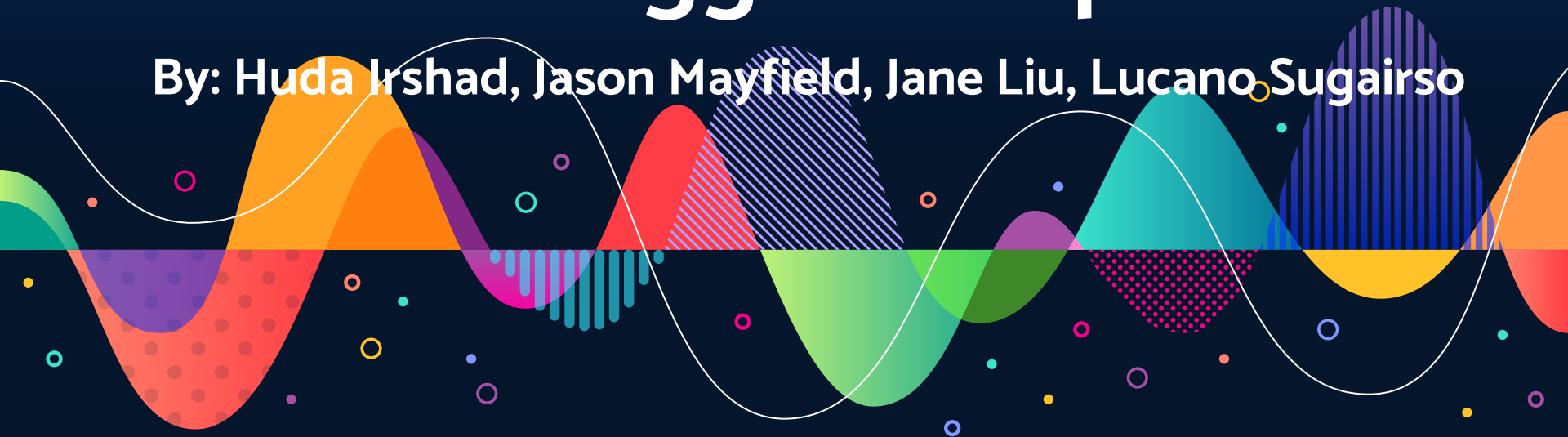


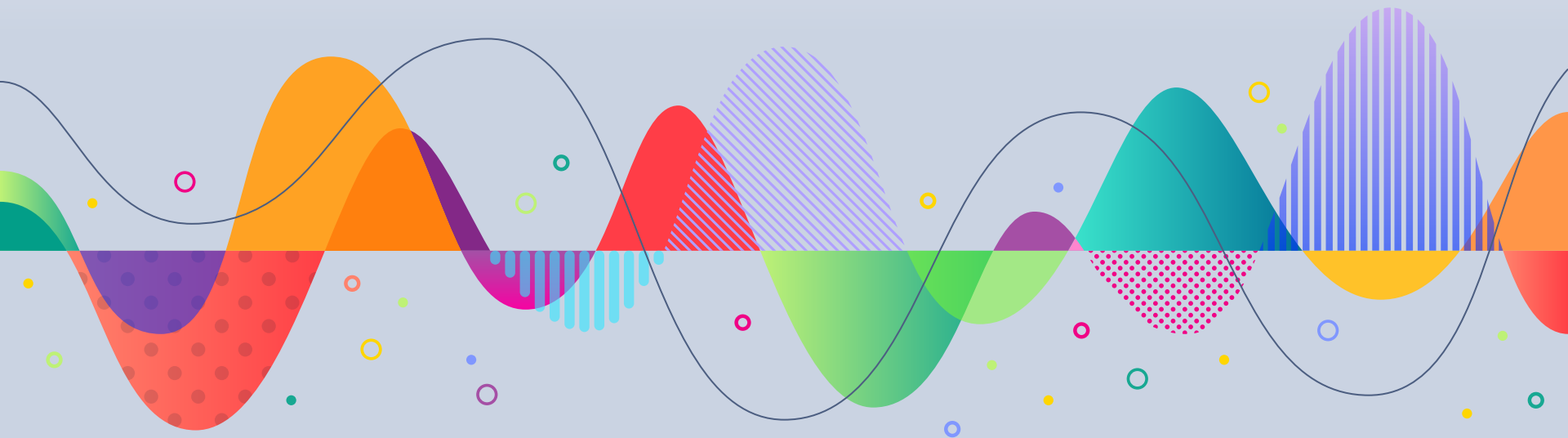
Automatic Stance Detection (Contradictory, My Dear Watson Kaggle Competition

By: Huda Irshad, Jason Mayfield, Jane Liu, Lucano Sugairso



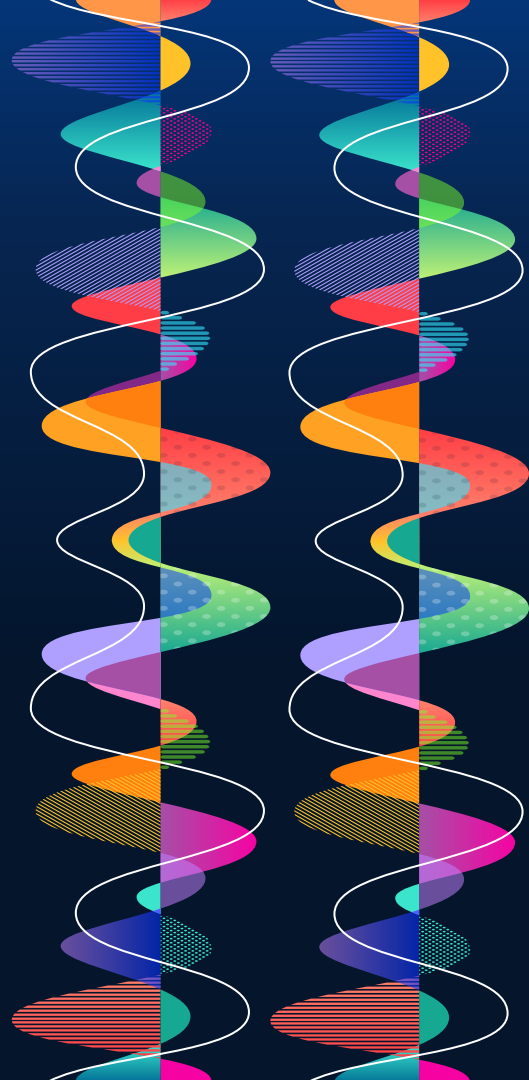
1.

The Problem to Solve



Situation We are solving for

- Given: Text
- Provide: Relation

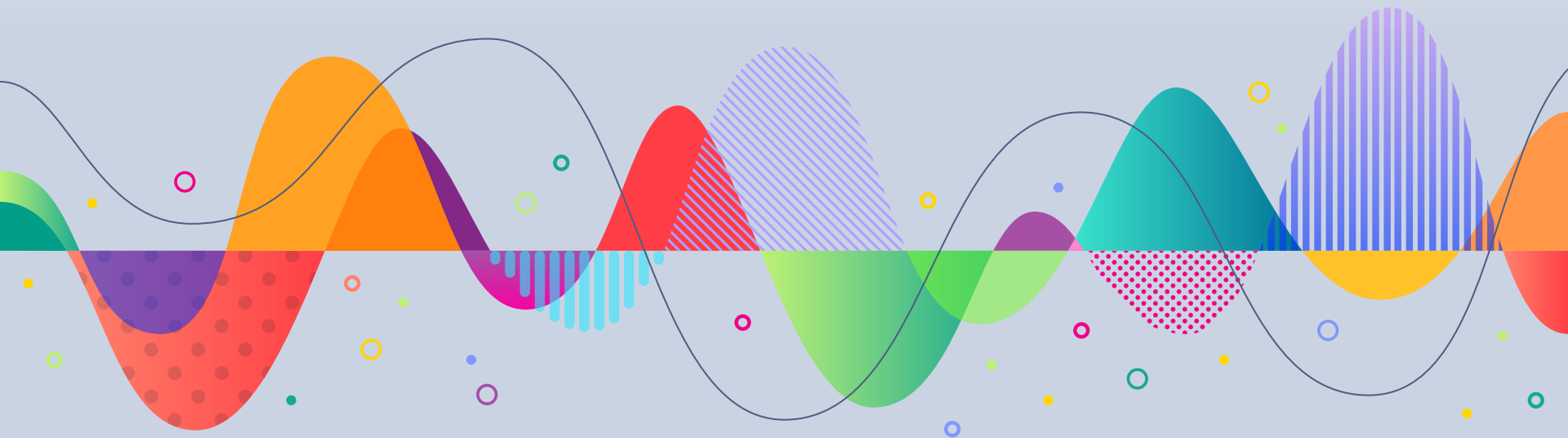


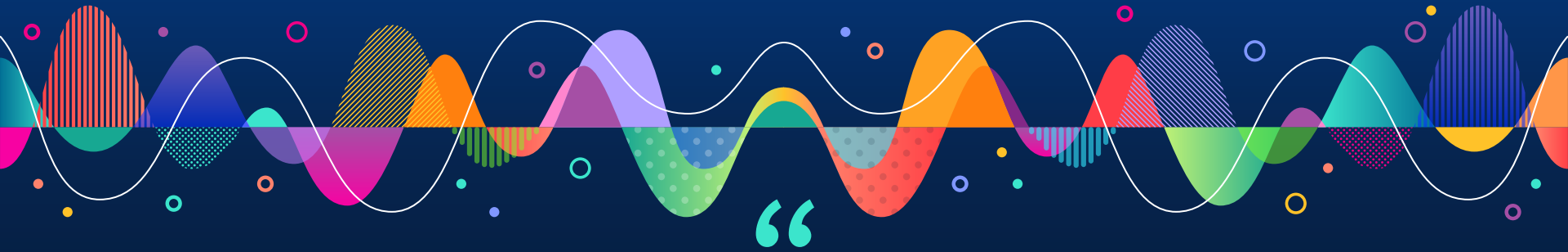
Motivating Problems



- ▷ Fake News Identification
- ▷ Text Analysis
- ▷ Fact Checking

2. Goals





Classify a pairing of sentences into the following categories

1. Entailment
2. Contradiction
3. Neutral

Examples

Premise

He came, he opened the door and I remember looking back and seeing the expression on his face, and I could tell that he was disappointed.

Hypothesis 1

Just by the look on his face when he came through the door I just knew that he was let down.

Entailment

Model

Hypothesis 2

He was trying not to make us feel guilty but we knew we had caused him trouble.

Neutral

Model

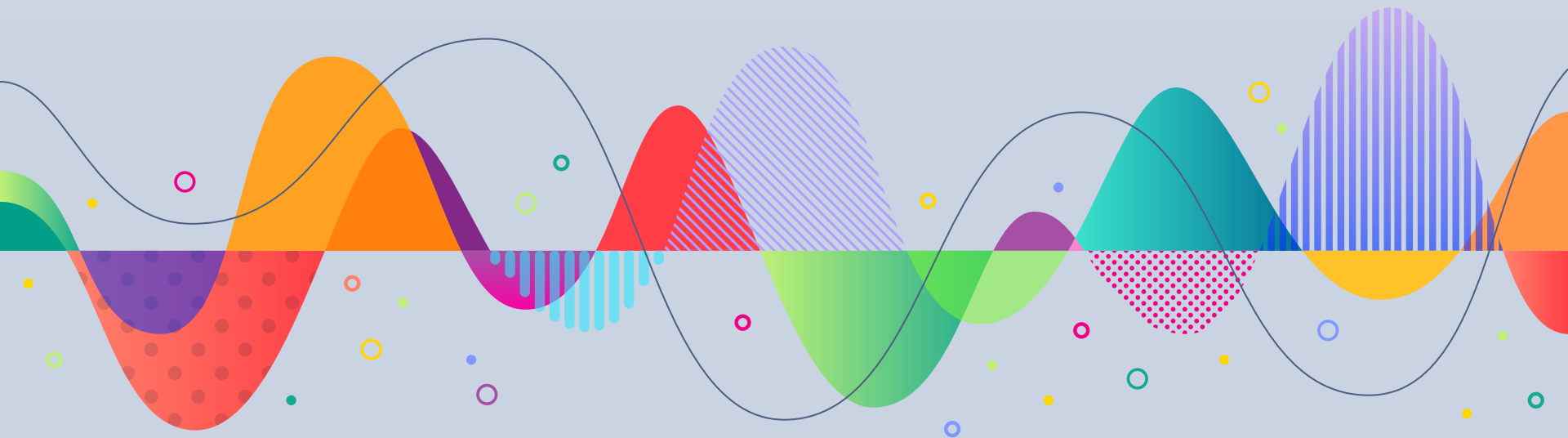
Hypothesis 3

He was so excited and bursting with joy that he practically knocked the door off it's frame.

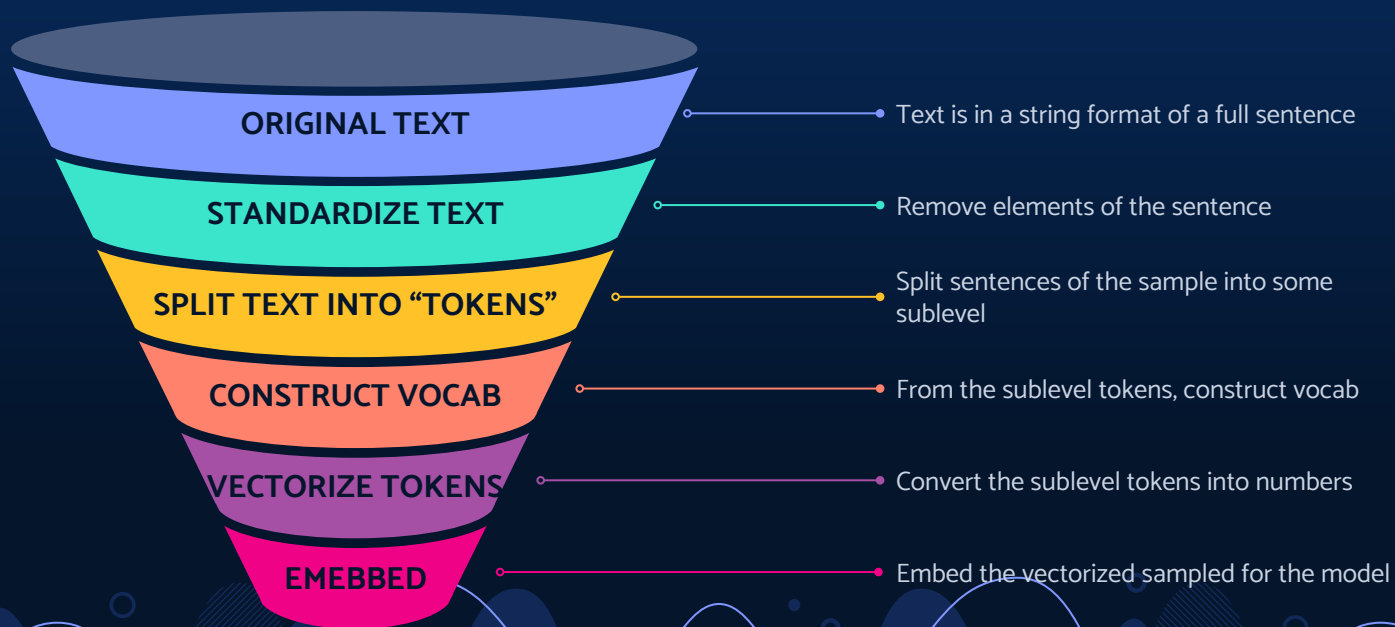
Contradiction

3.

Preprocessing of Data



Pre-processing of Data



Standardize Sentences

▷ Concerns

- ▶ Additional space taken from slight differences in English language; letter cases, punctuation, and contractions

▷ Solution

- ▶ Chose lowercase for all characters
- ▶ Expanded all contractions
- ▶ Remove punctuations (e.g., exclamation marks, commas, colons, etc.)

Original Text

“Hello! You’re welcome to our presentation.”



Lower Case

“hello! you’re welcome to our presentation.”



Remove Contractions

“hello! you are welcome to our presentation.”



Remove Punctuation

“hello you are welcome to our presentation

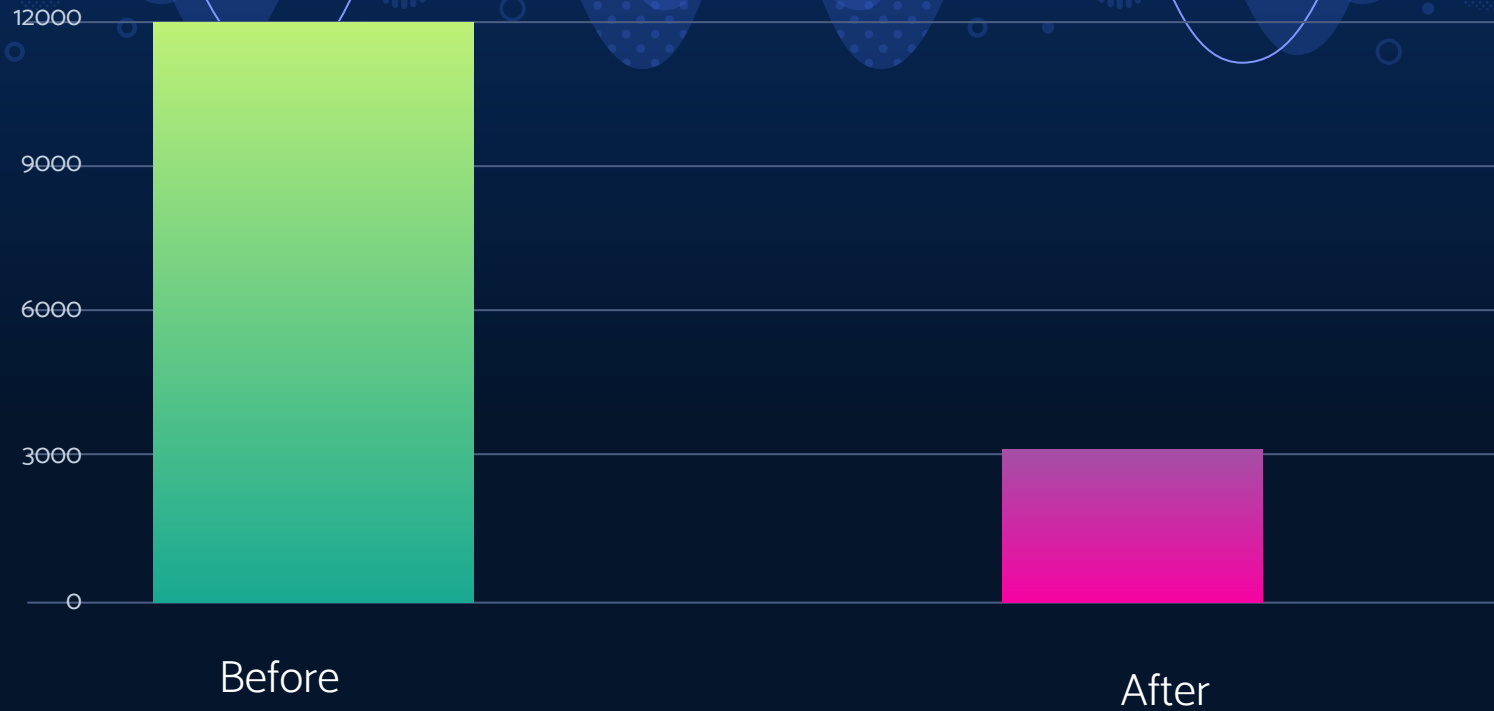


Creating Vocabulary

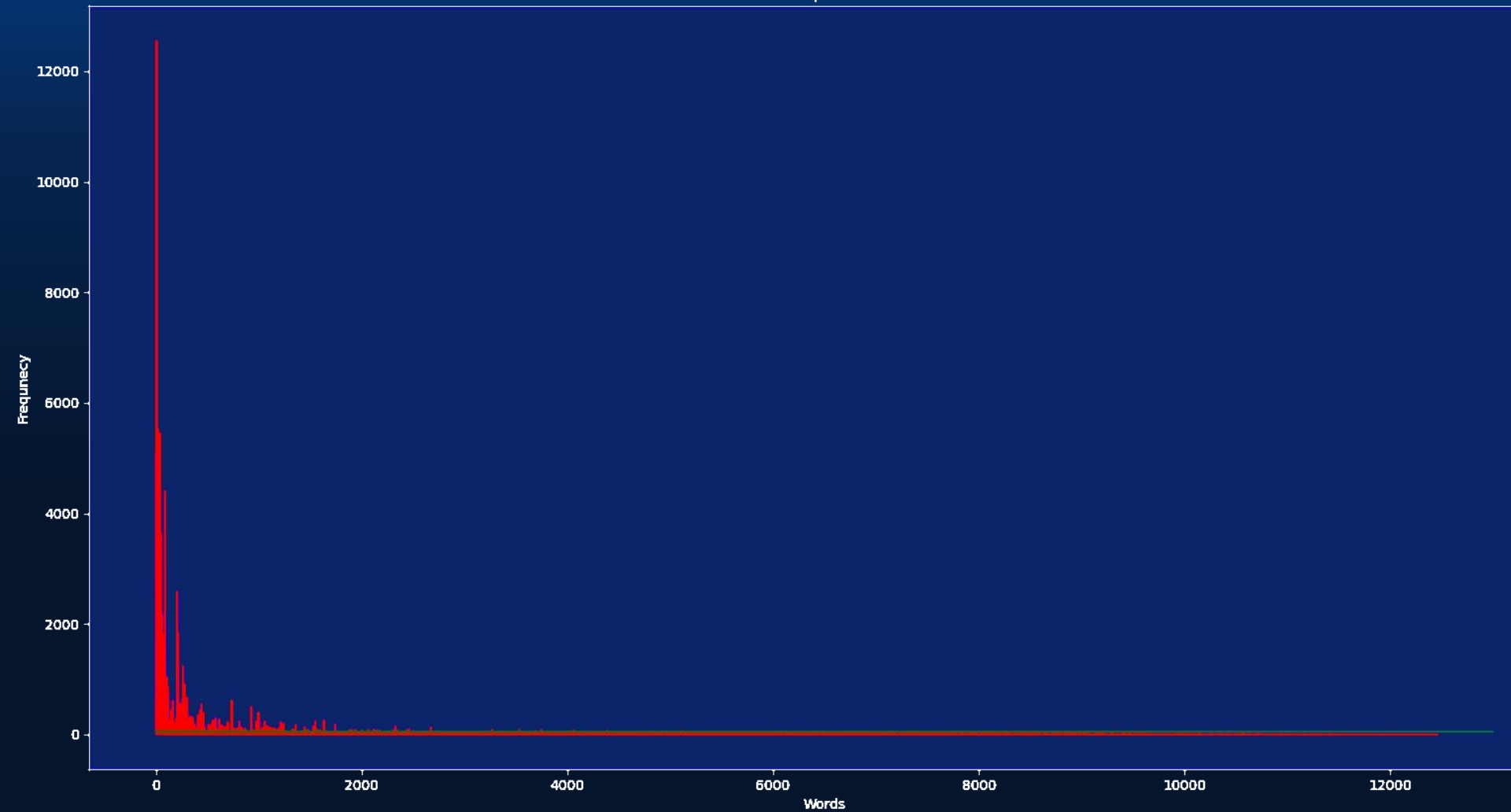
Concern: Too many words given

Solution : Cut off the less useful words

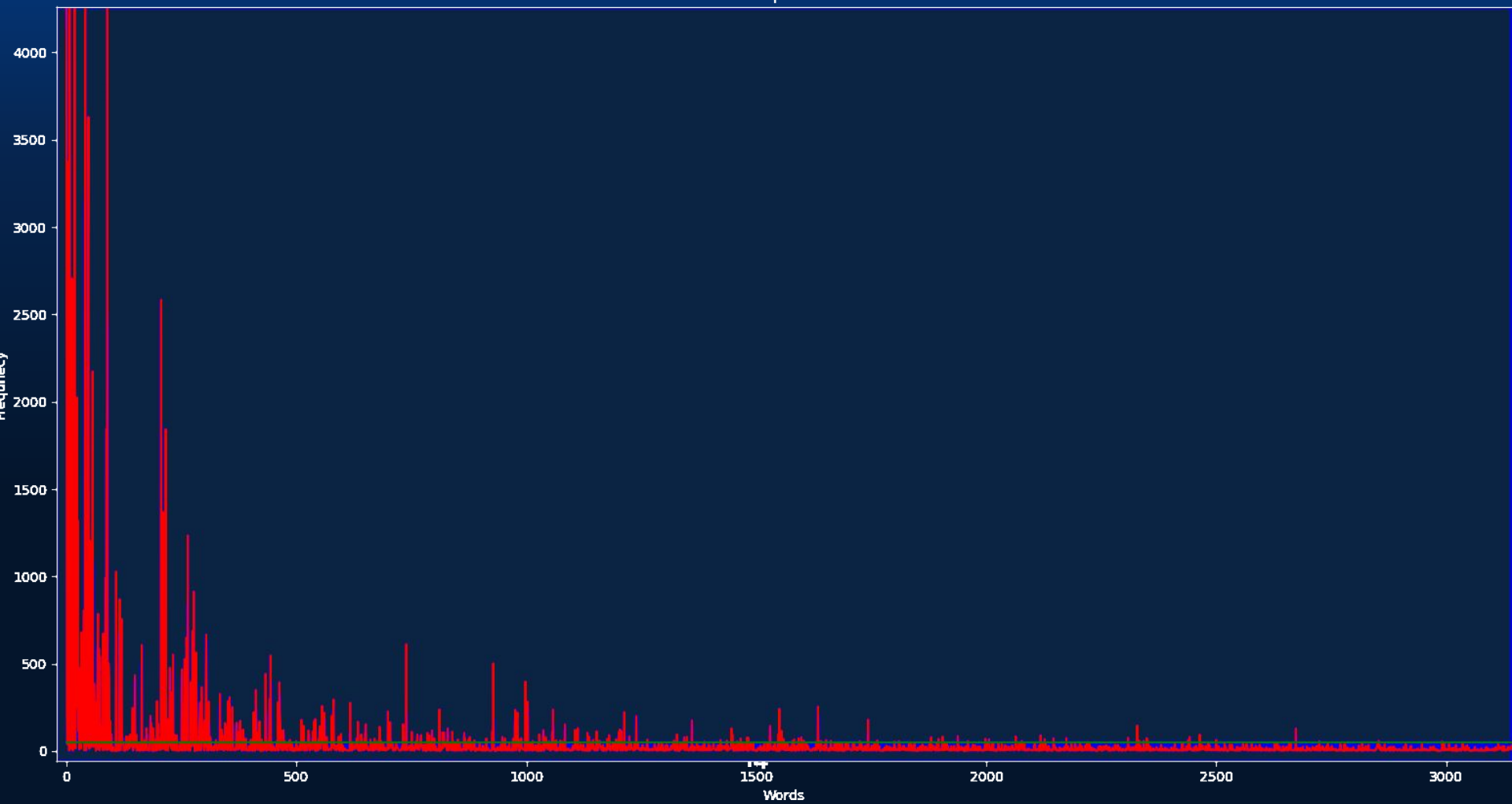
Creating Vocabulary



All unique words



All unique words



Additional Vocabulary Tokens

START & STOP

To mark the beginning and end of sentences.

To standardize the length of sentences for consistent inputs.

PAD

To replace all the words that were not used enough to be used in the dictionary.

UNKNOWN

Tokenization/ Embedding

this is my sample sentence for tokenization because it is fun

this is my sample sentence for tokenization because it is fun

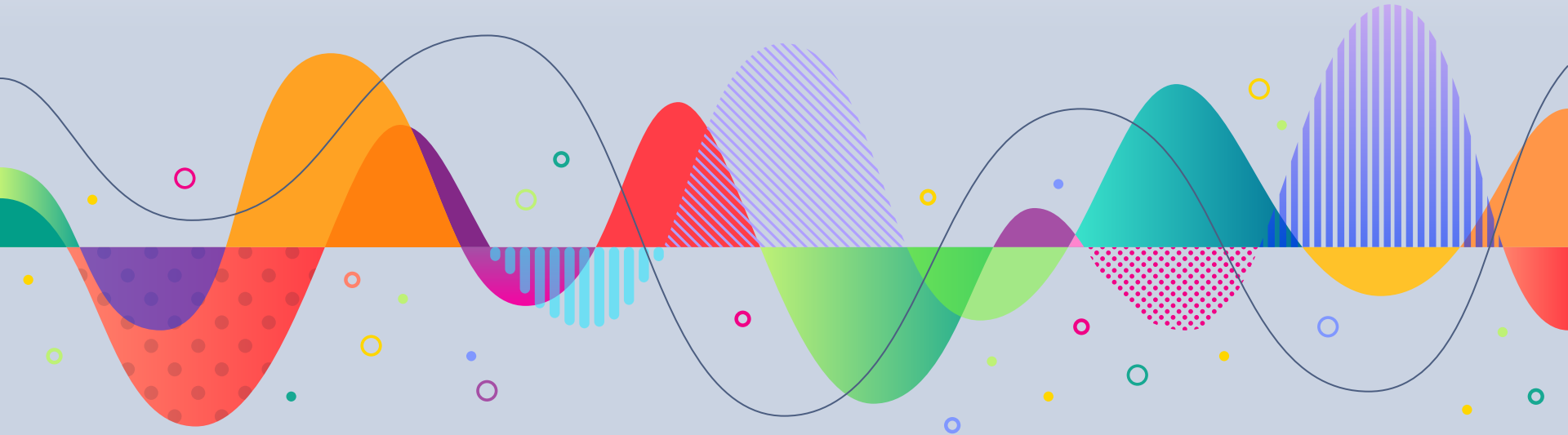
Dictionary

<ukn> : 0	<strt> : 3	Because : 6
<pad> : 1	tokenization : 4	fun : 7
<stp> : 2	sentence : 5	sample : 8

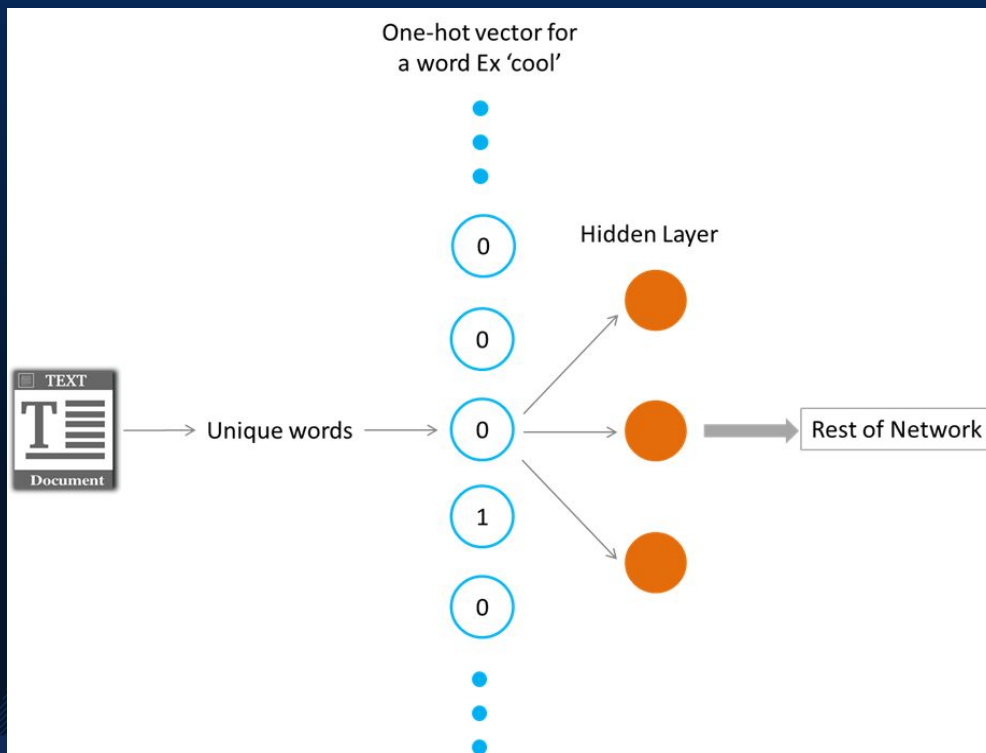
3 0 0 8 5 0 4 6 6 0 7 1 2

4.

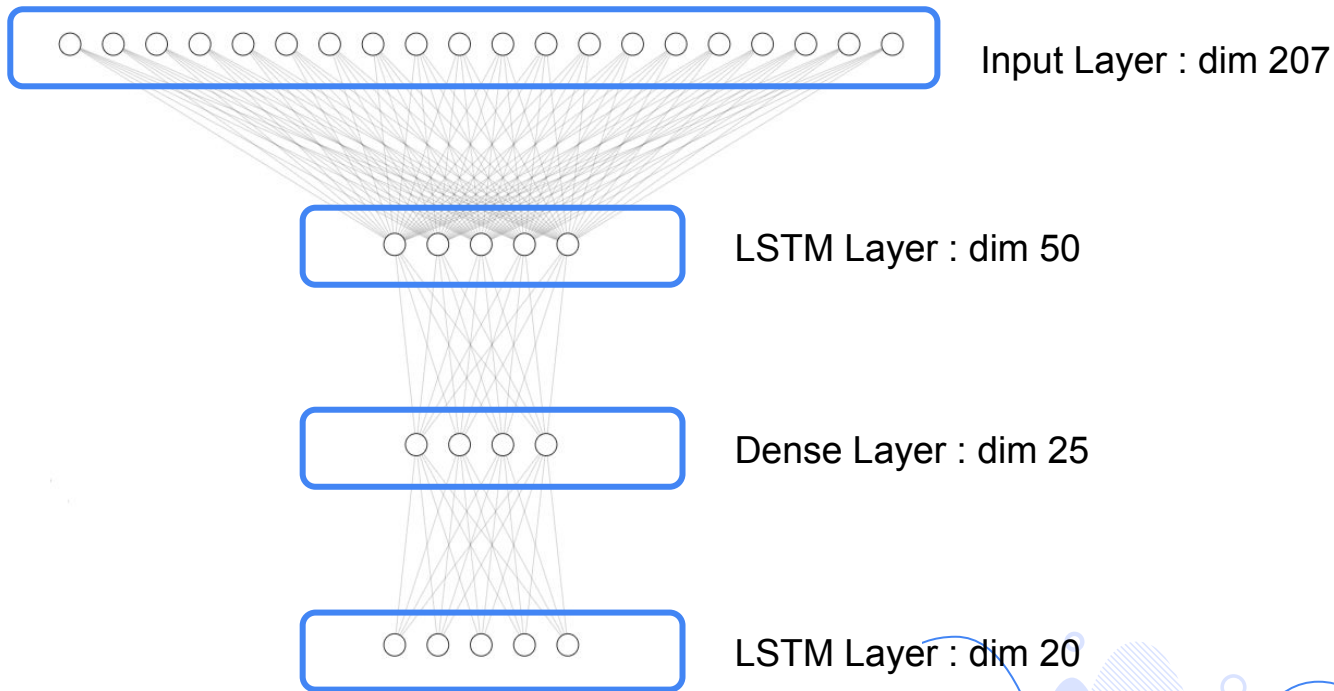
Data Processing



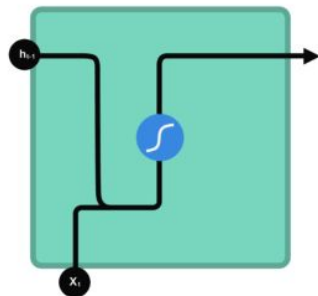
What Is Word Embedding



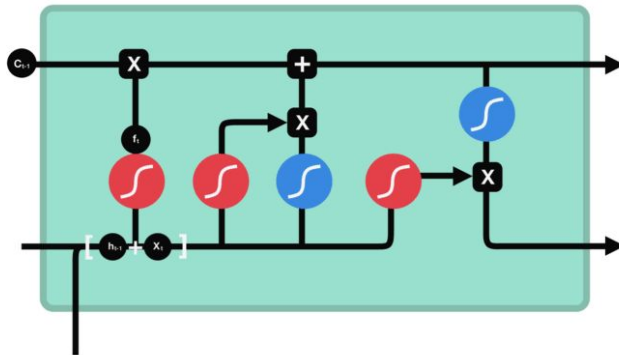
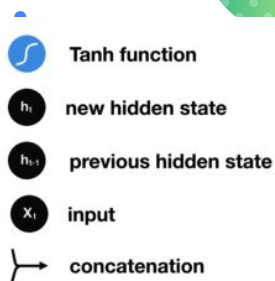
MODEL For each sentence



LSTM layer - Gates and Sigmoid Activation

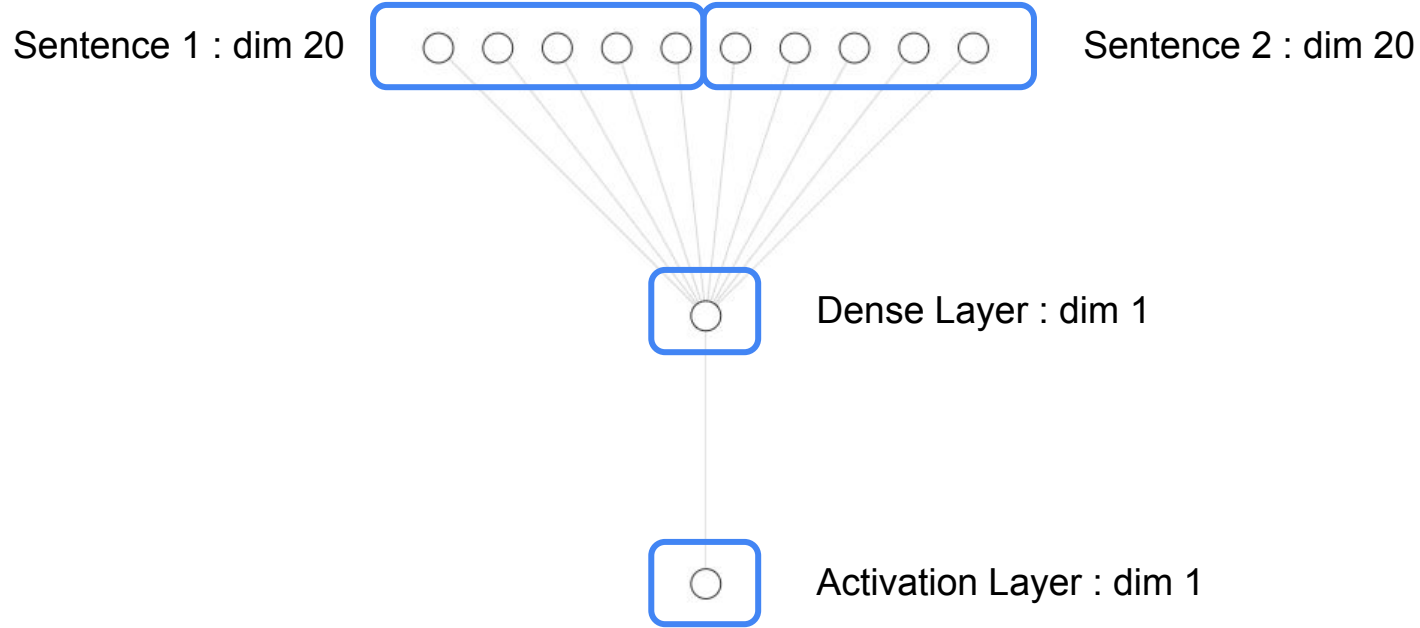


RNN Cell



LSTM Cell

MODEL To Combine two sentences processing



Results

Text(0.5, 1.0, 'Loss and Accuracy Per Epoch')

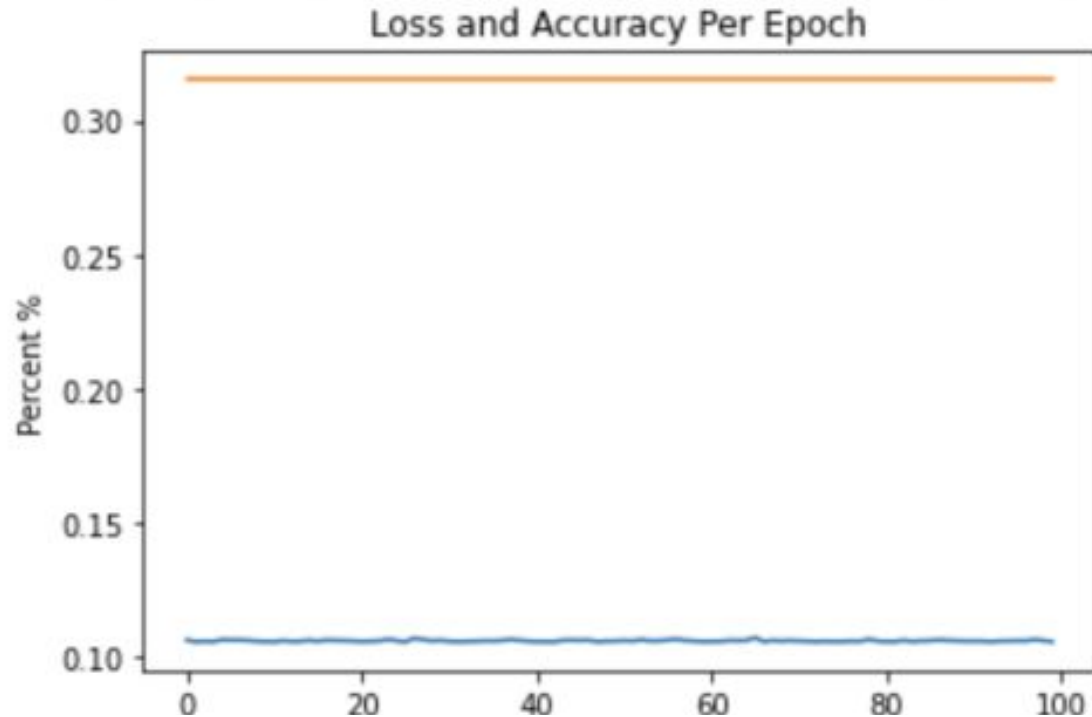


Figure: The plot to the left describes the model's accuracy in orange and loss in blue for each epoch when training. From this we can infer overfitting of the model and bottlenecks to the model, which we believe to be the immediate compression of the data from 207 dim -> 50 dim and 40 dim -> 1 dim for specific areas in the model

Results

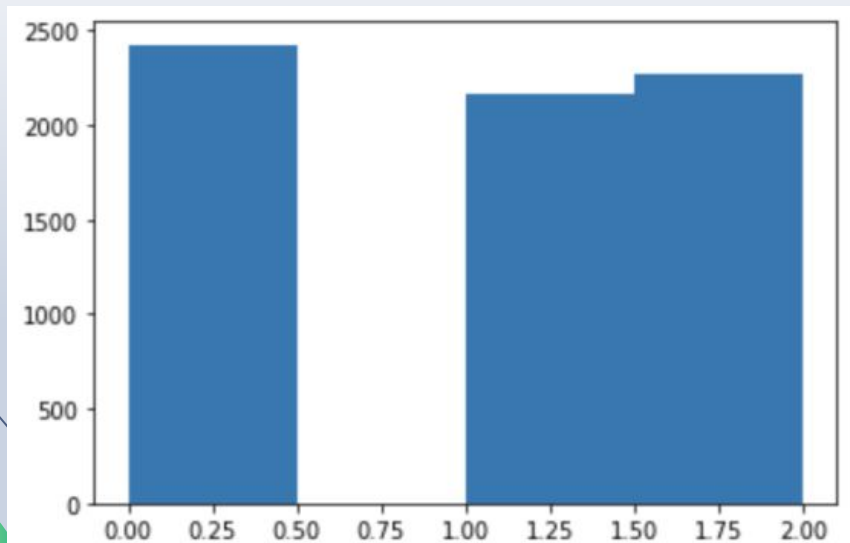


Figure: The histogram above shows a fair distribution of our training labels for our set

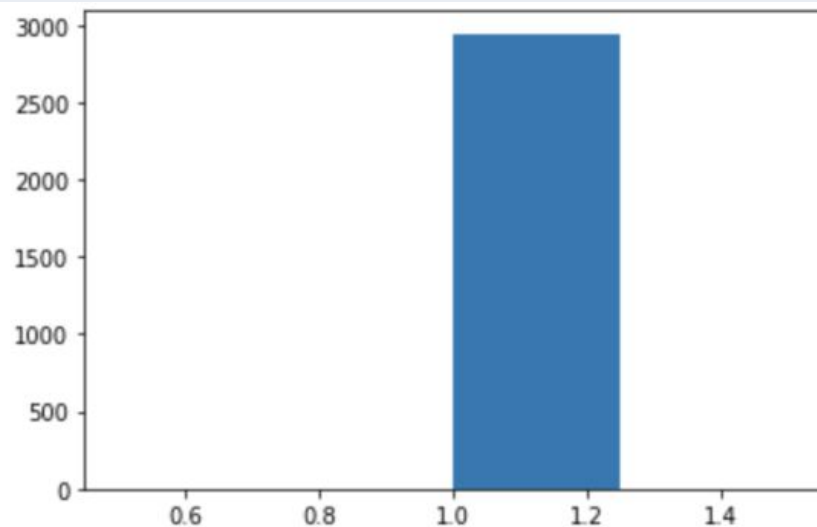


Figure: The histogram above shows an unfair distribution of the testing labels. Later realized the labels had to be self assigned.

5. Closing



Additional Work: Weaknesses

A decorative header featuring a black sine wave that oscillates across the top of the slide. Various colorful shapes, including triangles and circles, are scattered along the wave. Some of these shapes are filled with patterns like vertical stripes or dots. The colors include red, blue, orange, green, and purple.

“

- We reduced the testing data to English exclusively so our solution won't work on other languages.
- In writing there are areas between neutral and truth and contradiction that we cannot measure precisely
- As with many other machine learning problems, our data results should be improved with more training data.



LSTM model

<https://github.com/buomsoo-kim/Easy-deep-learning-with-Keras/blob/master/3.%20RNN/4-Advanced-RNN-3/4-1-lstm.py>

<https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>

Repo

https://github.com/huda-irs/cs542_NLP_Project.git

Paper on Overleaf

<https://www.overleaf.com/project/60da5bed106cd940c2d2f965>