**Project Milestone 1 – Overview**

**Abstract**

The topic for this project is to predict the news categorization. With the increase in the number of news it has become difficult for users to access news of their interest which makes it a necessity to categories news so that it could be easily accessed. Categorization refers to grouping that allows easier navigation among articles, internet new need to be categorized so that user would get recommendation of appropriate news based on their read news and categorization.

**How Categorization and or Classification help resolve this problem?** Categorization and classification allow humans to organize things, objects and ideas that exist around them and simplify their understanding of the word. Categorization is important in learning, prediction, inference and decision making.

The main idea of this project is to accurately identify the news category based on the headline and short summary so that the model will be used for recommending news to its user based on their past reading history.

By using Natural Language Processing (NLP) techniques to classify the text automatically by analyzing and then assigning the set of predefined tags or categories based on its context. NLP provides several techniques that would help text classification and categorization, for this project NLP techniques will be used to handle the problem statement (given below).

**Data Description** - This dataset contains around 200k news headlines from the year 2012 to 2018 obtained from HuffPost. The model trained on this dataset could be used to identify tags for untracked news articles or to identify the type of language used in different news articles.

Each news headline has a corresponding category. Categories and corresponding article have different counts. These categories are predefined at the time of collecting the data and assigned them accordingly.

The main idea behind this project is to find the solution of below problem statements and use the same model to recommend news articles to used based on their past reading history/habit.

**Problem Statement and Inspiration -**

1. Categorize news articles based on their headlines and short descriptions?
2. Do news articles from different categories have different writing styles?
3. A classifier trained on this dataset could be used on a free text to identify the type of language being used and use model for news recommendation.