RELATAS AI ML ASSIGNMENT

PROBLEM STATEMENT:

The company management has a business need to forecast the revenue of the company over the current month, quarter and coming quarters and year for financial planning and reporting.

The following forecasts are required:

- Forecast for the current month.
- Forecast for the current Quarter.
- Forecast for the current Financial Year
- Back Testing with accuracy for past known outcome data

OBSERVATIONS FROM DATA ANALYSIS:

1. The close values for products show no patterns or dependency on previous values.
2. Every sale has been classified into 3 categories (Won, Lost, In Progress). Where Won represents that the close value is to be considered towards the revenue, Lost sales will not be counted towards the total revenue and In Progress sales might or might not contribute.
3. The interactions between agents and contacts overlap for multiple sales and cannot be distinguished for individual sales.
4. All interactions involve only 181 texts.
5. Factors that might affect the outcome of stage of sale: Interactions, Date Difference, Product, Account, Close Value.

PROBLEM ANALYSIS:

- The revenue for a time period is to be calculated as follows:

  Total revenue = Close Values of 'Won' sales + Close Values of 'In Progress' sales that are predicted to be 'Won'.

- The interactions between agent and contact can be extracted by filtering the Agent_ID, Contact_ID, and the date when the interaction took place.
- These texts from interactions of a particular sale can be used to predict outcomes of an 'In Progress' sale.

PROPOSED SOLUTION:

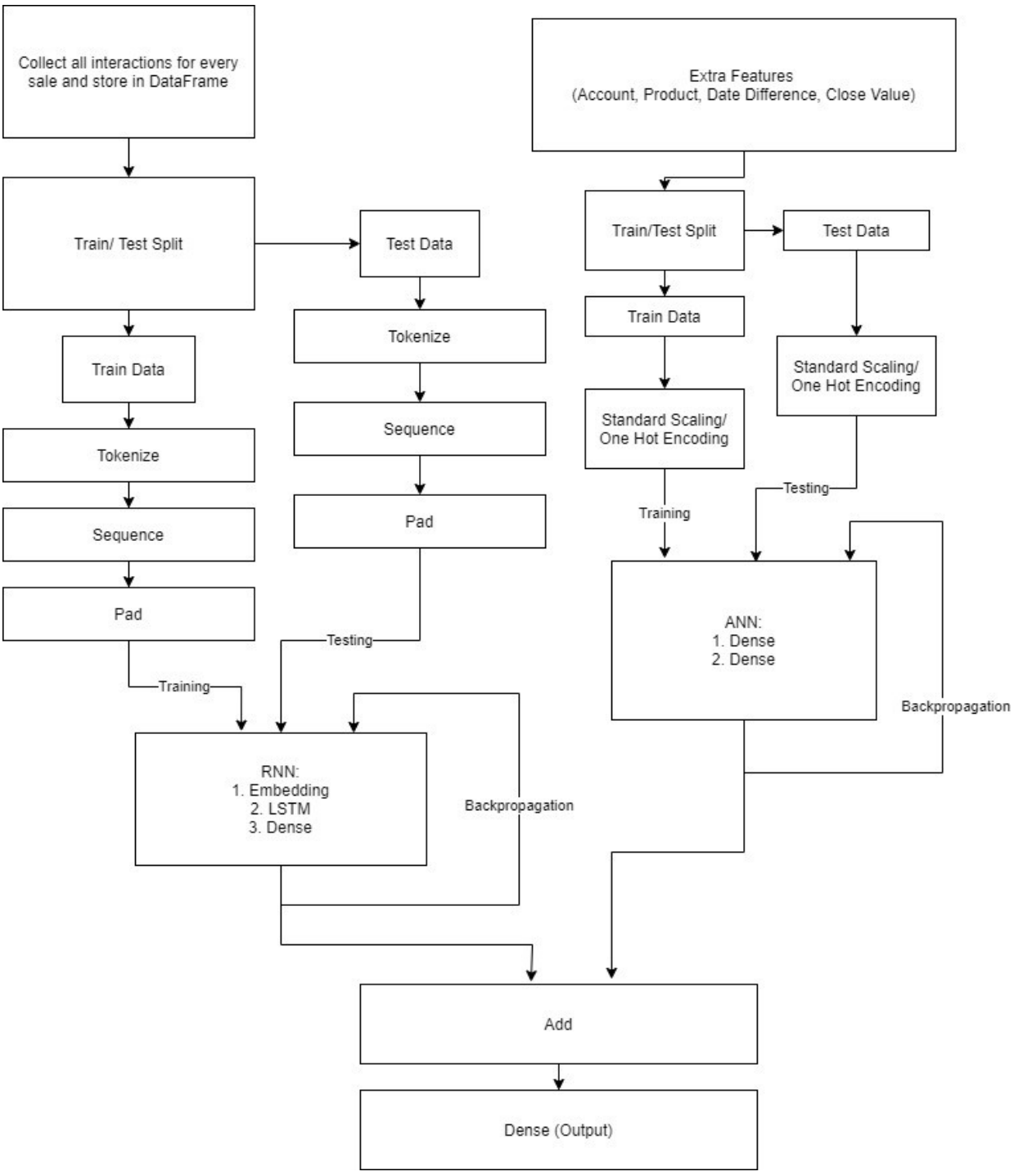
- An NLP Model is to be trained using extracted texts from every sale and the deal stage as labels.
- An ANN is to be trained with all the extra features about the sale.
- Outputs of the RNN & ANN to be concatenated to build a hybrid model.
- This model is to be used to predict the outcomes of the sales that are 'In Progress'.
- Finally, the predicted stages are used to calculate total close values for a given time period.

ASSUMPTIONS:

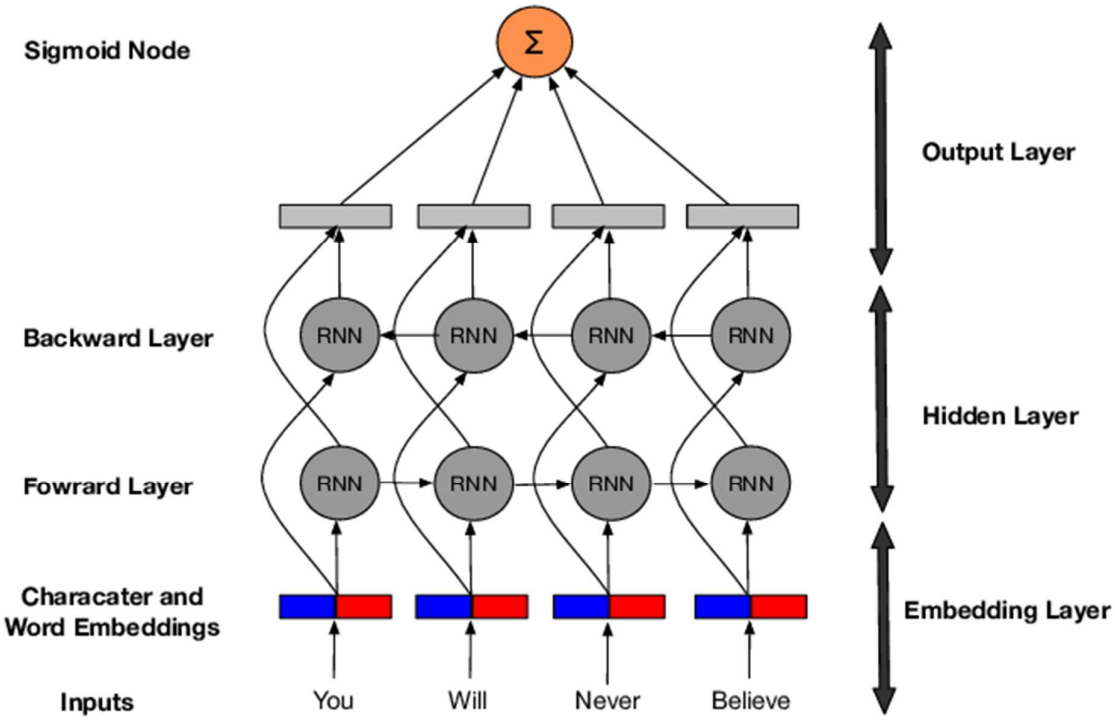
- Any interaction between the same Agent and Contact that takes place between the Create Date and Close Date pertains to that stage. (Even if the same interaction gets used for multiple sales).
- Every interaction is related to a sale and has an influence on the outcome.

COMPLETE PIPELINE:

1. A DataFrame will be created with all texts of a particular sale with details of that sale.
2. The texts will be processed (lowercase, remove stop words, remove punctuation, etc.)
3. Tokenize the texts and create a dictionary, Convert tokens to sequences, Pad sequences
4. These texts will pass through an RNN and the outputs will be added to the outputs of the ANN of extra features.
5. This added result will be given to a dense layer with SoftMax activation that will give us the predicted class.



RNN ARCHITECTURE:



FEATURE ENGINEERING & SELECTION:

1. I have assumed that the text interactions are a major factor having significant influence on the output (based on the given problem statement) and hence considered it for classification.
2. For continuous features like Date Difference and Close Value, a correlation plot with the Deal Stage variable will be plotted and selection will be done accordingly.
3. For categorical features like Account and Product, Chi Square Scores will be generated and selection will be done accordingly.

ALTERNATE ALGORITHMS FOR TEXT CLASSIFICATION:

1. KNN
2. Naïve Bayes
3. SVC
4. Random Forrest

TECHNOLOGIES TO BE USED:

1. Google Colab
2. Pandas, Numpy, MatPlotLib
3. Scikit-Learn
4. TensorFlow, Keras

Name: Kaustubh Sharma
E-Mail: sharma.kaustubh99@gmail.com
Date: 14 November 2020