Team KLM

1. Beeraka Krupa Kiranmai - 2021201022
2. Padigala Lakshman Sai - 2021201069
3. NVSS Maneesh Gupta - 2021201041

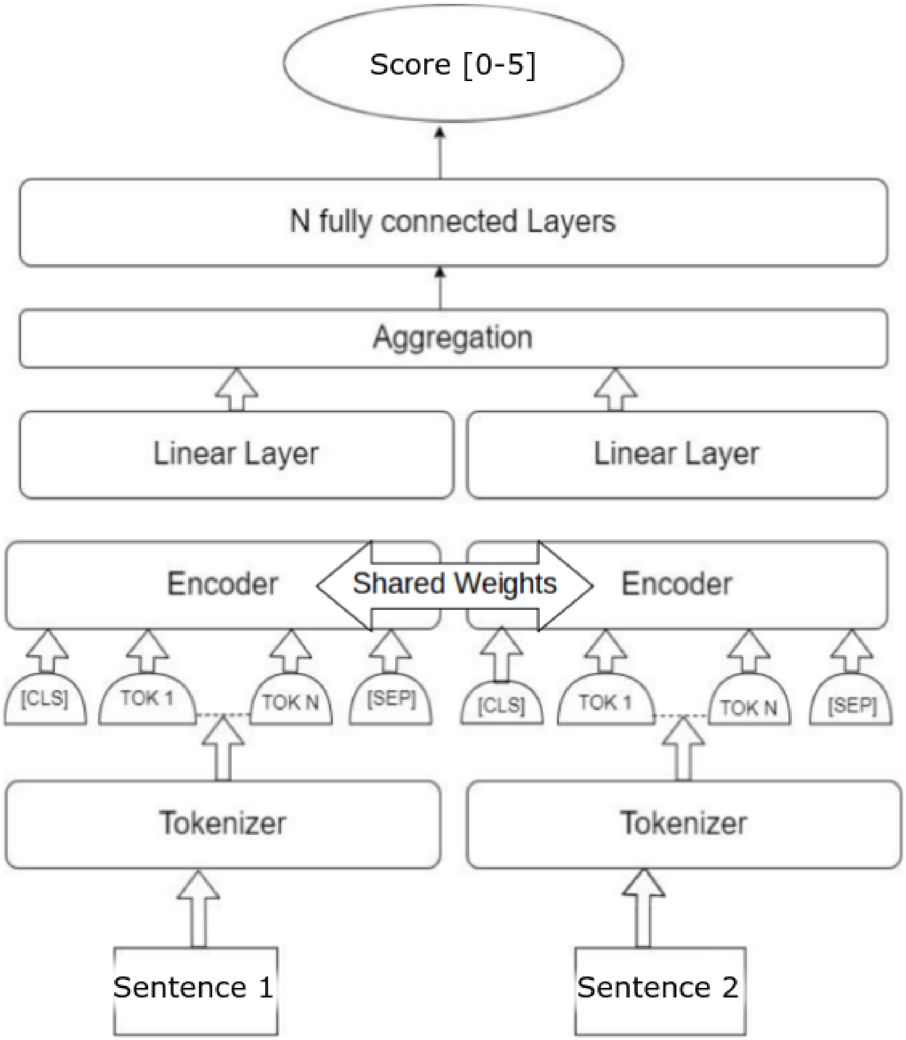
Semantic Textual Similarity

Intro to NLP

Work Done Till date:

1. Dataset Collection and preprocessing: In the Original paper for STS It is mentioned to use data from years 2012 to 2016 for training and in the year 2017 they have released 250 sentences with the scores corresponding to them. From the sources mentioned, we were able to collect the sentences and scores from each year and were able to combine them together with the help of small code snippets or sometimes manually going through them.
2. Built a base model to find similarity between the sentences using cosine similarity in two different ways, one is just by using bag-of-words as embeddings after some amount of preprocessing. And another one is by using bert-base-uncased pre-trained model for extracting embeddings for english-english pairs and then using cosine similarity for finding STS between pairs of sentences.
3. Another approach we came across during our literature survey is constructing a siamese architecture kind of thing for finding the similarity of the text. In this architecture we are using bert-base-uncased as the encoding layer. Here In this architecture after getting encodings of both the sentences using the same model and then adding a fully connected layer and then finding the final similarity score as a regression problem. We are unable to complete the following approach as we are getting some errors in code and we

will eventually try to solve them and move further. Our Idea is to use the same approach for pairs of english-spanish also we can use pre trained models like XLM-Roberta or DistilBert which serves for multilingual context.



4. Base models using cosine similarity have not shown any great results but helped us to understand the problem in depth.

Future Work

1. Complete the above approach for both english-english and spanish-english which is mentioned in 3rd point and check the scores with the original paper.
2. Select any other approach based on the results from the above experiments by using different pre-trained models and fix on some final model to implement.

In comparison to TimeLine:

As mentioned in the outline document , we are able to complete data collection and preprocessing. Also implemented Baseline models and started working on a different model for better results.