# INTRODUCTION TO NLP
## Interim Report Submission

**Team No.** - 59

**Team Name -** Team API

**Team Members**

1. Amogha Halhalli (Roll No. - 2021101007)
2. Ishit Bansal (Roll No. - 2021101083)
3. Pranav Gupta (Roll No. - 2021101095)

## Introduction

As a part of the Interim Project Submission, we trained embeddings of the words in the dataset against several models, including the Word2Vec Model and BERT Model. Then, we evaluated the Performance of the Model using the Pearson Correlation Function to compare the various embeddings which are generated by different Models which would at the end, lead to generation of higher semantic similarity scores.

## Datasets used

1. **SemEval Datasets (Training, Validation and Testing Sets)**

   **About the Dataset** - All the Files are in the form of .jsonl extension. Data is loaded from each of the files, train.jsonl, val.jsonl, test.jsonl and then data is cleaned from each of the files. The Training Set consists of 5749 pairs of sentences, the validation set consists of 1500 pairs of sentences while the testing set consists of 1379 such samples. The sentences are then broken down to words and converted to embeddings so that these can be used later for comparison of the 2 sentences.

# Exploratory Data Analysis

## Tokenization:

Applied following preprocessing steps as part of tokenization:

1. Removed punctuation
2. Replaced numbers by ‘number’
3. Replaced URLs found by ‘url’
4. Replaced Hashtags by ‘hashtag’
5. Replaced Emails by ‘email’
6. Replaced Mentions by ‘mention’
7. Converted Text to Lowercase
8. Split the Sentences into words
9. Removed Stopwords from text
10. Applied lemmatizer on the resultant words list
11. Finally used Stemming to obtain output

After performing these steps, the vocabulary was found to be 8258.

## Fine Tuning:

We explored various text representation methods and applied models like Bidirectional LSTM model and Attention Mechanism for better performance of Word2Vec and BERT Models. We evaluated the Pearson Correlation coefficient to examine the effects the process of fine-tuning has on the training of embeddings.

# Evaluation Metrics used

1. Pearson Correlation Coefficient

## Results obtained

| S. No. | Model Name | Optim Function | Train Data Pearson Coefficient | Validation Data Pearson Coefficient | Testing Data Pearson Coefficient |
|---|---|---|---|---|---|
| 1. | Word2Vec using Linear Regression | MSE | 0.47 | 0.16 | 0.19 |
| 2. | Word2Vec using GRU Regression | MSE | 0.75 | 0.40 | 0.37 |
| 3. | Word2Vec using BiLSTM Regression | MSE | 0.77 | 0.36 | 0.34 |
| 4. | Word2Vec using BiLSTM Attention | MSE | 0.77 | 0.37 | 0.37 |
| 5. | BERT untrained | None | 0.13 | -0.14 | 0.16 |
| 6. | BERT Fine-Tuned | MSE | 0.9 | 0.83 | 0.82 |