W4995 Project Proposal

Members: Tejas Badgujar (tcb2145), Ritayan Patra (rp3247), Sion Chun (sc3791), Meyhar sharma (ms7070)

Background and context to the problem statement:

Sarcasm is an interesting notion to capture and quantify in a Natural Language Processing setting, and implementing the same can be fairly difficult. Sarcasm often involves specific language patterns and a deep understanding of the language itself, so detecting it robustly can be fairly difficult, even for humans.

Doing so in settings such as public communication (e.g., headlines from the news) is fairly important, as discerning between sarcastic and non-sarcastic headlines could greatly change the writer's overall intent and the likely content of the article itself. It impacts which and how many readers are interested in reading it based on the headline.

As such, the goal of this machine learning project is to develop a NLP-based machine learning model that can detect whether or not (binary classification) a given headline for a news article is sarcastic.

Note that even though a current-generation LLM will likely be able to do this easily, the goal is to develop something lightweight (i.e., not involving billions of parameters) but still capable of performing the task well.

Identification and description of the data set(s) you are planning on using along with their source:

The data collected comes from an overly sarcastic, satirical news source (The Onion) and a straightforward, likely non-sarcastic news source (HuffPost) and can be found here https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection.

Proposed ML techniques you are proposing on applying to solve the problem:

As mentioned above, since this is a binary classification problem, we will use binary classifiers for the project. We will explore the simpler ones (e.g., logistic and Tree-Based Classifiers), as well as possibly neural networks (e.g., CNNs, RNNs, and Attention-based models) should we deem them relevant and viable.

However, as with any NLP-based project, the key part of the project will be the feature extraction and creation. The crux of the problem is trying to extract information relevant to our machine-learning problem from the raw text headlines. This could be frequencies of specific sarcastic-indicating words, the general word embedding of the text using some open-source model, etc. As we have not deeply covered NLP-based methodologies in class yet, we have not delved deeply into this section yet; however, we intend to use the things we are taught in class as a baseline for the machine learning model we will develop.