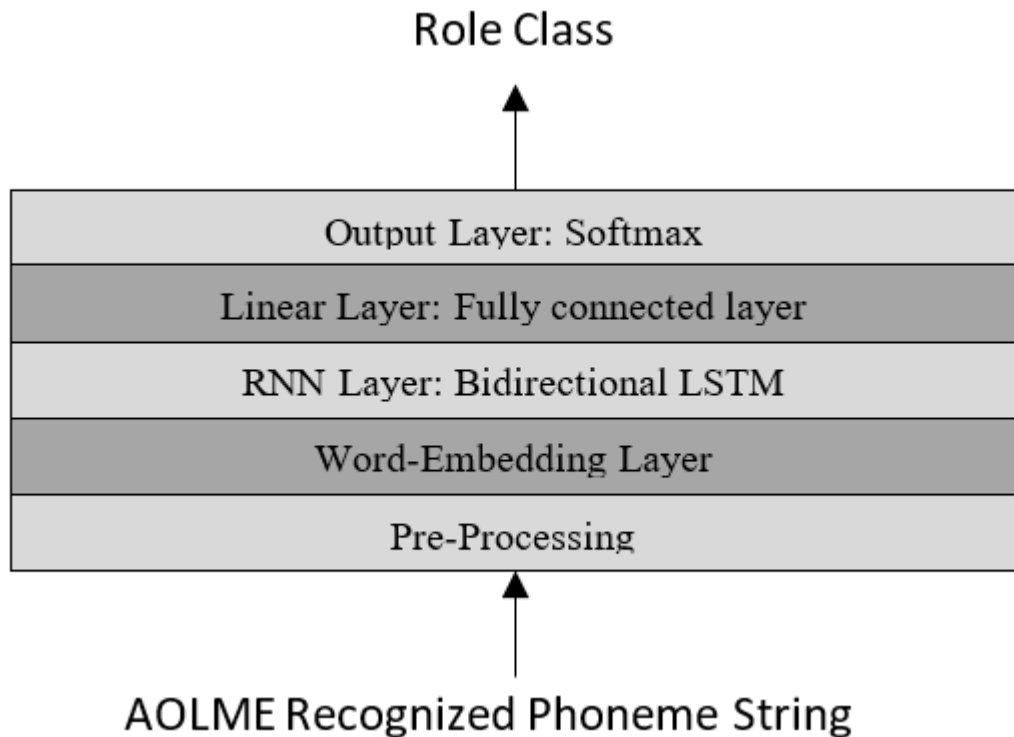


Roles Classifier

The model implemented here follows the following architecture diagram:



We start with the import declarations:

```
In [31]: from collections import Counter
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from torch.utils.data import Dataset, DataLoader
from BiLstmClassifier import BiLstmFixedLength, BiLstmVariableLength, BiLstmClassifier

import numpy as np
import pandas as pd
import spacy
import torch
import torch.nn.functional as F
```

0. Hyper Parameters Definition

Defining the correct hyper-parameters to refine the performance of the model is very important. We must consider some situations around the values to be assigned.

0.1 Epochs

We must define a correct number of epochs to allow the model to learn from the dataset. If this is too few, the model will not learn anything, if the value is too high, we can fall into over-fitting or wasting processing time where the model can't learn anymore.

0.2 Batch Size

It defines the amount of data for each batch to be used on each training epoch. Because we have no much data, the value here is small.

0.3 Embedded Layer Dimension

Word embeddings are always around 50 and 300 in length, longer embedding vectors don't add enough information and smaller ones don't represent the semantics well enough. For this model, we are using small sentences for most of the cases.

0.4 Hidden Layer Dimension

This parameter represents how complex is the language used in the conversations. For example, if the sentences belong to a literature book from Shakespeare, it probably will use a sophisticated language, and we can assign a value of 70. On the other hand, if the sentences belong to simple chat talking about movies, it is maybe simpler, and we can assign a value of 30.

```
In [32]: EPOCHS = 100
BATCH_SIZE = 5 # Small batches because the dataset is not bigger than 1000
HIDDEN_LAYER_DIM = 60 # AOLME is not too complex language, it represents simple chat
EMBEDDED_LAYER_DIM = 50
```

1. Load dataset

We are going to use the dataset generated by the Jupyter Notebook ["AOLME Datasets Generator" \(main.ipynb\)](#).

```
In [33]: roles = pd.read_csv('output/balanced_372.csv')
print(f'Dataset Size: {roles.shape}\n')
print(roles.head())
```

Dataset Size: (372, 2)

	Role	Text
0	Student	you like how its like
1	Student	its like youre obsessed with
2	Student	and this one
3	Student	no
4	Student	i dont like summary

2. Pre-Processing

2.1. Mapping 'Roles' labels to numbers for vectorization

```
In [34]: mapping = {'Student': 0, 'Co-Facilitator': 1, 'Facilitator': 2}
roles['Role'] = roles['Role'].apply(lambda x: mapping[x])
roles.head()

# Load English words model package
tok = spacy.load('en')

def tokenize(text: str):
    """
    This method tokenizes a sentence, considering the text is already lower
    ASCII, and punctuation has been removed
    :param text: The sentence to be tokenized
    :return: A list containing each word of the sentence
    """
    return [token.text for token in tok.tokenizer(text)]
```

2.2. Dataset cleaning and Sentence Vectorizing

```

In [35]: # Count number of occurrences of each word
counts = Counter()
for index, row in roles.iterrows():
    counts.update(tokenize(row['Text']))

# Deletes words appearing only once
print(f'Number of Words before cleaning: {len(counts.keys())}')
for word in list(counts):
    if counts[word] < 2:
        del counts[word]
print(f'Number of Words after cleaning: {len(counts.keys())}\n')

# Creates vocabulary
vocab2index = {'': 0, 'UNK': 1}
words = ['', 'UNK']
for word in counts:
    vocab2index[word] = len(words)
    words.append(word)

def encode_sentence(text, vocabulary_map, n=70):
    """
    Encodes the sentence into a numerical vector, based on the vocabulary
    :param text: The sentence
    :param vocabulary_map: A map assigning a number to each word in the
    :param n: Required vector size
    :return: Vectorized sentence and length
    """
    tokenized = tokenize(text)
    vectorized = np.zeros(n, dtype=int)
    enc1 = np.array([vocabulary_map.get(w, vocabulary_map["UNK"]) for w in tokenized])
    length = min(n, len(enc1))
    vectorized[:length] = enc1[:length]
    return vectorized, length

# Creates a new column into Dataset: each sentence expressed as a numerical vector
roles['Vectorized'] = roles['Text'].apply(lambda x: np.array(encode_sentence(x, vocab2index)))
print(roles.head())

```

Number of Words before cleaning: 662

Number of Words after cleaning: 347

	Role	Text \	Vectorized
0	0	you like how its like	[2, 3, 4, 5, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
1	0	its like youre obsessed with	[5, 3, 2, 1, 1, 6, 0, 0, 0, 0, 0, 0, 0, 0, ...]
2	0	and this one	[7, 8, 9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
3	0	no	[10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]
4	0	i dont like summary	[11, 12, 13, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...]

```
<ipython-input-35-5b2887f873d4>:38: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a list-or-tuple of lists-or-tuples-or ndarrays with different lengths or shapes) is deprecated. If you meant to do this, you must specify 'dtype=object' when creating the ndarray
```

```
roles['Vectorized'] = roles['Text'].apply(lambda x: np.array(encode_sentence(x, vocab2index)))
```

Check if the dataset is balanced

```
In [36]: Counter(roles['Role'])
```

```
Out[36]: Counter({0: 124, 1: 124, 2: 124})
```

2.3 Split into training and validation partitions

```
In [37]: X = list(roles['Vectorized'])
y = list(roles['Role'])
X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size=0.1)

class RolesDataset(Dataset):
    """
    Simple PyTorch Dataset wrapper defined by an array of vectorized sentences
    """
    def __init__(self, input_x, input_y):
        self.X = input_x
        self.y = input_y

    def __len__(self):
        return len(self.y)

    def __getitem__(self, idx):
        return torch.from_numpy(self.X[idx][0].astype(np.int32)), self.y[idx]

training_ds = RolesDataset(X_train, y_train)
validation_ds = RolesDataset(X_valid, y_valid)
```

2.4 Training and Validation Functions

```

In [38]: def train_model(input_model, epochs=10, lr=0.001, verbose=True):
        """
        Trains the input model
        :param verbose: Prints each batch iteration
        :param input_model: Input Model
        :param epochs: The number of training epochs
        :param lr: Learning Rate
        :return: training loss, validation loss, validation accuracy, and val
        """

        parameters = filter(lambda p: p.requires_grad, input_model.parameters)
        optimizer = torch.optim.Adam(parameters, lr=lr)

        for i in range(epochs):
            input_model.train()
            sum_loss = 0.0
            total = 0

            # Iterates on Training DataLoader
            for x, y, l in training_dl:
                x = x.long()
                y = y.long()
                y_pred = input_model(x, l)
                optimizer.zero_grad()
                loss = F.cross_entropy(y_pred, y)
                loss.backward()
                optimizer.step()
                sum_loss += loss.item() * y.shape[0]
                total += y.shape[0]

            val_loss, val_acc, val_rmse = get_metrics(input_model, validation_dl)

            if verbose and (i + 1) % 20 == 1:
                print(f"Epoch {i}: training loss %.3f, valid. loss %.3f, valid. accuracy %.3f, val
                    sum_loss / total, val_loss, val_acc, val_rmse))

        print(f"FINAL: training loss %.3f, valid. loss %.3f, valid. accuracy %.3f, val
            sum_loss / total, val_loss, val_acc, val_rmse))

        return sum_loss / total, val_loss, val_acc, val_rmse

def get_metrics(input_model, valid_dl):
    """
    Obtains current validation metrics
    :param input_model: Input Model
    :param valid_dl: Validation PyTorch DataLoader
    :return:
    """

    input_model.eval()
    correct = 0
    total = 0
    sum_loss = 0.0
    sum_rmse = 0.0

    # PyTorch uses CrossEntropy function to implement Softmax on the same
    for x, y, l in valid_dl:

```

```
x = x.long()
y = y.long()
y_hat = input_model(x, l)
loss = F.cross_entropy(y_hat, y)
pred = torch.max(y_hat, 1)[1]
correct += (pred == y).float().sum()
total += y.shape[0]
sum_loss += loss.item() * y.shape[0]
sum_rmse += np.sqrt(mean_squared_error(pred, y.unsqueeze(-1)))
return sum_loss / total, correct / total, sum_rmse / total
```

```
In [39]: vocab_size = len(words)
training_dl = DataLoader(training_ds, batch_size=BATCH_SIZE, shuffle=True)
validation_dl = DataLoader(validation_ds, batch_size=BATCH_SIZE)
```

BiLSTM - Fixed Length Input

We can see the implemented model class in [BiLstmClassifier class \(BiLstmClassifier.py\)](#).

BiLstmFixedLength has the following features:

- Word-Embedding Layer. # Embeddings: Vocabulary Size, Embeddings size: 50
- Bi-directional LSTM Layer. Input size: 50, Hidden size: 60
- Linear Layer. Fully connected layer, Input size: 60, 3 output features (roles)
- Dropout: 0.7
- Fixed Length Input (see [encode_sentence](#) function)

```
In [40]: model_fixed = BiLstmFixedLength(vocab_size, EMBEDDED_LAYER_DIM, HIDDEN_L
print(f'\nBiLSTM - Fixed Length: {EPOCHS} epochs, Learning Rate: 0.1')
print('=====')
train_model(model_fixed, epochs=EPOCHS, lr=0.1)
print(f'\nBiLSTM - Fixed Length: {EPOCHS} epochs, Learning Rate: 0.05')
print('=====')
train_model(model_fixed, epochs=EPOCHS, lr=0.05)
print(f'\nBiLSTM - Fixed Length: {EPOCHS} epochs, Learning Rate: 0.01')
print('=====')
train_model(model_fixed, epochs=EPOCHS, lr=0.01)
```

BiLSTM - Fixed Length: 100 epochs, Learning Rate: 0.1

```
=====
Epoch 0: training loss 1.576, valid. loss 1.693, valid. accuracy 0.29
3, and valid. RMSE 1.157
Epoch 20: training loss 1.515, valid. loss 1.574, valid. accuracy 0.38
7, and valid. RMSE 1.031
Epoch 40: training loss 1.668, valid. loss 1.701, valid. accuracy 0.41
3, and valid. RMSE 0.996
Epoch 60: training loss 1.512, valid. loss 1.237, valid. accuracy 0.37
3, and valid. RMSE 1.070
Epoch 80: training loss 1.477, valid. loss 1.821, valid. accuracy 0.36
0, and valid. RMSE 1.029
FINAL: training loss 1.453, valid. loss 1.917, valid. accuracy 0.373,
and valid. RMSE 1.128
```

BiLSTM - Fixed Length: 100 epochs, Learning Rate: 0.05

```
=====
Epoch 0: training loss 1.306, valid. loss 1.541, valid. accuracy 0.44
0, and valid. RMSE 1.059
Epoch 20: training loss 1.115, valid. loss 1.260, valid. accuracy 0.45
3, and valid. RMSE 0.972
Epoch 40: training loss 1.173, valid. loss 1.299, valid. accuracy 0.45
3, and valid. RMSE 1.038
Epoch 60: training loss 1.126, valid. loss 1.232, valid. accuracy 0.45
3, and valid. RMSE 0.969
Epoch 80: training loss 1.122, valid. loss 1.349, valid. accuracy 0.40
0, and valid. RMSE 1.038
FINAL: training loss 1.146, valid. loss 1.503, valid. accuracy 0.387,
and valid. RMSE 1.130
```

BiLSTM - Fixed Length: 100 epochs, Learning Rate: 0.01

```
=====
Epoch 0: training loss 1.047, valid. loss 1.394, valid. accuracy 0.37
3, and valid. RMSE 1.117
Epoch 20: training loss 0.906, valid. loss 1.169, valid. accuracy 0.41
3, and valid. RMSE 1.061
Epoch 40: training loss 0.914, valid. loss 1.267, valid. accuracy 0.41
3, and valid. RMSE 1.138
Epoch 60: training loss 0.861, valid. loss 1.199, valid. accuracy 0.45
3, and valid. RMSE 1.072
Epoch 80: training loss 0.863, valid. loss 1.211, valid. accuracy 0.42
7, and valid. RMSE 1.067
FINAL: training loss 0.860, valid. loss 1.267, valid. accuracy 0.400,
and valid. RMSE 1.115
```


Out[40]: (0.859525477243995, 1.2670233289400736, tensor(0.4000), 1.114778289189035)

BiLSTM - Variable Length Input

We can see the implemented model class in [BiLstmClassifier class \(BiLstmClassifier.py\)](#).

BiLstmVariableLength has the following features:

- Word-Embedding Layer. # Embeddings: Vocabulary Size, Embeddings size: 50
- Bi-directional LSTM Layer. Input size: 50, Hidden size: 60
- Linear Layer. Fully connected layer, Input size: 60, 3 output features (roles)
- Dropout: 0.7
- Variable Length Input. Uses PyTorch's [pack_padded_sequence](https://pytorch.org/docs/stable/generated/torch.nn.utils.rnn.pack_padded_sequence.html) (https://pytorch.org/docs/stable/generated/torch.nn.utils.rnn.pack_padded_sequence.html) to create sequences of variable length.

```
In [41]: model = BiLstmVariableLength(vocab_size, EMBEDDED_LAYER_DIM, HIDDEN_LAYER_DIM)

print(f'\nBiLSTM - Variable Length: {EPOCHS} epochs, Learning Rate: 0.1')
print('=====')
train_model(model, epochs=EPOCHS, lr=0.1)
print(f'\nBiLSTM - Variable Length: {EPOCHS} epochs, Learning Rate: 0.05')
print('=====')
train_model(model, epochs=EPOCHS, lr=0.05)
print(f'\nBiLSTM - Variable Length: {EPOCHS} epochs, Learning Rate: 0.01')
print('=====')
train_model(model, epochs=EPOCHS, lr=0.01)
```

BiLSTM - Variable Length: 100 epochs, Learning Rate: 0.1

=====

Epoch 0: training loss 1.518, valid. loss 1.593, valid. accuracy 0.360, and valid. RMSE 1.091
 Epoch 20: training loss 1.545, valid. loss 1.737, valid. accuracy 0.400, and valid. RMSE 0.969
 Epoch 40: training loss 1.463, valid. loss 1.413, valid. accuracy 0.467, and valid. RMSE 0.990
 Epoch 60: training loss 1.475, valid. loss 1.683, valid. accuracy 0.387, and valid. RMSE 0.958
 Epoch 80: training loss 1.641, valid. loss 1.434, valid. accuracy 0.333, and valid. RMSE 1.108
 FINAL: training loss 1.487, valid. loss 1.790, valid. accuracy 0.347, and valid. RMSE 1.071

BiLSTM - Variable Length: 100 epochs, Learning Rate: 0.05

=====

Epoch 0: training loss 1.327, valid. loss 1.527, valid. accuracy 0.333, and valid. RMSE 1.128
 Epoch 20: training loss 1.143, valid. loss 1.363, valid. accuracy 0.440, and valid. RMSE 0.996
 Epoch 40: training loss 1.097, valid. loss 1.462, valid. accuracy 0.373, and valid. RMSE 1.107
 Epoch 60: training loss 1.103, valid. loss 1.313, valid. accuracy 0.427, and valid. RMSE 1.034
 Epoch 80: training loss 1.012, valid. loss 1.311, valid. accuracy 0.440, and valid. RMSE 0.968
 FINAL: training loss 0.945, valid. loss 1.253, valid. accuracy 0.453, and valid. RMSE 1.021

BiLSTM - Variable Length: 100 epochs, Learning Rate: 0.01

=====

Epoch 0: training loss 1.017, valid. loss 1.262, valid. accuracy 0.467, and valid. RMSE 1.039
 Epoch 20: training loss 0.835, valid. loss 1.152, valid. accuracy 0.427, and valid. RMSE 1.034
 Epoch 40: training loss 0.815, valid. loss 1.267, valid. accuracy 0.387, and valid. RMSE 1.056
 Epoch 60: training loss 0.899, valid. loss 1.217, valid. accuracy 0.413, and valid. RMSE 1.089
 Epoch 80: training loss 0.848, valid. loss 1.189, valid. accuracy 0.400, and valid. RMSE 1.026
 FINAL: training loss 0.859, valid. loss 1.240, valid. accuracy 0.400, and valid. RMSE 1.075

Out[41]: (0.8594971093465182, 1.2398587961991627, tensor(0.4000), 1.0751769500291373)

BiLSTM - with pretrained GloVe Word Embeddings

We can see the implemented model class in [BiLstmClassifier class \(BiLstmClassifier.py\)](#).

BiLstmGloveVector has the following features:

- Word-Embedding Layer. # Embeddings: Vocabulary Size, Embeddings size: 50
- Bi-directional LSTM Layer. Input size: 50, Hidden size: 60
- Linear Layer. Fully connected layer, Input size: 60, 3 output features (roles)
- Dropout: 0.7
- Uses pretrained GloVe Word Embeddings to initialize weights based on its vocabulary.

```
In [42]: def load_glove_vectors():
    """Load the glove Global Vectors for Word Representation"""
    word_vectors = {}

    with open("./glove/glove.6B.50d.txt", encoding="utf8") as f:
        for line in f:
            split = line.split()
            word_vectors[split[0]] = np.array([float(x) for x in split[1:]])
    return word_vectors

def get_embedding_matrix(word_counts, emb_size=50):
    """ Creates embedding matrix from word vectors"""
    vocab_size = len(word_counts) + 2
    vocab_to_idx = {}
    vocab = ["", "UNK"]
    W = np.zeros((vocab_size, emb_size), dtype="float32")
    W[0] = np.zeros(emb_size, dtype='float32') # adding a vector for padding
    W[1] = np.random.uniform(-0.25, 0.25, emb_size) # adding a vector for padding
    vocab_to_idx["UNK"] = 1
    i = 2

    for word in word_counts:
        if word in word_vecs:
            W[i] = word_vecs[word]
        else:
            W[i] = np.random.uniform(-0.25, 0.25, emb_size)
            vocab_to_idx[word] = i
            vocab.append(word)
            i += 1
    return W, np.array(vocab), vocab_to_idx
```

```
In [43]: word_vecs = load_glove_vectors()
pretrained_weights, vocab, vocab2index = get_embedding_matrix(counts, EM

model = BiLstmGloveVector(vocab_size, EMBEDDED_LAYER_DIM, HIDDEN_LAYER_D

print(f'\nBiLSTM - with pretrained GloVe Word Embeddings: {EPOCHS} epoch
print('=====
train_model(model, epochs=EPOCHS, lr=0.1)
print(f'\nBiLSTM - with pretrained GloVe Word Embeddings: {EPOCHS} epoch
print('=====
train_model(model, epochs=EPOCHS, lr=0.05)
print(f'\nBiLSTM - with pretrained GloVe Word Embeddings: {EPOCHS} epoch
print('=====
train_model(model, epochs=EPOCHS, lr=0.01)
```

BiLSTM - with pretrained GloVe Word Embeddings: 100 epochs, Learning Rate: 0.1

```
=====
Epoch 0: training loss 1.317, valid. loss 1.128, valid. accuracy 0.41
3, and valid. RMSE 1.074
Epoch 20: training loss 1.036, valid. loss 1.335, valid. accuracy 0.2
80, and valid. RMSE 1.271
Epoch 40: training loss 1.208, valid. loss 1.439, valid. accuracy 0.3
33, and valid. RMSE 1.083
Epoch 60: training loss 1.190, valid. loss 1.397, valid. accuracy 0.3
87, and valid. RMSE 0.975
Epoch 80: training loss 1.216, valid. loss 1.493, valid. accuracy 0.3
07, and valid. RMSE 1.217
FINAL: training loss 1.128, valid. loss 1.808, valid. accuracy 0.373,
and valid. RMSE 0.986
```

BiLSTM - with pretrained GloVe Word Embeddings: 100 epochs, Learning Rate: 0.05

```
=====
Epoch 0: training loss 1.076, valid. loss 1.491, valid. accuracy 0.32
0, and valid. RMSE 1.010
Epoch 20: training loss 1.039, valid. loss 1.442, valid. accuracy 0.3
33, and valid. RMSE 1.020
Epoch 40: training loss 1.032, valid. loss 1.324, valid. accuracy 0.3
47, and valid. RMSE 0.991
Epoch 60: training loss 0.931, valid. loss 1.424, valid. accuracy 0.3
60, and valid. RMSE 1.167
Epoch 80: training loss 0.941, valid. loss 1.559, valid. accuracy 0.3
20, and valid. RMSE 1.030
FINAL: training loss 0.923, valid. loss 1.361, valid. accuracy 0.413,
and valid. RMSE 1.024
```

BiLSTM - with pretrained GloVe Word Embeddings: 100 epochs, Learning Rate: 0.01

```
=====
Epoch 0: training loss 0.969, valid. loss 1.324, valid. accuracy 0.40
0, and valid. RMSE 1.057
```

```
Epoch 20: training loss 0.828, valid. loss 1.346, valid. accuracy 0.347, and valid. RMSE 1.001
Epoch 40: training loss 0.851, valid. loss 1.301, valid. accuracy 0.360, and valid. RMSE 1.085
Epoch 60: training loss 0.858, valid. loss 1.337, valid. accuracy 0.307, and valid. RMSE 1.169
Epoch 80: training loss 0.796, valid. loss 1.273, valid. accuracy 0.373, and valid. RMSE 1.083
FINAL: training loss 0.807, valid. loss 1.467, valid. accuracy 0.333, and valid. RMSE 0.972
```

```
Out[43]: (0.8068194387838094, 1.4672222097714742, tensor(0.3333), 0.9717885474348189)
```

Testing Several Files

```

In [44]: @torch.no_grad()
def get_all_preds(model, loader):
    all_preds = torch.tensor([])
    for batch in loader:
        images, labels = batch

        preds = model(images)
        all_preds = torch.cat(
            (all_preds, preds)
            ,dim=0
        )
    return all_preds

def get_num_correct(preds, labels):
    return preds.argmax(dim=1).eq(labels).sum().item()

file_size = [150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 645]
accuracy_fixed = []
accuracy_variable = []
accuracy_glove = []

for i in file_size:
    BATCH_SIZE = int(i * 0.5)
    file_name = f'output/balanced_{i}.csv'
    roles = pd.read_csv(file_name)
    mapping = {'Student': 0, 'Co-Facilitator': 1, 'Facilitator': 2}
    roles['Role'] = roles['Role'].apply(lambda x: mapping[x])
    counts = Counter()

    for index, row in roles.iterrows():
        counts.update(tokenize(row['Text']))

    for word in list(counts):
        if counts[word] < 2:
            del counts[word]

    vocab2index = {'': 0, 'UNK': 1}
    words = ['', 'UNK']
    for word in counts:
        vocab2index[word] = len(words)
        words.append(word)

    roles['Vectorized'] = roles['Text'].apply(lambda x: np.array(encode_

X = list(roles['Vectorized'])
y = list(roles['Role'])
X_train, X_valid, y_train, y_valid = train_test_split(X, y, test_size=

training_ds = RolesDataset(X_train, y_train)
validation_ds = RolesDataset(X_valid, y_valid)

vocab_size = len(words)
training_dl = DataLoader(training_ds, batch_size=BATCH_SIZE, shuffle=True)
validation_dl = DataLoader(validation_ds, batch_size=BATCH_SIZE)

print('\n*****')

```

```

print(f'* Processing file: {file_name} *')
print('*****')

print(f'\nBiLSTM - Fixed Length Input')
print('=====')
model_fixed = BiLstmFixedLength(vocab_size, EMBEDDED_LAYER_DIM, HIDDEN_LAYER_DIM)
train_model(model_fixed, epochs=EPOCHS, lr=0.1, verbose=False)
train_model(model_fixed, epochs=EPOCHS, lr=0.05, verbose=False)
_, _, validation_accuracy, _ = train_model(model_fixed, epochs=EPOCHS, lr=0.05, verbose=False)
accuracy_fixed.append(validation_accuracy)

print(f'\nBiLSTM - Variable Length Input')
print('=====')
model_variable = BiLstmVariableLength(vocab_size, EMBEDDED_LAYER_DIM, HIDDEN_LAYER_DIM)
train_model(model_variable, epochs=EPOCHS, lr=0.1, verbose=False)
train_model(model_variable, epochs=EPOCHS, lr=0.05, verbose=False)
_, _, validation_accuracy, _ = train_model(model_variable, epochs=EPOCHS, lr=0.05, verbose=False)
accuracy_variable.append(validation_accuracy)

print(f'\nBiLSTM - with pretrained GloVe Word Embeddings')
print('=====')
word_vecs = load_glove_vectors()
pretrained_weights, vocab, vocab2index = get_embedding_matrix(counts, word_vecs)
model = BiLstmGloveVector(vocab_size, EMBEDDED_LAYER_DIM, HIDDEN_LAYER_DIM, pretrained_weights)
train_model(model, epochs=EPOCHS, lr=0.1, verbose=False)
train_model(model, epochs=EPOCHS, lr=0.05, verbose=False)
_, _, validation_accuracy, _ = train_model(model, epochs=EPOCHS, lr=0.05, verbose=False)
accuracy_glove.append(validation_accuracy)

FINAL: training loss 0.338, valid. loss 1.815, valid. accuracy 0.429,
and valid. RMSE 1.140
FINAL: training loss 0.282, valid. loss 2.056, valid. accuracy 0.514,
and valid. RMSE 0.956
FINAL: training loss 0.255, valid. loss 1.964, valid. accuracy 0.500,
and valid. RMSE 0.964

BiLSTM - with pretrained GloVe Word Embeddings
=====
FINAL: training loss 0.685, valid. loss 1.414, valid. accuracy 0.429,
and valid. RMSE 1.000
FINAL: training loss 0.444, valid. loss 1.755, valid. accuracy 0.500,
and valid. RMSE 0.964
FINAL: training loss 0.372, valid. loss 1.916, valid. accuracy 0.471,
and valid. RMSE 0.956

*****
* Processing file: output/balanced_400.csv *
*****

```

Graphical Performance Analysis

In the following plots we can see the how the model behaves when it is trained with different amounts of data.

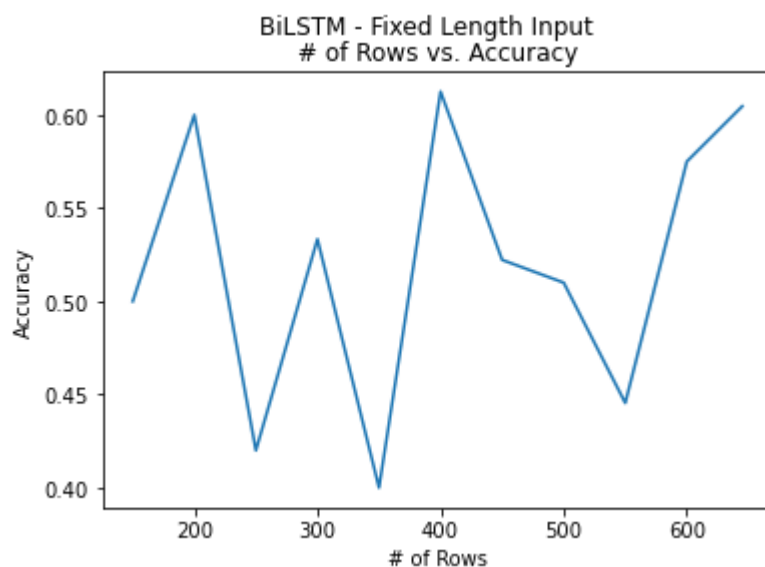
In [45]: %matplotlib inline

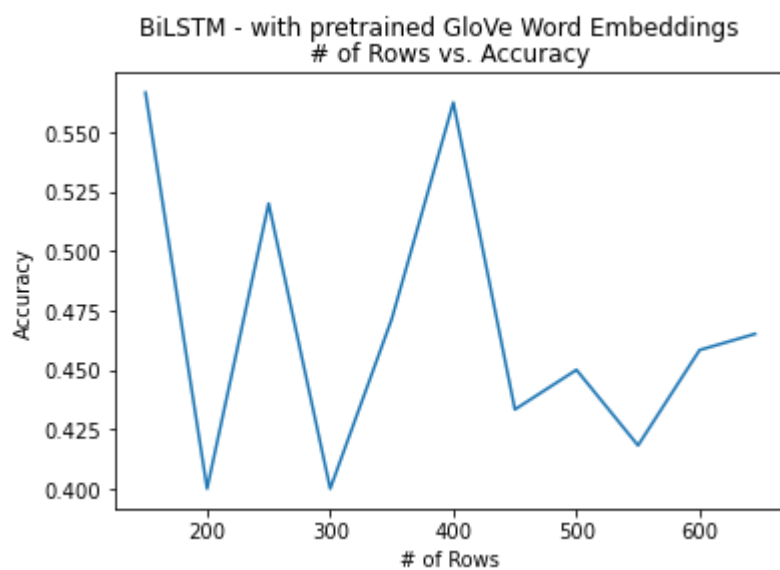
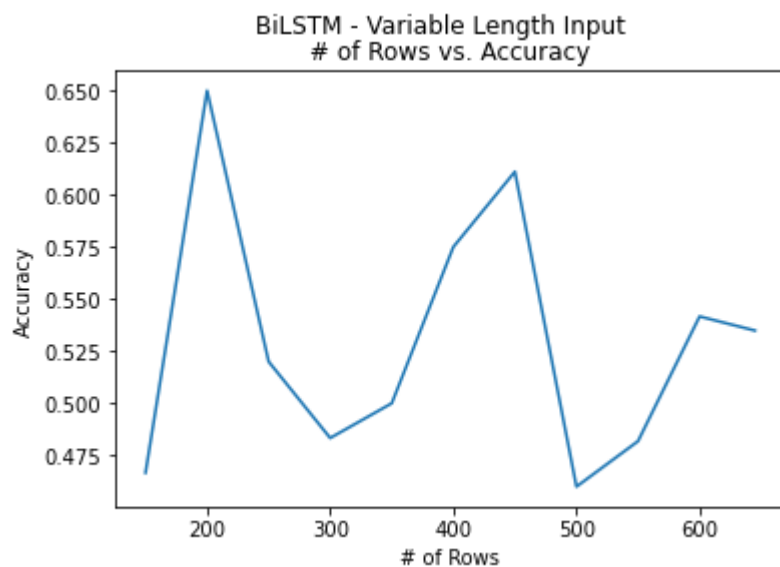
```
import matplotlib.pyplot as plt

plt.plot(file_size, accuracy_fixed)
plt.title('# of Rows vs. Accuracy')
plt.suptitle('BiLSTM - Fixed Length Input')
plt.xlabel('# of Rows')
plt.ylabel('Accuracy')
plt.show()

plt.plot(file_size, accuracy_variable)
plt.title('# of Rows vs. Accuracy')
plt.suptitle('BiLSTM - Variable Length Input')
plt.xlabel('# of Rows')
plt.ylabel('Accuracy')
plt.show()

plt.plot(file_size, accuracy_glove)
plt.title('# of Rows vs. Accuracy')
plt.suptitle('BiLSTM - with pretrained GloVe Word Embeddings')
plt.xlabel('# of Rows')
plt.ylabel('Accuracy')
plt.show()
```





Conclusions

- The model with the best performance is **BiLSTM - with pretrained GloVe Word Embeddings**, as we can see that it has a most stable performance as the model is trained with more data.
- The model reaches an approximated Accuracy of 60% for the selected model, and it will probably improve when it is trained with a bigger dataset.
- It is important to see that the first iteration with the smallest dataset with 150 rows reaches a high accuracy. This is happening because over-fitting caused by too few rows, but the accuracy starts to show a more real behavior with bigger datasets.

In []:

