RNN:
NEXT WORD PREDICTION

Wilfred Djumin DAAA/FT/2B/05

TASK

1. Build a next word predictor given a sequence of words
2. Predict next 10 words for each given input of words
3. Find ways to evaluate generated sequence of words

Overview

- Data Loading + EDA
- Preprocessing (Tokenization, Input Output Pairs)
- Modelling (SimpleRNN , GRU , LSTM ,etc.)
- Evaluation (Accuracy , METEOR , Perplexity , Readability)
- Improving on Models (Bi-LSTM & Bi-GRU)
- Conclusions

EVALUATION METHODS

Evaluation Methods

- **Model Accuracy** (Train and Validation)**:** As we are performing sequence generation, we should try to complement accuracy with other metrics, as generated text may not necessarily be classified into predefined categories
- **METEOR Score:** considers precision, recall, stemming, synonymy, and word order, more on **Quality of generated text**, higher the better
- **Perplexity** : Helps evaluate the **fluency and coherence** of generated sequences, the **lower the score the better**
- **Readability** (Flesch Score): Assesses the **ease** with which a text can be **read** and understood
- **Human Evaluation:** Creativity and appropriateness are challenging to be measured purely on metrics

Function: predict_next_N_words

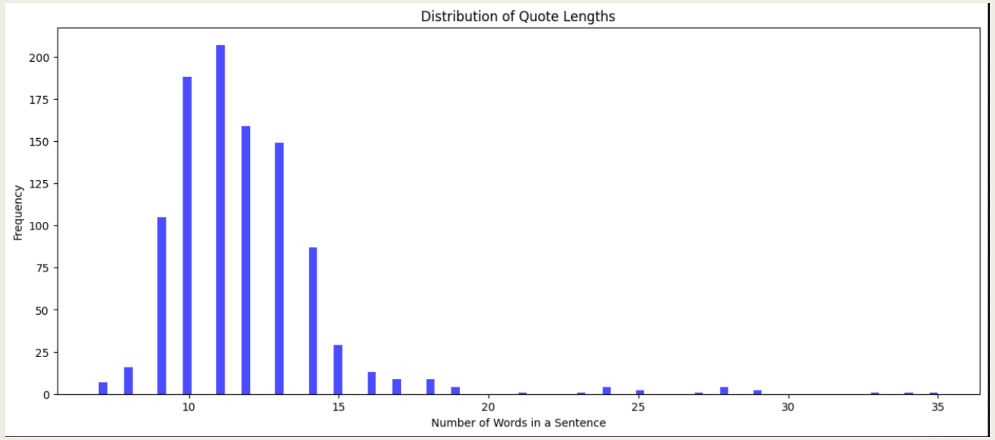
- This function generates a sequence of words given an initial input text using a trained RNN .
- Input Parameters:
  - *model: Trained RNN model.*
  - *input_text: Initial text for generating the sequence (seed texts in this case).*
  - *N_words: Number of words to generate (default: 10).*
  - *input_length: Maximum sequence length (default: max_sequence_len-1).*
  - *temperature: Controls randomness in the output (default: 1.0)*

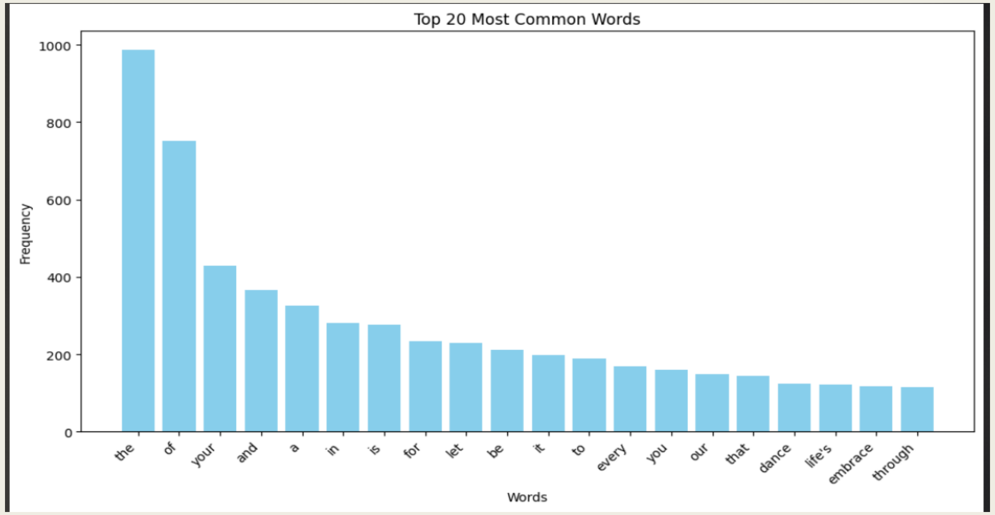
The temperature parameter allows for more “creative” generated texts by setting it >1.0 while also getting “precise” texts <1.0

DATA LOADING & EDA

Data Loading & EDA

- Our data contains 1000 quotes of 10 different ‘themes’ , each having 100 quotes per theme
- Most quotes fall under 20 words, as seen from the distribution plot, while the maximum word count for a quote is 35
- Most common words includes “the” , “of” and “your”





PREPROCESSING

Tokenization & Input Output Pairs

- For our case, we would try retaining punctuations like apostrophes, semicolons, and commas as they play a part in sentence sequence and coherence
- To generate more Input Output pairs, we can make use of a loop, iterating through till the max sequence length of our data (35), and will consider every possible combination of lists ranging from 2 to 35 words which helps us get more data
- We then pad them with the length of 35
- We split X and y into train and validation sets with this shapes =>

```
X_train shape: (48247, 33)
y_train shape: (48247,)
X_val shape: (20678, 33)
y_val shape: (20678,)
```

- Then apply One Hot Encoding to y_train and y_val

MODELLING

SimpleRNN

Model Architecture

- Embedding (128)
- SimpleRNN (64)
- Dropout (0.3)
- Dense

“Precise” prediction (0.2 Temperature):

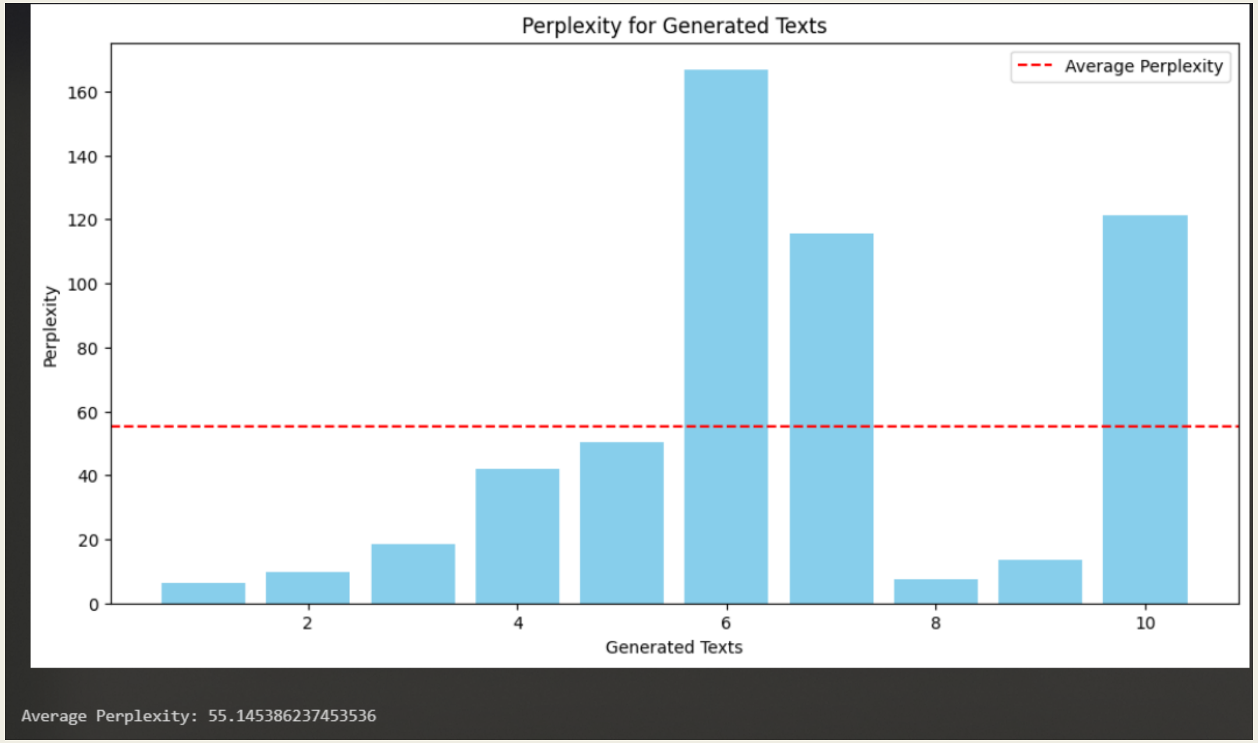
“radiate some positivity, for it is the heartbeat of transformation that our”

“Creative” prediction (2.0 Temperature):

'radiate some growth; it is the canvas, your reality spread wide open’

- We notice that for the more “creative” sequence, it is more random put less meaningful as compared to the “precise” sequence which has a decent sentence structure
- Train Accuracy of 0.75 , Validation Accuracy of 0.76
- METEOR Scores of “Precise” text (0.375) was also higher than “Creative” text (0.285) which is expected since the “Creative” sequences are more random , which results in overall poorer quality

Perplexity of SimpleRNN



- The Model mainly struggles to continue sequences for seed text 6, 7 & 10 meaning that model is less certain and less accurate in predicting the next word

- Average Perplexity of 55.14

6: let your time and energy is the realization of your heart lead a day is
7: every person is a garden of kindness we plant its burdens be the
10: morning and evening would make it is the foundation of every action and decision shaping the

MODEL IMPROVEMENTS

Bi-GRU

allows model to capture both past and future input sequences , wherelse regular GRU only has access to past input sequences

We would be using **KerasTuner** & **GridSearch** to tune:

- Learning Rate
- Dropout rate
- No. of Layers
- Nodes

“Precise” prediction (0.2 Temperature):

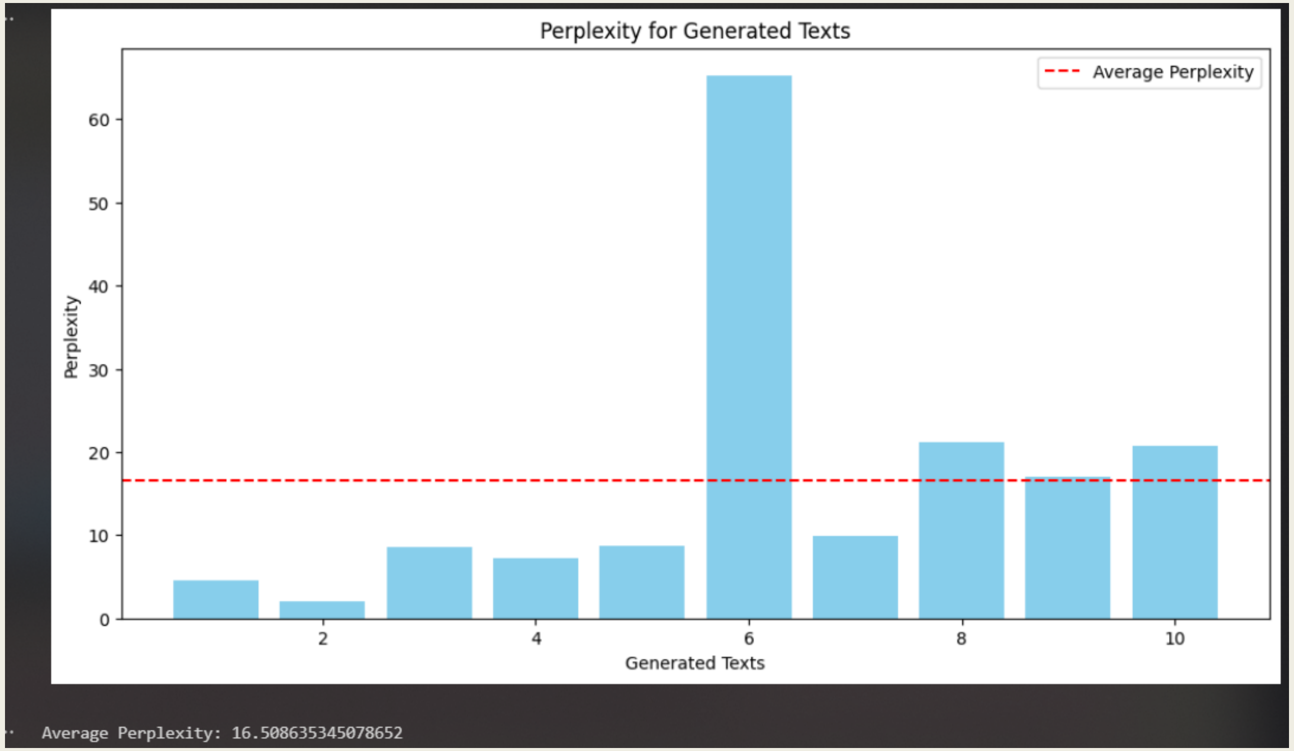
“radiate some confidence, and let it be the foundation of your greatness”

“Creative” prediction (1.5 Temperature):

'radiate some confidence, and let it be the armor that shields your’

- We do notice that the generated sequences are more coherent and well structured, as compared to the ones that SimpleRNN generated
- Train accuracy of 0.78 , Validation accuracy of 0.77 , improved a little vs SimpleRNN
- METEOR Scores of “Precise” text (0.38) was also higher than “Creative” text (0.36) which is expected since the “Creative” sequences are more random , which results in overall poorer quality
- “Precise” text also was more readable, having a higher FLESCH score of 66 (compared to 60)

Perplexity of Bi-GRU



- Like SimpleRNN, Bi-GRU Model mainly struggles to continue sequences for seed text 6 , as the perplexity was the highest, which is common as seen in SimpleRNN

- Average Perplexity of 16.50 is lower than that of the SimpleRNN (55.14) , indicating a

6: let your time and energy ripples out, creating abundance around you leave behind a precious

CONCLUSIONS

Conclusions

- Overall, GRU performed better than the SimpleRNN, but LSTM by itself did perform worse than SimpleRNN
- However, using Bidirectional GRU / LSTM does help the model improve
- Although most of our model's accuracy were not as high(around 75% range), we shouldn't solely look at accuracy metrics in the case of text generation, as it does not cover the fluency, coherence and relevance to the input or context
- However, we could probably increase our model's performance if we had performed some text augmentation
- It can be quite challenging to evaluate the model's "creativity" using a metric, and usually, human evaluation works better

THANK YOU