**CSCE 5290: Natural Language Processing**

**Project Proposal**

Group 15: Nikita Lokhmachev, Paul Phillips, Likhitha Gullapalli, Leela Dodda

Project Name: Semantic Textual Similarity Comparison

• **Introduction**

In Natural Language Processing, the task of Semantic Textual Similarity plays a significant role as it has many applications in the field of AI. It has made a large contribution to search engines, anti-plagiarism systems, conversational and commercial chatbots as well as document clustering.1

However, while humans can tell two texts how similar two texts are based on their context, for computers, it is a difficult task. The similarity can be measured by comparing word-level and context-level relationships of texts or sentences. The former can be achieved relatively easily by incorporating heuristics and machine learning algorithms, whereas the latter requires supervised, self-supervised, or unsupervised deep learning models that require a lot of computational resources.

• **Motivation**

The number of heuristics, machine learning approaches, and deep learning approaches that can be combined in different ways to achieve reasonable and satisfactory results is practically limitless, which makes the problem of text similarity a large field of research. Therefore, it is vital to compare several existing methods with each other and potentially come up with our own by combining them.2

• **Significance**

This project has significance in labeling and organizing text documents. If a program can read in multiple text documents and recognize which ones are making similar points, then it can organize these documents under a single label. Recognizing textual similarity is also significant in search engines. If a person wants to use a search engine to look for the answer to a question, then textual similarity can be used to identify documents will answer it, even if the documents word the question in a different way. This project has significance in improving the capabilities of conversational chatbots. People in ordinary conversations can say the same thing in multiple different ways. A program that is made to simulate a conversation will need to be able to do this if the same point needs to be made more than once. Repeating the same point in the same way would make the conversation seem mechanical, so the ability to do a semantically similar response with different words would make the conversation feel more authentic.

• **Objectives**

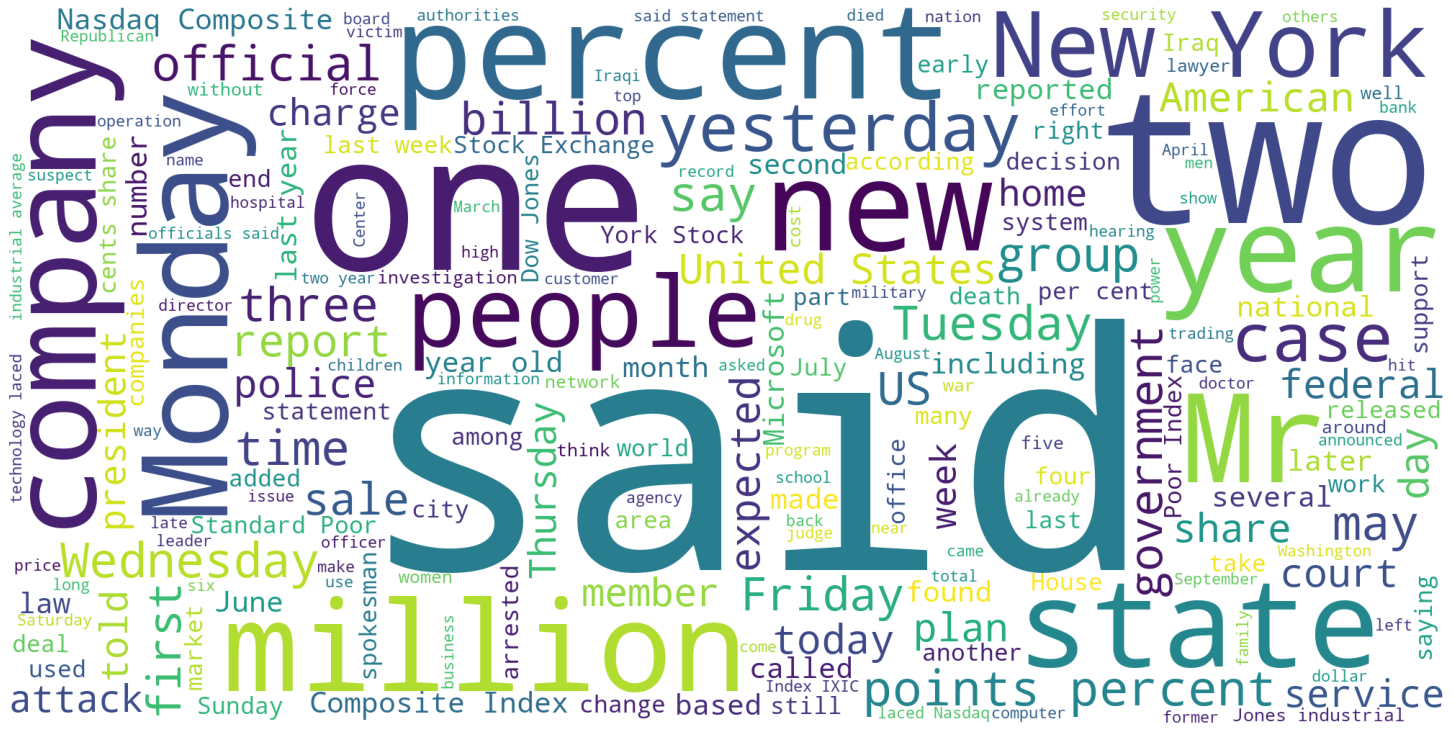
The main goal of text similarity models is to compute how close two texts are in terms of their meaning (semantic similarity) and the words they use (lexical similarity).

Our team is planning to split into two groups: those who use standard machine learning approaches and those who incorporate deep learning algorithms. Ultimately, since there are 4 members in our team, we should get 3-4 models and try to stack them together to achieve even better results.

• **Features**

We chose a relatively small dataset called Microsoft Research Paraphrase Corpus.3 It consists of 5800 pairs of sentences that have been extracted from news sources on the web. Each pair is manually labelled as either having paraphrased or semantic equivalence. Since the dataset is small, it will be easy to test different models on.

To analyze and visualize the information in the dataset, we created a wordcloud of the combination of all texts in it.



• **GitHub**

The project github link is: <https://github.com/n1kkqt/nlp_text_similarity>

• **References**

[1] Ganesan, K. (2021). What is text similarity? *The Business Case for AI.* Retrieved from <https://kavita-ganesan.com/what-is-text-similarity/#.YUTYsbhKhPY>

[2] Sieg, A. (2018, Jul 4). Text Similarities: Estimate the degree of similarity between two text. *Medium.* Retrieved from <https://medium.com/@adriensieg/text-similarities-da019229c894>

[3] (2005, Mar 3). Microsoft Research Paraphrase Corpus. *Microsoft*. <https://www.microsoft.com/en-us/download/details.aspx?id=52398>