**CS410 Project Proposal**

Team: Spiderman

Term: Fall 2022

Members: Sumanth Sureshbabu (sumanth4@illinois.edu)

Vasantha Kumar Dhanasekaran \* (vkd2@illinois.edu)

Rajesh Grandhi (rajeshg2@illinois.edu)

**What is your free topic?**

Sentiment Analysis to predict sarcasm

**Please give a detailed description. What is the task?**

Project is to build and train a classifier. Classifier will be built using neural network. We are planning to use ‘tensorflow’ library to build and train the model.

Tasks are:

1. Get the data
2. Clean and transform the data
3. Exploratory data analysis
4. Feature engineering and selection
5. Machine learning using the training dataset
6. Evaluate the trained model
7. Document results/ Deploy the model

**Why is it important or interesting?**

It’s interesting since our model will detect whether sarcasm is involved in a review/comment/news article. It’s also important to detect sarcasm as the article can be tagged right.

**What is your planned approach?**

Planned approach is to collect data from an open source and then split it as 70:30 to train and test the model. Once the model is built, we can evaluate it against any text and observe the results.

**What tools, systems or datasets are involved?**

**Tools**: Tensorflow library

**System**: Classifier (Neural network with sigmoid activation function)

**Datasets**: We are using the google sarcasm dataset. Past studies in Sarcasm Detection mostly make use of Twitter datasets collected using hashtag-based supervision but such datasets are noisy in terms of labels and language. Furthermore, many tweets are replies to other tweets and detecting sarcasm in these requires the availability of contextual tweets.

To overcome the limitations related to noise in Twitter datasets, this News Headlines dataset for Sarcasm Detection is collected from two news website. TheOnion aims at producing sarcastic versions of current events and we collected all the headlines from News in Brief and News in Photos categories (which are sarcastic). We collect real (and non-sarcastic) news headlines from HuffPost.

This dataset has following advantages over the existing Twitter datasets:

* Since news headlines are written by professionals in a formal manner, there are no spelling mistakes and informal usage. This reduces the sparsity and also increases the chance of finding pre-trained embeddings.
* Furthermore, since the sole purpose of TheOnion is to publish sarcastic news, we get high-quality labels with much less noise as compared to Twitter datasets.
* Unlike tweets which are replies to other tweets, the news headlines we obtained are self-contained. This would help us in teasing apart the real sarcastic elements.

**What is the expected outcome?**

Expected outcome is a model which can predict whether the author wrote a text sarcastically

**How are you going to evaluate your work?**

We are planning to use the basic measures:

* Classification accuracy
* Precision and Recall
* Area under curve

**Which programming language do you plan to use?** Python

**Please justify that the workload of your topic is at least 20\*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.**

|  |  |
| --- | --- |
| **Task** | **Duration** |
| Identify dataset Research – Data collection | 6 hours |
| Clean and transform the data – Data preparation | 5 hours |
| Explore the dataset – Exploratory analysis | 8 hours |
| Feature engineering and selection – Feature engineering | 8 hours |
| Building the classifier - Development | 10 hours |
| Evaluate the model – Testing and validation | 5 hours |
| Deploy in tensorflow - Implementation | 11 hours |
| Overall documentation & presentation | 10 hours |