**CSCE 5290: Natural Language Processing**

**Project Proposal**

**Group 10**

**Title:** Language Understanding

Sentence to Sentence semantic similarity  
Can you identify question pairs that have the same intent or meaning?  
**Dataset:** Quora question pairs with similar questions marked

**Members:**

1. Saikiran reddy kondreddy

2. Rohith Sai kandelli

3. Harshith Nagula

4. Sai Hrithik Sanugoju

**GitHub link:** https://github.com/rohithsai0630/Natural-language-processing-.git

**1. Motivation**

The main aim of the project is to understand whether two proposed questions have the same meaning, even if they are rephrased in different words. This issue is quite common on platforms like Quora, customer support systems, and community forums, for example, as users want to know the meaning of something and they have two similar questions. This situation is quite annoying, as sometimes every third question in such systems can be reworded, so answers could have some similar features. The system aims to automatically determine whether the proposed questions are the same, so the question-answering system is less cumbersome, and the users can smoothly work with the question boards.

**2. Significance**

In natural language processing, the recognition of questions that are semantically similar has a wide range of uses in applications.

- With duplicate question detection, automated content moderation guides people toward existing solutions.

- For already-paid questions, virtual assistants can recognize different ways of asking the same question, thereby sparing repetition and saving people's time.

- Chatbots and QA System: Enhanced ability to understand user intent means that answers will be more accurate and more useful.

This project represents a step forward in terms of sentence-level semantic similarity. Contributions will be made to current techniques in NLP, such as BERT and Sentence-BERT which are widely used for context understanding in texts.

**3. Objectives**

1. Model Building: We will try to build an effective model that decides two or more questions are similarly equivalent.
2. Fine-tuning: Adapt BERT or Sentence-BERT fine-tuned models to better perform on the Quora question pairs dataset.
3. Evaluation Metrics: Brands use proven models, accurate, precise, recall, f1 value and maybe even cosine-similarity all over the world to analyze their product.
4. Scalability: Make sure that the model can process large datasets efficiently so it will work well across different contexts. Then we can bring it into reality in products and services for ordinary people to appreciate.

**4. Features**

The following key components will be delivered in this project:

1. Data Preprocessing Pipeline: A data preprocessing system that tokenizes, cleans and prepares the question pairs; it could handle missing data, make tokenization rather than transformations to a vector space, normalize the length of sentences.
2. Semantic Similarity Model: The transformer-based model, perhaps BERT or Sentence-BERT, is trained and fine-tuned to detect semantic similarity between questions.
3. Model Evaluation: we will be using standard measures like precision, recall, accuracy, F-1 score, cosine similarity will be used for model performance measurement doing work to ensure the detection accuracy of identical question pairs.

**Milestones:**

Phase 1: Data set overview, handling missing values, and pre-processing.

Phase 2: Model selection and training.

Phase 3: Model fine-tuning and performance testing. & Final Assessment & Report Generation.

**5. The Dataset**

The Quora Question Pairs Dataset will be used. Containing over 400,000 question pairs and annotations.

Size: 404,290 question pairs and 2,345,796 question pairs in the training and test set respectively.

Pair type: Each pair of questions must have a binary label 1 means the questions are similar, 0 in this context describes one question as non-similar.

Pre-processing: we will clean the dataset by remove missing values and special characters and tokenize it to ensure that our models receive normalized text inputs. We also employ padding and truncation so that all sequence lengths are of uniform length.

**6. Visualization**

A diagram of a process

Description automatically generated

*Figure 1 Workflow Description*

Data Collection: At this step, the Quora Question Pairs dataset is used to collect the dataset. The raw data is pairs of questions with their corresponding labels showing whether it's a duplicate or not.

Data Preprocessing: This involves cleaning and preparing the data for model training. Preprocessing includes handling missing data, tokenizing text and normalizing the question pairs.

Model Selection: Once data will be ready, we'll play around with performance-based machine training models (such as BERT or Sentence-BERTs). During this stage pre-trained models are fine-tuned

Training and Validation: The selected model will now start to be trained using the preprocessed data. To make sure the system works effectively on earlier untested test cases, the training data is split into two parts training and validation sets.

Evaluation: After training the data, the parameters like accuracy, precision, F1 score, recall and other important parameters are used to access the models performance. The models performance is improved by adjusting the hyperparameters.

Final Model: The model will be considered final and prepared for deployment or additional analysis if it has performed to a sufficient degree.

**References**

Reimers, N. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*. <https://fq.pkwyx.com/default/https/aclanthology.org/D19-1410.pdf>

Rani, S., Kumar, A., & Kumar, N. (2022). Eliminating data duplication in CQA platforms using deep neural model. *Computational Intelligence and Neuroscience*, *2022*(1), 2067449. <https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/2067449>

<https://www.kaggle.com/competitions/quora-question-pairs/overview>

<https://www.quora.com/>