ISyE 6740 – Summer 2021

Project Proposal

Team Member Names:

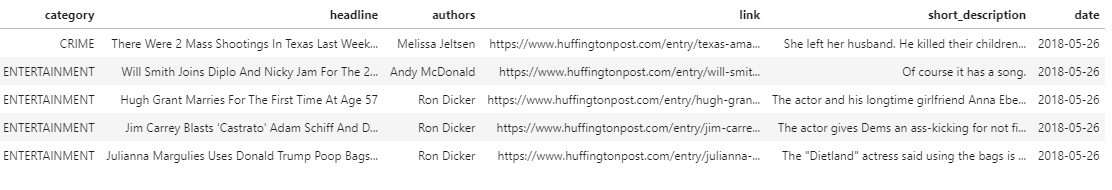
Project Title:

**Problem statement**

With the rapid development of digital platforms, the amount of online news published every day has grown drastically over the past few years. How must they be distributed to the appropriate audience in a timely and accurate fashion has been a trending research topic. News classification using machine learning methodologies has proven to be quite successful in providing information of interest in real-time.

Our group will utilize several algorithms to perform such news classification using data collected from Kaggle. We first perform vigorous exploration data analysis and choose several appropriate algorithms and finally We will compare the performance of these models.

**Data Source**

The input dataset consists of 200853 news from the year 2012 to 2018 obtained from HuffPost. The labels provided for this classification problem includes a total of 41 categories, such as Crimes, Politics, Wellness, Entertainment, etc. The text features are news title and descriptions. we will be applying NLP techniques to these columns and convert them into bag-of-words. Another feature that we might use is Author, which we assume most authors will only focus on a smaller number of topics. Suggested by Oliver et al.(2014), adding aggregation features to non-aggregated features could improve the accuracy of classification. Thus, we will create aggregated features such as number of Crime news by author, topics that has not been published by author, etc. However further data exploration is needed to be certain if they will be helpful to our model.

After the initial data exploration, we have a few concerns regarding the labels provided by the data owner. For example, Good News and WEIRD NEWS are extremely broad category that may lower our model accuracy. Also, WORLDPOST and THE WORLDPOST, CULTURE & ARTS and ARTS & CULTURE mean the same thing and should be group together. Therefore, further in-depth data exploration and testing with models should be conducted along with proper visualizations to uncover the true categories and help us decide whether if we should merge some of the provided labels.

**Methodology**

1. EDA
2. Data Preparation and cleaning: language-specific preprocessors that removed stop words and/or stemmed words. Split into train, validation, and test sets. Take account of unequal categories.
3. Text preprocessing: TF-IDF, (embedding?) (Bert?) etc.
4. Modeling: linear models or non-linear models based on EDA
5. Model tunning

**Evaluation and Final Results**

Our evaluation metrics are classification accuracy and F1 scores.

**Reference**

O. Schulte and K. Routley, "Aggregating predictions vs. aggregating features for relational classification," *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2014, pp. 121-128, doi: 10.1109/CIDM.2014.7008657.