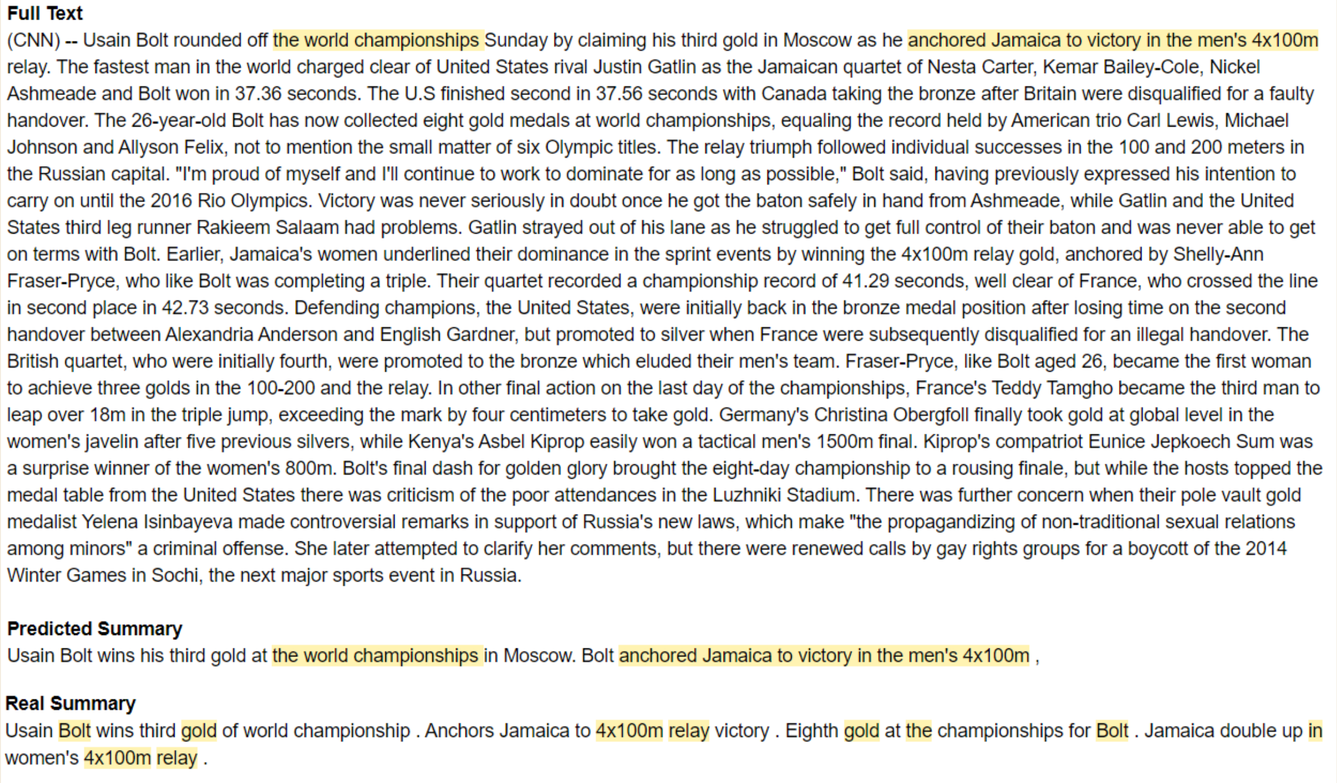
Predicted Summary

Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple.

MODEL TRAINING (contd..)

BART- Bidirectional Autoregressive Transformer

Output :



Full Text

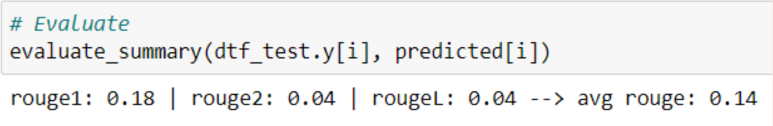
(CNN) -- Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m relay. The fastest man in the world charged clear of United States rival Justin Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel Ashmeade and Bolt won in 37.36 seconds. The U.S finished second in 37.56 seconds with Canada taking the bronze after Britain were disqualified for a faulty handover. The 26-year-old Bolt has now collected eight gold medals at world championships, equaling the record held by American trio Carl Lewis, Michael Johnson and Allyson Felix, not to mention the small matter of six Olympic titles. The relay triumph followed individual successes in the 100 and 200 meters in the Russian capital. "I'm proud of myself and I'll continue to work to dominate for as long as possible," Bolt said, having previously expressed his intention to carry on until the 2016 Rio Olympics. Victory was never seriously in doubt once he got the baton safely in hand from Ashmeade, while Gatlin and the United States third leg runner Rakieem Salaam had problems. Gatlin strayed out of his lane as he struggled to get full control of their baton and was never able to get on terms with Bolt. Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple. Their quartet recorded a championship record of 41.29 seconds, well clear of France, who crossed the line in second place in 42.73 seconds. Defending champions, the United States, were initially back in the bronze medal position after losing time on the second handover between Alexandria Anderson and English Gardner, but promoted to silver when France were subsequently disqualified for an illegal handover. The British quartet, who were initially fourth, were promoted to the bronze which eluded their men's team. Fraser-Pryce, like Bolt aged 26, became the first woman to achieve three golds in the 100-200 and the relay. In other final action on the last day of the championships, France's Teddy Tamgho became the third man to leap over 18m in the triple jump, exceeding the mark by four centimeters to take gold. Germany's Christina Obergfoll finally took gold at global level in the women's javelin after five previous silvers, while Kenya's Asbel Kiprop easily won a tactical men's 1500m final. Kiprop's compatriot Eunice Jepkoech Sum was a surprise winner of the women's 800m. Bolt's final dash for golden glory brought the eight-day championship to a rousing finale, but while the hosts topped the medal table from the United States there was criticism of the poor attendances in the Luzhniki Stadium. There was further concern when their pole vault gold medalist Yelena Isinbayeva made controversial remarks in support of Russia's new laws, which make "the propagandizing of non-traditional sexual relations among minors" a criminal offense. She later attempted to clarify her comments, but there were renewed calls by gay rights groups for a boycott of the 2014 Winter Games in Sochi, the next major sports event in Russia.

Predicted Summary

Usain Bolt wins his third gold at the world championships in Moscow. Bolt anchored Jamaica to victory in the men's 4x100m ,

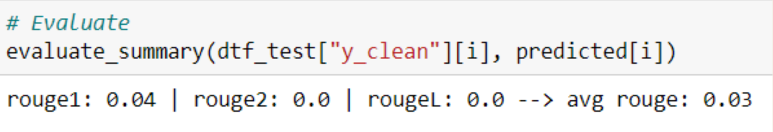
Real Summary

Usain Bolt wins third gold of world championship . Anchors Jamaica to 4x100m relay victory . Eighth gold at the championships for Bolt . Jamaica double up in women's 4x100m relay .



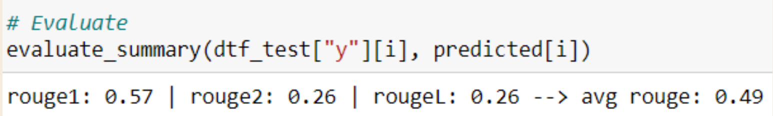
```
# Evaluate
evaluate_summary(dtf_test.y[i], predicted[i])
```

```
rouge1: 0.18 | rouge2: 0.04 | rougeL: 0.04 --> avg rouge: 0.14
```



```
# Evaluate
evaluate_summary(dtf_test["y_clean"][i], predicted[i])
```

```
rouge1: 0.04 | rouge2: 0.0 | rougeL: 0.0 --> avg rouge: 0.03
```



```
# Evaluate
evaluate_summary(dtf_test["y"][i], predicted[i])
```

```
rouge1: 0.57 | rouge2: 0.26 | rougeL: 0.26 --> avg rouge: 0.49
```

TEXT CLASSIFICATION (named entity recognition NER)

Text classification is the process of **categorizing** the text into a group of words. By using NLP, text classification can automatically analyze text and then assign a set of predefined tags or categories based on its context. NLP is used for sentiment analysis, topic detection, and language detection.
There is mainly three text classification approach-

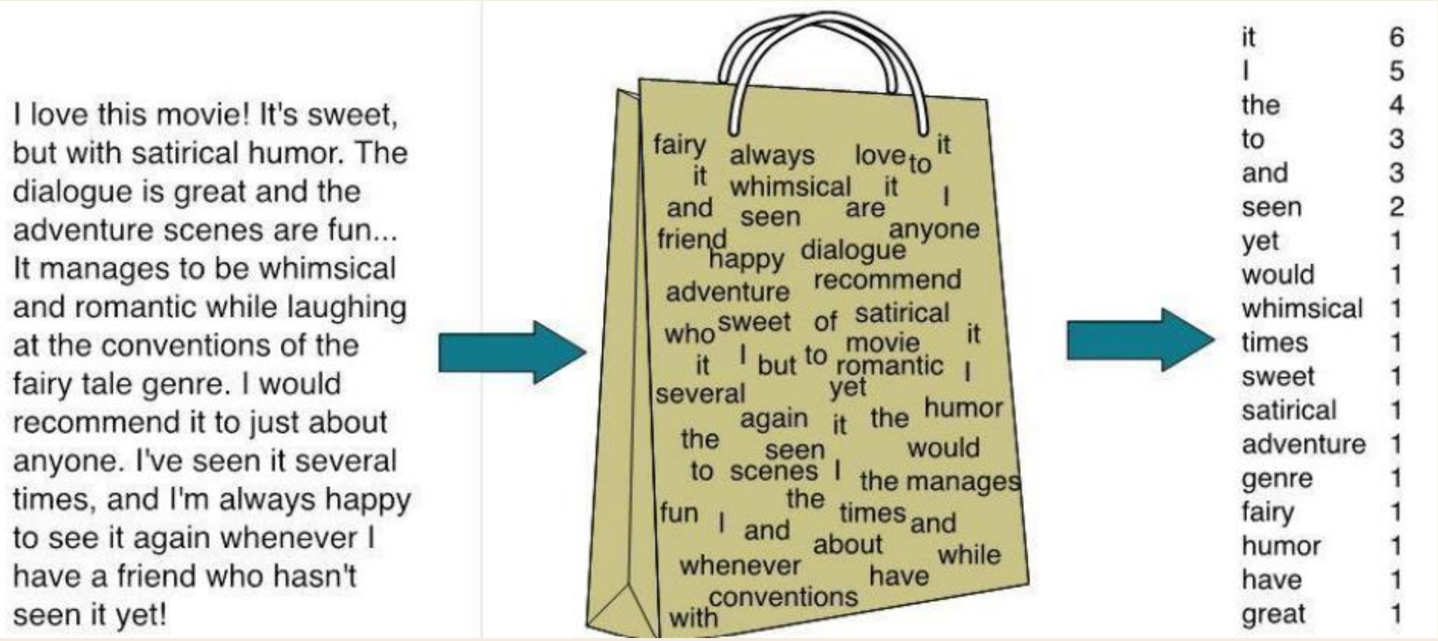
- Rule-based System
- Machine System
- Hybrid System

In the **rule-based approach**, texts are separated into an organized group using a set of handicraft linguistic rules. Those **handicraft linguistic rules** contain users to define a list of words that are characterized by groups. For example, words like **Narendra Modi** and **Arvind Kejriwal** would be categorized into **politics**. People like **Virat Kohli** and **Sachin Tendulkar** would be categorized into **sports**.

Machine-based classifier learns to make a classification based on past observation from the data sets. User data is prelabeled as train and test data. It collects the classification strategy from the previous inputs and learns continuously. Machine-based classifier usage a **bag of words** for **feature extension**.
In a **bag of words**, a **vector represents the frequency of words** in a predefined dictionary of a word list.

The third approach to text classification is the **Hybrid Approach**. Hybrid approach usage **combines** a **rule-based** and **machine Based approach**. Hybrid based approach usage of the rule-based system to create a tag and use machine learning to train the system and create a rule. Then the machine-based rule list is compared with the rule-based rule list. If something does not match on the tags, humans improve the list manually. It is the best method to implement text classification

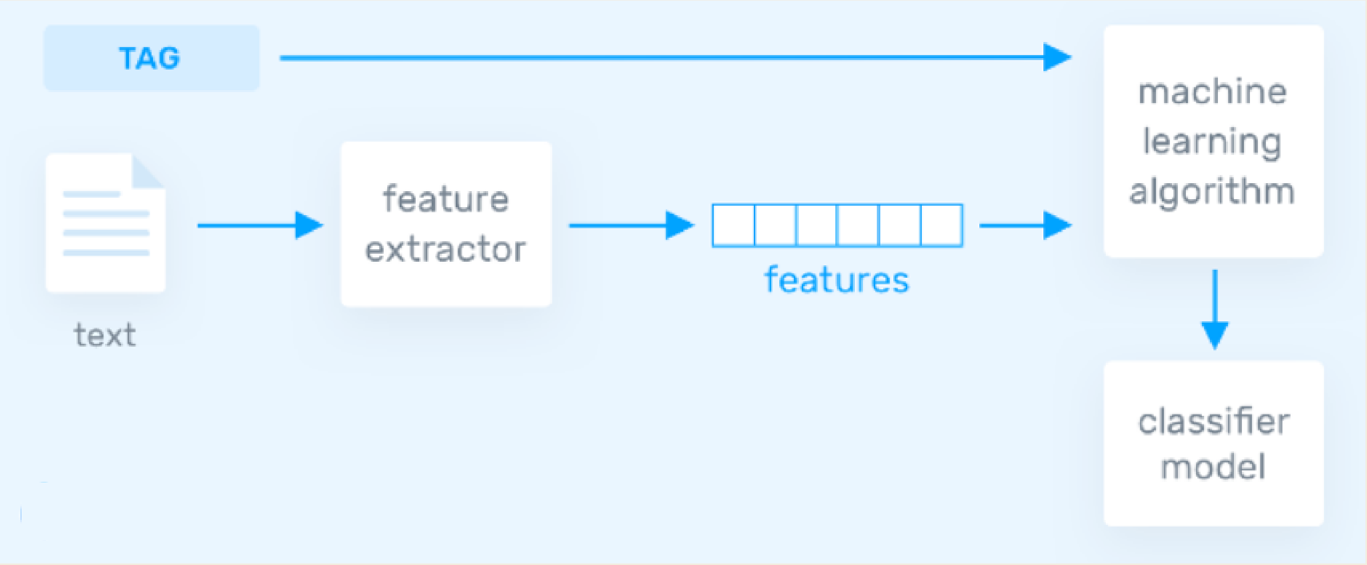
APPROACHES



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

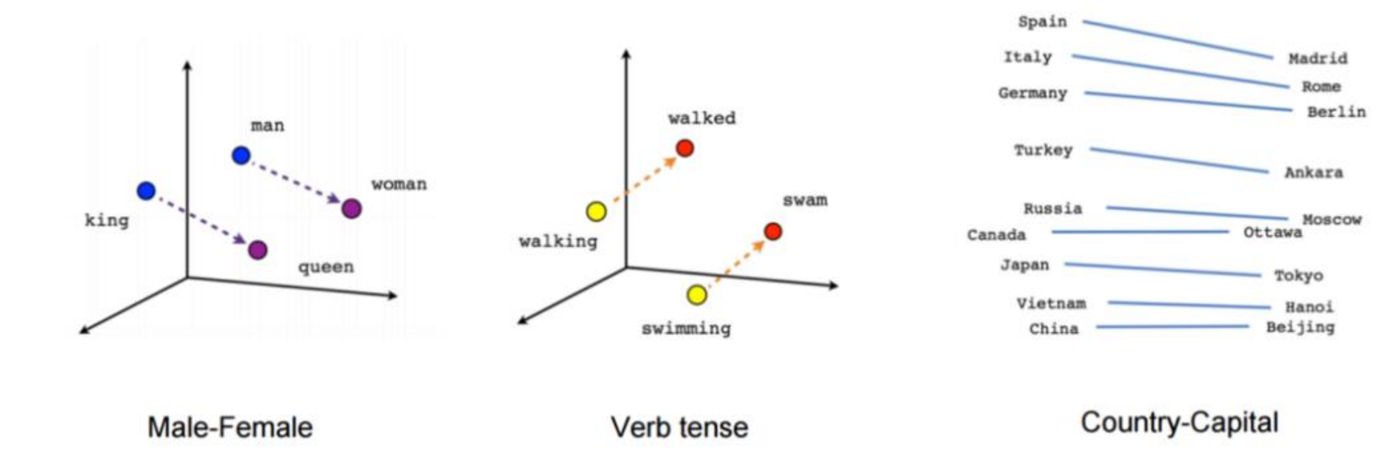
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

← Bag Of Words

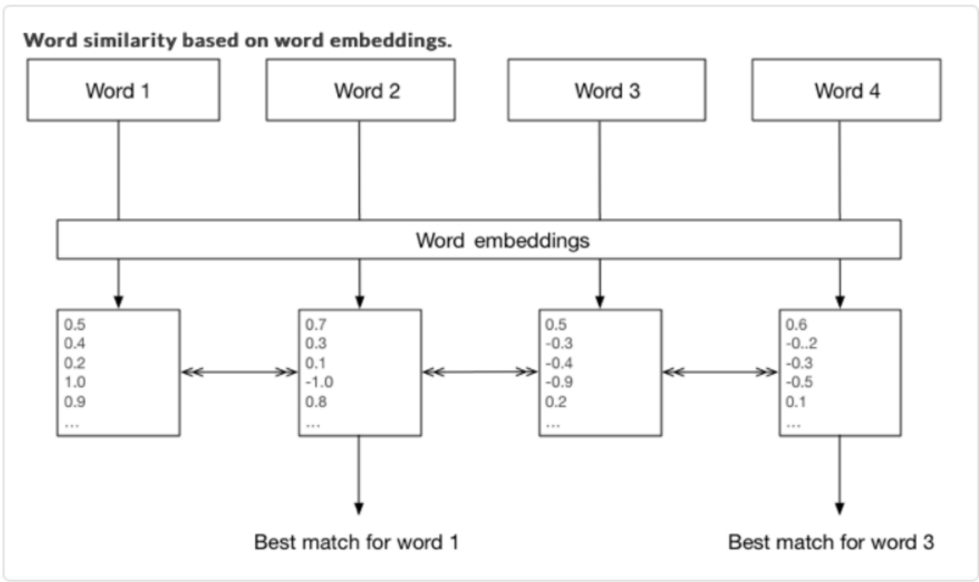


Process Methodology of feature extraction →

APPROACHES

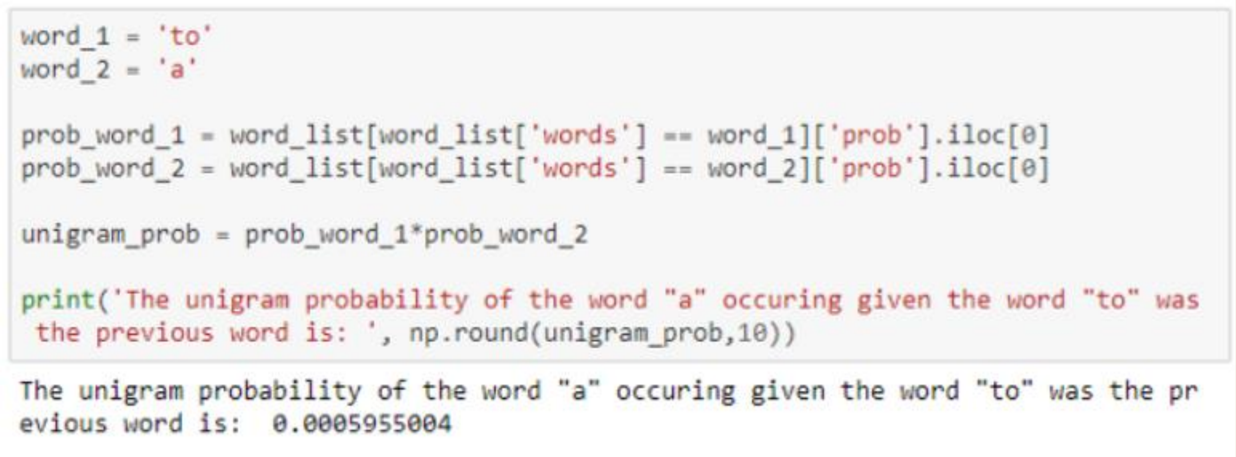


Vectorizing (Word2Vec)



Word Embeddings

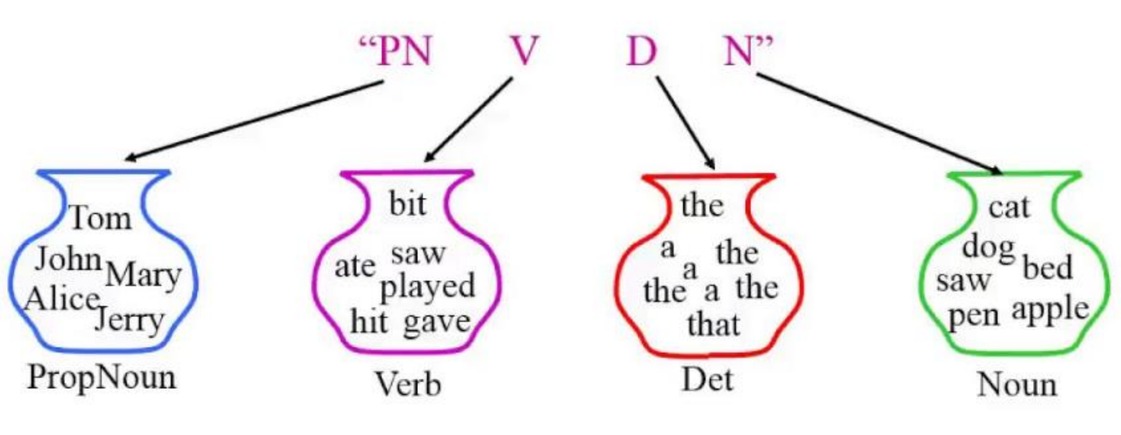
APPROACHES



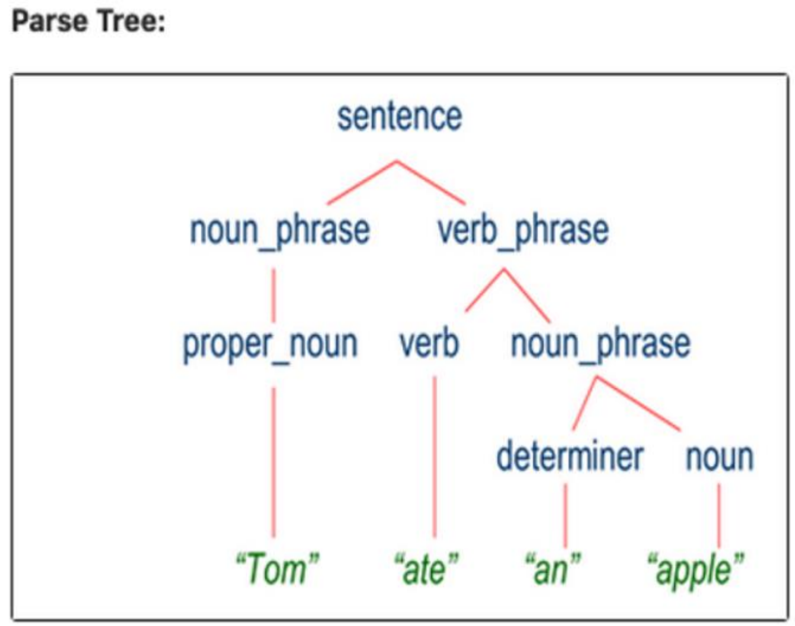
```
word_1 = 'to'
word_2 = 'a'

prob_word_1 = word_list[word_list['words'] == word_1]['prob'].iloc[0]
prob_word_2 = word_list[word_list['words'] == word_2]['prob'].iloc[0]

unigram_prob = prob_word_1*prob_word_2

print('The unigram probability of the word "a" occuring given the word "to" was
 the previous word is: ', np.round(unigram_prob,10))
```

```
The unigram probability of the word "a" occuring given the word "to" was the pr
evious word is:  0.0005955004
```

Probabilistic Method



Label Sequencing



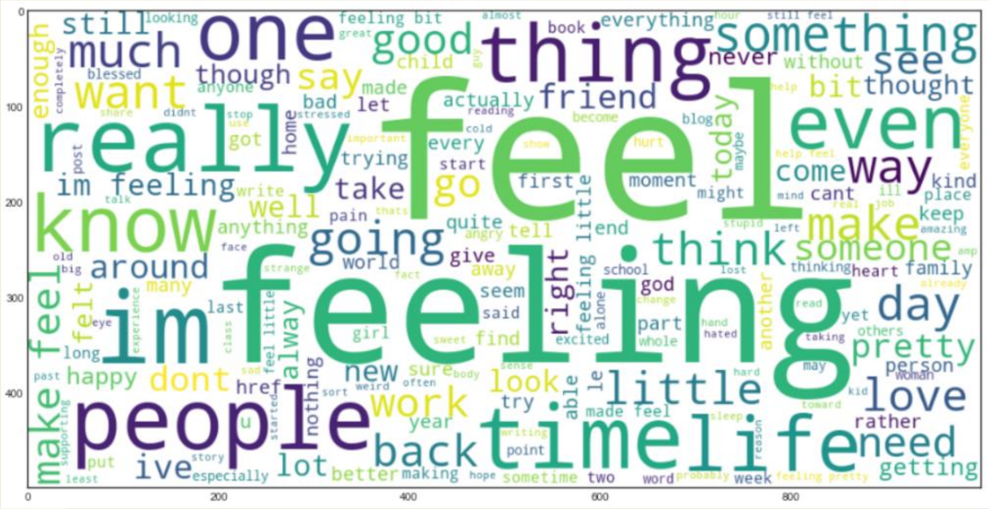
SENTIMENT ANALYSIS

Sentiment Analysis, means to **identify the view or emotion** behind a situation. It basically means to analyze and find the emotion or intent behind a piece of text or speech or any mode of communication.

- User Review 1: I love this cheese sandwich, it’s so delicious.
- User Review 2: This chicken burger has a very bad taste.
- User Review 3: I ordered this pizza today.

The first review is definitely a **positive** one and it signifies that the customer was really happy with the sandwich.
The second review is **negative**, and hence the company needs to look into their burger department.
And, the third one doesn’t signify whether that customer is happy or not, and hence we can consider this as a **neutral** statement.

Word Cloud is a data visualization technique used to depict text in such a way that, the more frequent words appear enlarged as compared to less frequent words. This gives us a little insight into, how the data looks after being processed through all the steps until now.

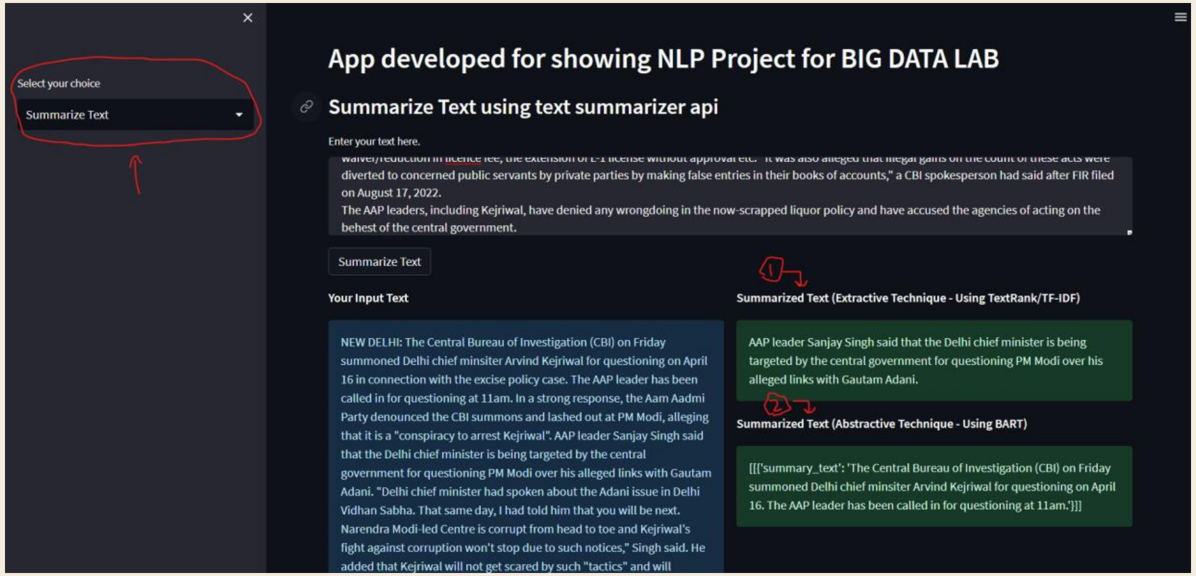


UI DEVELOPMENT & DEPLOYMENT USING STREAMLIT

Streamlit is an open-source app framework for Machine Learning and Data Science teams. Create beautiful web apps in minutes. Other popular frameworks are **Flask** and **Django**. But the issue with using these frameworks is that we should have some knowledge of HTML, CSS, and JavaScript.

App & Code walkthrough (next…)

Feature 1 : Select Summarize Text from dropdown & enter the your text. Afterwards click on Summarize text button



App developed for showing NLP Project for BIG DATA LAB

Summarize Text using text summarizer api

Enter your text here.

waiver/reduction in licence fee, the extension of L-1 license without approval etc. "It was also alleged that illegal gains on the count of these acts were diverted to concerned public servants by private parties by making false entries in their books of accounts," a CBI spokesperson had said after FIR filed on August 17, 2022.
The AAP leaders, including Kejriwal, have denied any wrongdoing in the now-scrapped liquor policy and have accused the agencies of acting on the behest of the central government.

Your Input Text

NEW DELHI: The Central Bureau of Investigation (CBI) on Friday summoned Delhi chief minsiter Arvind Kejriwal for questioning on April 16 in connection with the excise policy case. The AAP leader has been called in for questioning at 11am. In a strong response, the Aam Aadmi Party denounced the CBI summons and lashed out at PM Modi, alleging that it is a "conspiracy to arrest Kejriwal". AAP leader Sanjay Singh said that the Delhi chief minister is being targeted by the central government for questioning PM Modi over his alleged links with Gautam Adani. "Delhi chief minister had spoken about the Adani issue in Delhi Vidhan Sabha. That same day, I had told him that you will be next. Narendra Modi-led Centre is corrupt from head to toe and Kejriwal's fight against corruption won't stop due to such notices," Singh said. He added that Kejriwal will not get scared by such "tactics" and will

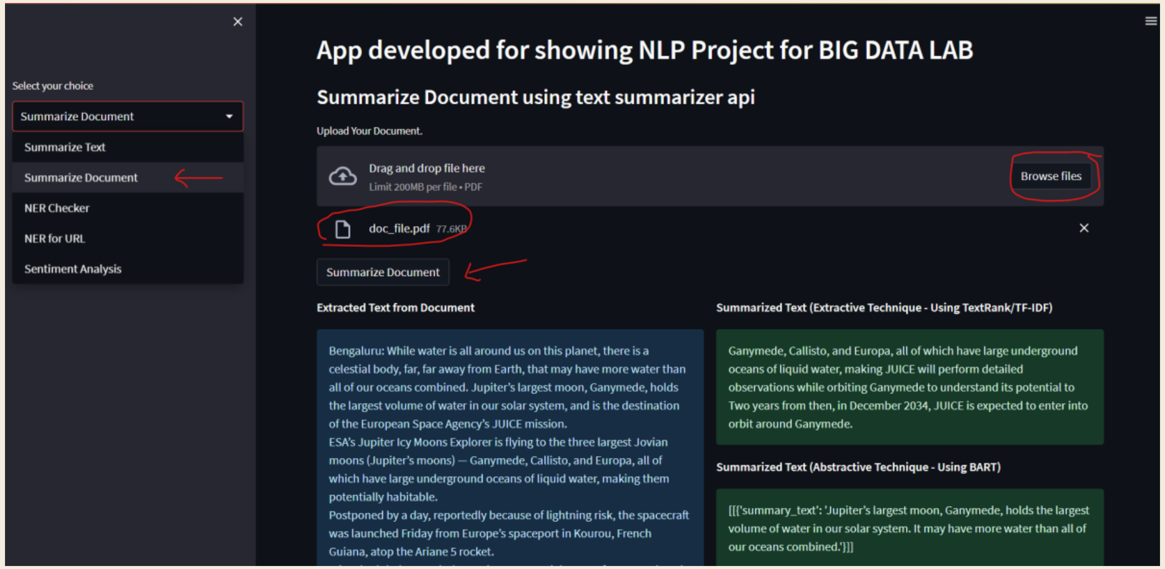
Summarized Text (Extractive Technique - Using TextRank/TF-IDF)

AAP leader Sanjay Singh said that the Delhi chief minister is being targeted by the central government for questioning PM Modi over his alleged links with Gautam Adani.

Summarized Text (Abstractive Technique - Using BART)

```
[[{'summary_text': 'The Central Bureau of Investigation (CBI) on Friday summoned Delhi chief minsiter Arvind Kejriwal for questioning on April 16. The AAP leader has been called in for questioning at 11am.'}]]
```

Feature 2 : Select Summarize Doc from dropdown & choose your file. Afterwards click on Summarize text button



App developed for showing NLP Project for BIG DATA LAB

Summarize Document using text summarizer api

Upload Your Document.

doc_file.pdf 77.6KB

Extracted Text from Document

Bengaluru: While water is all around us on this planet, there is a celestial body, far, far away from Earth, that may have more water than all of our oceans combined. Jupiter's largest moon, Ganymede, holds the largest volume of water in our solar system, and is the destination of the European Space Agency's JUICE mission.
ESA's Jupiter Icy Moons Explorer is flying to the three largest Jovian moons (Jupiter's moons) — Ganymede, Callisto, and Europa, all of which have large underground oceans of liquid water, making them potentially habitable.
Postponed by a day, reportedly because of lightning risk, the spacecraft was launched Friday from Europe's spaceport in Kourou, French Guiana, atop the Ariane 5 rocket.

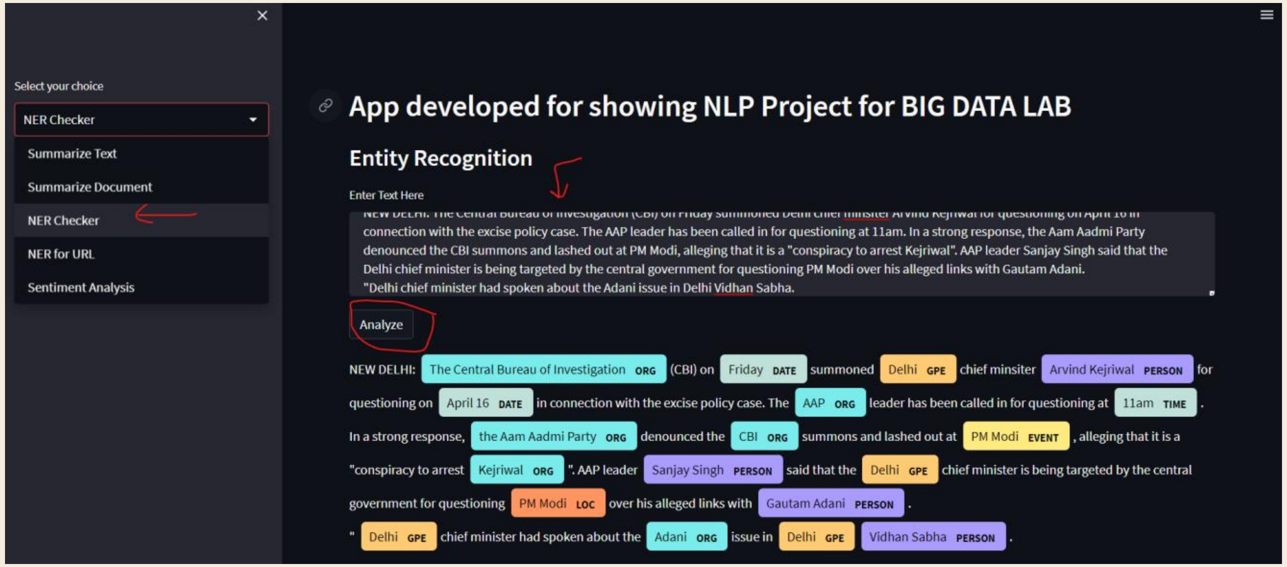
Summarized Text (Extractive Technique - Using TextRank/TF-IDF)

Ganymede, Callisto, and Europa, all of which have large underground oceans of liquid water, making JUICE will perform detailed observations while orbiting Ganymede to understand its potential to Two years from then, in December 2034, JUICE is expected to enter into orbit around Ganymede.

Summarized Text (Abstractive Technique - Using BART)

[[{'summary_text': 'Jupiter's largest moon, Ganymede, holds the largest volume of water in our solar system. It may have more water than all of our oceans combined.'}]]

Feature 3 : Select NER Analyze from dropdown & enter the your text. Afterwards click on Analyze button.



App developed for showing NLP Project for BIG DATA LAB

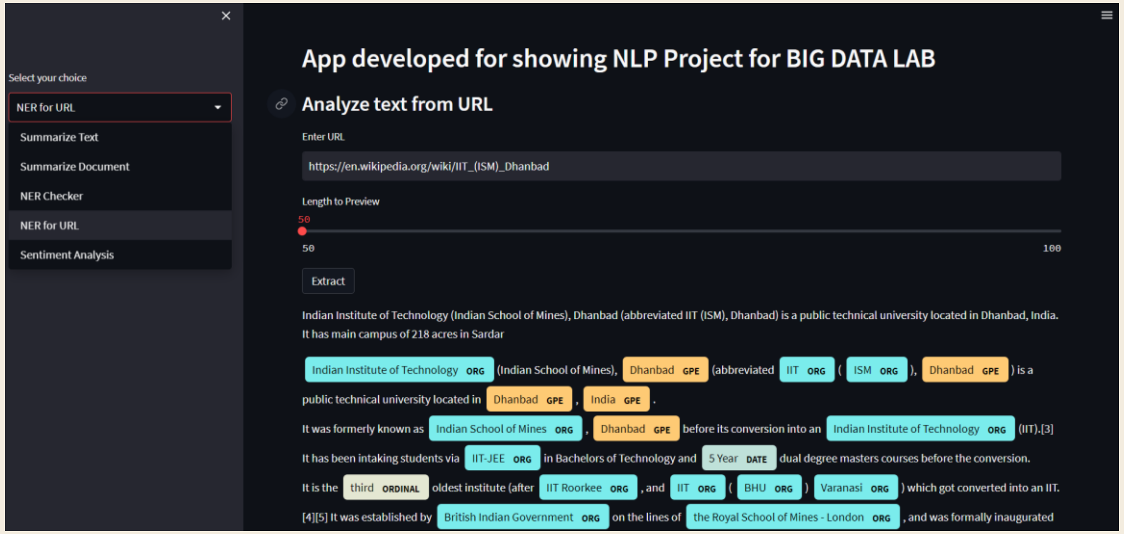
Entity Recognition

Enter Text Here

NEW DELHI: The Central Bureau of Investigation (CBI) on Friday summoned Delhi chief minsiter Arvind Kejriwal for questioning on April 16 in connection with the excise policy case. The AAP leader has been called in for questioning at 11am. In a strong response, the Aam Aadmi Party denounced the CBI summons and lashed out at PM Modi, alleging that it is a "conspiracy to arrest Kejriwal". AAP leader Sanjay Singh said that the Delhi chief minister is being targeted by the central government for questioning PM Modi over his alleged links with Gautam Adani.
"Delhi chief minister had spoken about the Adani issue in Delhi Vidhan Sabha.

NEW DELHI: The Central Bureau of Investigation ORG (CBI) on Friday DATE summoned Delhi GPE chief minsiter Arvind Kejriwal PERSON for questioning on April 16 DATE in connection with the excise policy case. The AAP ORG leader has been called in for questioning at 11am TIME . In a strong response, the Aam Aadmi Party ORG denounced the CBI ORG summons and lashed out at PM Modi EVENT , alleging that it is a "conspiracy to arrest Kejriwal ORG ". AAP leader Sanjay Singh PERSON said that the Delhi GPE chief minister is being targeted by the central government for questioning PM Modi LOC over his alleged links with Gautam Adani PERSON .
" Delhi GPE chief minister had spoken about the Adani ORG issue in Delhi GPE Vidhan Sabha PERSON .

Feature 4 : Select NER for link from dropdown & enter the url. Afterwards click on Extract button.



App developed for showing NLP Project for BIG DATA LAB

Analyze text from URL

Enter URL

https://en.wikipedia.org/wiki/IIT_(ISM)_Dhanbad

Length to Preview

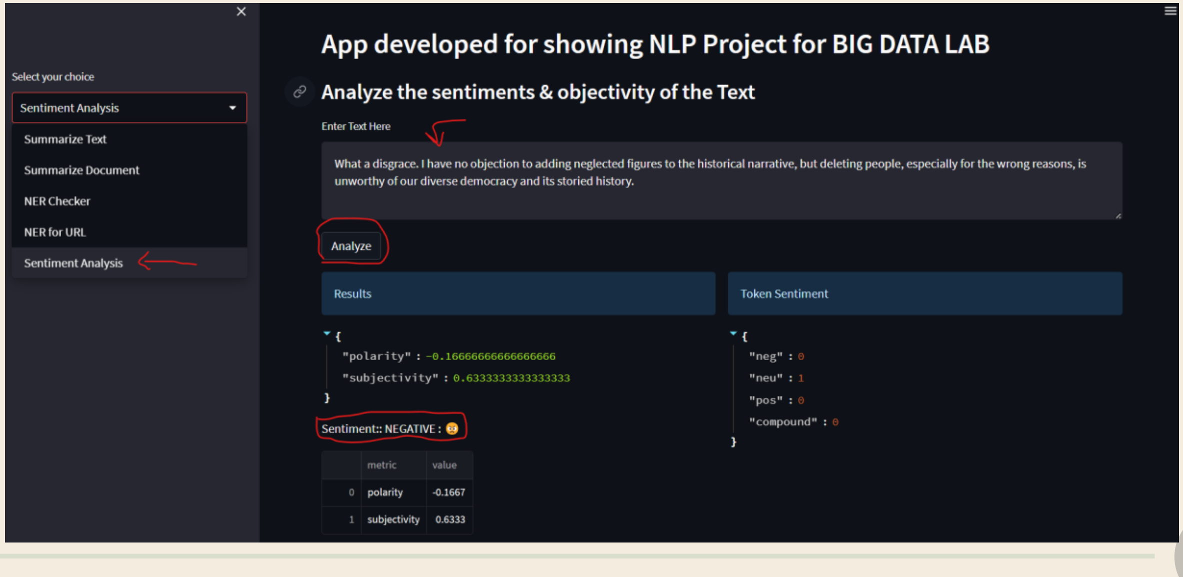
50

50 100

Indian Institute of Technology (Indian School of Mines), Dhanbad (abbreviated IIT (ISM), Dhanbad) is a public technical university located in Dhanbad, India. It has main campus of 218 acres in Sardar

Indian Institute of Technology ORG (Indian School of Mines), Dhanbad GPE (abbreviated IIT ORG (ISM ORG), Dhanbad GPE) is a public technical university located in Dhanbad GPE , India GPE .
It was formerly known as Indian School of Mines ORG , Dhanbad GPE before its conversion into an Indian Institute of Technology ORG (IIT).[3]
It has been intaking students via IIT-JEE ORG in Bachelors of Technology and 5 Year DATE dual degree masters courses before the conversion.
It is the third ORDINAL oldest institute (after IIT Roorkee ORG , and IIT ORG (BHU ORG) Varanasi ORG) which got converted into an IIT.
[4][5] It was established by British Indian Government ORG on the lines of the Royal School of Mines - London ORG , and was formally inaugurated

Feature 5 : Select Sentiment Analyzer from dropdown enter the tweet from twitter. Afterwards click on Analyze button.



App developed for showing NLP Project for BIG DATA LAB

Analyze the sentiments & objectivity of the Text

Enter Text Here

What a disgrace. I have no objection to adding neglected figures to the historical narrative, but deleting people, especially for the wrong reasons, is unworthy of our diverse democracy and its storied history.

Results

```
{
  "polarity" : -0.16666666666666666
  "subjectivity" : 0.6333333333333333
}
```

Sentiment:: NEGATIVE : 😟

	metric	value
0	polarity	-0.1667
1	subjectivity	0.6333

Token Sentiment

```
{
  "neg" : 0
  "neu" : 1
  "pos" : 0
  "compound" : 0
}
```

THANK YOU