GROUP 4 FINAL PROJECT PROPOSAL

Group Members:

- Alexis Kaldany
- Ricardo Diaz
- Sagar Tripathi

Problem:

Our project examines COVID articles and tries to classify them, there are large quantities of covid articles which can be overwhelming if a person is trying to look for a specific article in a category (business, tech, etc). We are looking to simplify the search for articles by categories plus comparing accuracy from two NLP methods to evaluate which works better to classify articles.

Dataset:

COVID News Articles (2020 - 2022) | Kaggle

The dataset encapsulates approximately half a million news articles collected over a period of 2 years during the Coronavirus pandemic onset and surge. It contains several articles with the headline, content, and their respective category.

NLP Method:

There are two main NLP methods we are using.

1. Classical Classification Model
2. Seq2seq Model

Packages:

Pandas, Torch, Transformers (DistilBert), HuggingFace.

We need pandas to read, and manipulate the dataset, torch to use neural networks for our modes, and HuggingFace to use summarization for the Seq2seq Model

Performance:

We will evaluate and compare the results based on their accuracy and their confusion matrix.

Time Frame for the Project:

NLP Final Project — Project Schedule

Start Week	Nov 1, 2022

Week	1	2	3	4	5	6	7	Notes
Starting	Nov 1	Nov 8	Nov 15	Nov 22	Nov 29	Dec 6	Dec 12	
EDA and Research								All members
Modeling								All members
Evaluate Results and Reports								All members