**PROJECT PROPOSAL: News Classification and Summarization**

**1. Vision and Goal**

1.1. What problems are you going to be tackling on your project?

The project is an application for news classification and news summarization. There are two specific services. One service will classify a news article into various categories like sports, politics etc. Another service is to summarize news articles. If a user wants to get news summary on let’s say sports, he can select the sports category and he will be provided with latest news summary on sports from various sources. This is not primarily multi-documents summarization but extractive summarization of individual news articles that falls under the sports category. User also has an option to upload a news article to get just a summary of it. (So, if user do not want to read a lengthy news article on “covid vaccination”, they can simply upload the news article and get the summary of it as an output)

* 1. . Why is that an interesting/useful application of Data Science?

Extractive summarization is one of the challenging fields in data science and specifically Natural Language Processing. There are many algorithms available to summarize documents and it would be an interesting experiment to see how effective certain algorithm could be to for summarization. This project also has plenty of future scope which includes making this summarization service into a more generic service. It offers people to randomly pick up a news and classify it into certain categories as well, which gives them a perspective about a certain news article.

**2. Use of Data and Models**

What data and models are you envisioning training to address that?

**2.1.** For classification of news articles, we are looking into various supervised learning algorithms including logistic regression. For summarization, we want to explore TextRank algorithm. The advantage of using TextRank algorithm is that it is an unsupervised algorithm and hence there is no requirement for training the model. We can calculate the accuracy of our summarization by a metric called [ROUGE](https://www.freecodecamp.org/news/what-is-rouge-and-how-it-works-for-evaluation-of-summaries-e059fb8ac840/) which is used for automatic evaluation of summaries.

**2.2.** Data used for this application is ad hoc. We are looking into [BBC News Summary dataset](https://www.kaggle.com/pariza/bbc-news-summary) from Kaggle and various other widely used datasets by researchers mainly [Newsroom](https://github.com/lil-lab/newsroom) and [CNNDailyMails](https://github.com/abisee/cnn-dailymail).

**3. User Interface**

What will the user-interface that packages your model(s) look like and how will you make it user-friendly for someone to leverage your work?

There will be a website that will be 2 sections.

**Section 1:**

Provides user the option to upload a news article and select the option to categorize it or summarize it based on the option that the user chooses. (We can try to provide the user with both options as well, given the timeline, we cannot say if we can integrate both our services.)

**Section 2:**

User has the option to select a category and she/he can get summaries of news on that category from various news sources. For example, User can select COVID, and he/she can get summary of the news articles that are relevant to this category.

**Note:**

Static dataset will used for both the service. This service has an excellent future scope where we can get online dataset as well. But considering the timeline of our project, this service is mainly a prototype that can classifies or summarizes news articles. But user has the option to upload a news article and get summary on it using our service mentioned in Part 3, section 1.