# Data Selection Proposal

1. Dataset found at: https://www.kaggle.com/snapcrack/all-the-news
This dataset was chosen as it contains a large amount of data (over 100k articles) ranging from 15 different news sources which would allow me to compare them well. The dataset is textual and is a great resource to perform sentiment analysis on.

2. i) The data is already nicely separated into the text that I would like to analyze (the article's content) and metadata such as the article's author and time of publishing.

3. ii) I will be categorizing data into a set of emotions (e.g. happy, sad, trust, fear, worry, etc.) which corresponds to the main emotion of the article's text. I will be working with sentiment analysis algorithms, a subset of nature language processing methods. Another concurrent idea I have is to try to group articles based on similarity of subject. This also relies on NLP algorithms; more research will be done to determine exactly what algorithms and methods to utilize, such as a Recurrent Neural Network (RNN).

3. iii) I will be attempting to showcase the project's results as a webapp, ideally an interactive one where users will be able to explore the results by news source, by year, all data, or by emotion. Individual articles will be able to be shown based alongside their predicted emotion category and their predicted probabilities for each category.