Yu Lu Liu

# Final Project MAIS 202 Deliverable 1: Project Proposal

## Dataset

Sarcasm is a form of communication that is sometimes difficult to detect, especially in text form, where there is a lack of tonal cues. My project aims to detect sarcasm in news headlines. In other word, I would attempt to classify news headlines into two categories: sarcastic and not sarcastic. The dataset I would be using is the News Headlines dataset for Sarcasm Detection from R. Misra and A. Prahal (2019). I chose this dataset for its high usability. Unlike datasets collected from Reddit or Twitter, there's no spelling mistake nor slang words to worry about. Each of the 28619 headlines has also been clearly labelled.

## Methodology

I. Data Preprocessing

The dataset is organized into 4 columns: the URL source of the news, the headline, the sarcasm label, and the news company who published the news. I will only keep the headlines with their corresponding labels. Then, I will tokenize the headlines and remove common stop words such as "the" which offers no valuable information.

II. Machine Learning Model

I will attempt to implement a Recurrent Neural Network (RNN) that will output a category: sarcastic or not sarcastic. I chose RNN because it is shown to be more effective than Convolutional Neural Network to handle sequential data (Yin et al., 2017). This will allow the model to hopefully get a better grasp of the context of each word, which is crucial to detect sarcasm.

III. Final Conceptualization

I wish to implement a simple webapp. The input would be news headlines entered by the user, and the output would be a display of whether the headline inputted was sarcastic or not.

## References:

Rishabh Misra, Prahal Arora. 2019. Sarcasm Detection using Hybrid Neural Network. (Kaggle dataset: https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection/metadata)

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schutze. 2017. Comparative Study of CNN and RNN for Natural Language Processing. arXiv preprint arXiv:1702.01923.