December 6^{th}, 2023

PARAPHRASE IDENTIFICATION

Ala, Venkata Nithya - ala.v@northeastern.edu

Vura, Shravya - vura.s@northeastern.edu

Donepudi, Lakshmi Sowmya - donepudi.l@northeastern.edu

PROJECT OUTLINE

- This project aims to develop a paraphrase identifier using Long Short-Term Memory (LSTMs) networks, focusing on analyzing semantic relationships in sentences.

- The Paraphrase Identifier has implications in various NLP tasks like question-answering systems, text summarization, and plagiarism detection.

KEY CHALLENGE

- Current state-of-the-art models utilize computationally heavy transformers and attentive networks that require GPUs.

- We present a Bidirectional LSTM (BiLSTM) network that does not demand high computational capacities.

PROJECT APPROACH

- A deep bidirectional LSTM network with 2 BiLSTM layers that understand sentences by generating context vectors for each word was developed.

- BiLSTM layers made it easier to extract context from words that come before (preceding) and after (following) one another in a phrase.

CORPUS

Corpus Source	Microsoft Research Paraphrase Corpus
Content	Sentence pairs
Corpus size	5800 Records
Data Splitting	4076 Training Records, 1725 for test data

DATA PRE-PROCESSING

Tokenization	Build Vocabulary	Calculated IDF	Converting Words to Indices	Label Extraction
To standardize text formatting and filter out unnecessary characters	To list unique words used in the dataset	To assess their importance across the corpus	To transform words into numerical representations	To categorize sentence pairs based on their relationships.

MODEL ARCHITECTURE

Input Layer	Shared Embedding layer	Bidirectional LSTM Layers	Concatenation of LSTM outputs and Dense Layer	Output Layer
There are two input layers, one for each sentence in the pair	Converts the word indices into 50-dimensional dense vectors, capturing the semantic properties of the words	The first LSTM layer has 100 units, returns context vectors of words The second LSTM layer, has 50 units, returns context of sentence	Outputs from the second LSTM layer for both sentences are concatenated Combined data passes through a dense layer with 64 units and ReLU (Rectified Linear Unit) activation.	The final layer is a dense layer with one unit and a sigmoid activation function

SIAMESE NETWORK WITH BIDIRECTIONAL LSTM LAYERS

Key Features

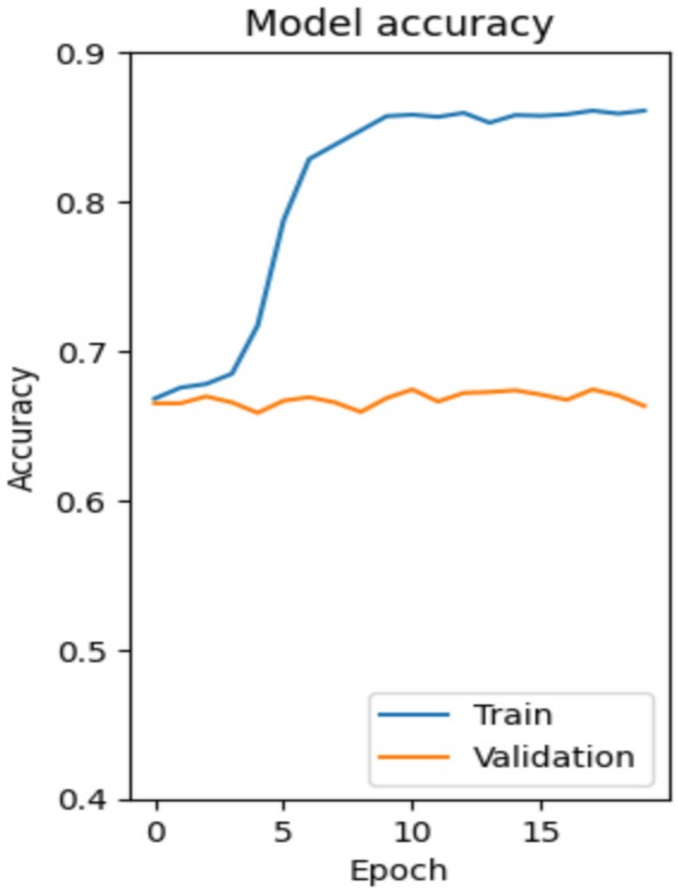
- Shared Embedding Layer
- Two Bidirectional LSTM Layers
- Dense Layer for Output Processing

Training Process

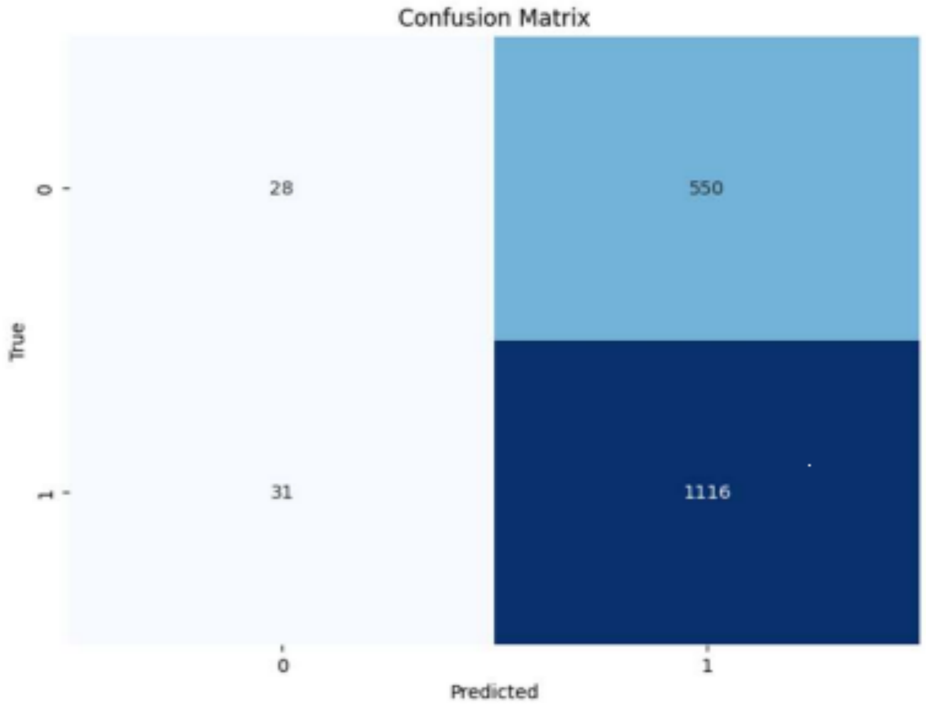
- Compiled with Adam Optimizer and Binary Crossentropy Loss
- Trained over 20 epochs with a batch size of 64

RESULTS

Evaluated on MSRP test set



Training and validation accuracy plot

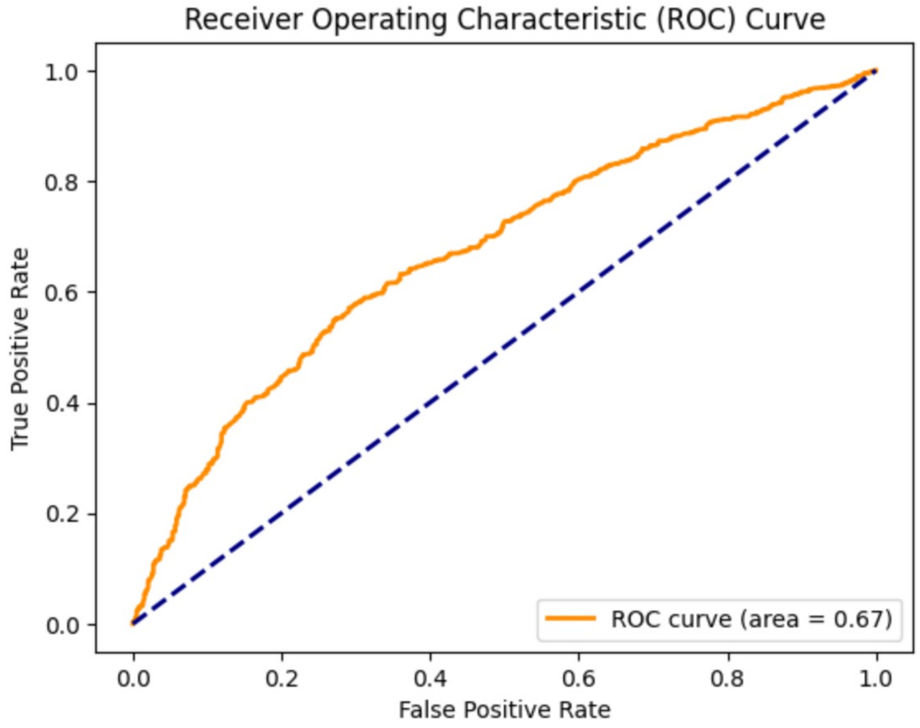


Confusion Matrix

RESULTS

Classification Report:

	precision	recall	f1-score	support
0	0.47	0.05	0.09	578
1	0.67	0.97	0.79	1147
accuracy			0.66	1725
macro avg	0.57	0.51	0.44	1725
weighted avg	0.60	0.66	0.56	1725



Comments: Fairly good model with high recall rate for paraphrases. Additional data required to improve performance.

EXPLORING EFFECTS OF L2 REGULARIZATION

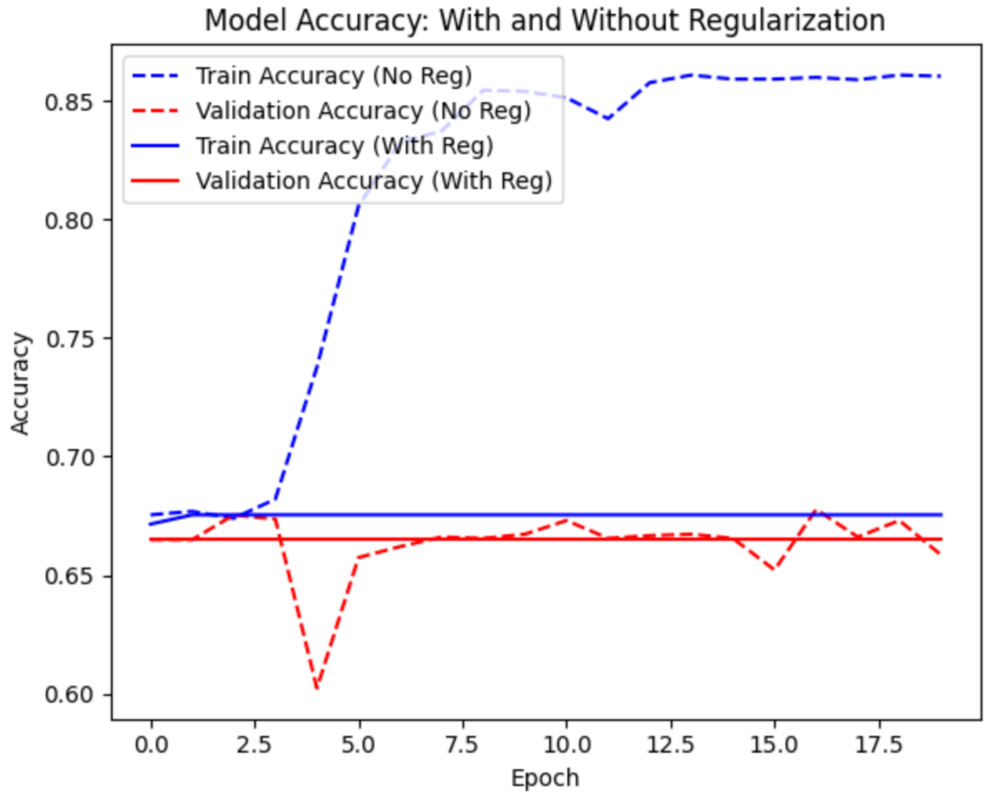
Key Features

- Shared Embedding Layer
- Two Bidirectional LSTM Layers with L2 regularization
- Dense Layer with L2 regularization

Training Process

- Compiled with Adam Optimizer and Binary Crossentropy Loss
- Trained over 20 epochs with a batch size of 64

L2 REGULARIZATION EFFECTS



Training and validation accuracy plot

CONCLUSION

- The model excels in identifying semantic similarities with high recall rates and commendable accuracy.

- While L2 regularization was explored to enhance generalization, it was found to be non-essential for this corpus.

- While the presented model shows promising capabilities, other research directions include exploring alternative regularization strategies and investigating additional data with rich linguistic features.

- The key challenge of identifying paraphrases without the need of high computational capacities is successfully addressed.

THANK YOU