410 Project Proposal (Fall 2022)

Team name: MATH

Team members:

Anqi Chen ac89 (captain)
Tianhao Luo tluo3
Hang Zhang hangz2
Md Majharul Islam Rajib mrajib2

1. **What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?**

Overview of task:

We would like to build a classification algorithm based on labeled text data, and the classification algorithm will be either sentiment analysis or sarcasm detection (we most likely will go for sarcasm detection, but also want to have a plan B if things don't go as planned).

More details can be found if you follow the links in the 'Data' subsection.

Tools & models:

It is interesting because it will use some models to analyze the feelings of human beings, purely from the text. Also, if we are able to detect sarcasm, it would be even more interesting since this would be more subtle than sentiment analysis. Even more, it can serve as (part of) a fake news detector for platforms such as Twitter/Facebook/Tiktok etc.

The packages we plan to use include numpy, pandas, matplotlib, seaborn, nltk, keras, pytorch, wordcloud (not an exhaustive list yet). The models we plan to use include simple machine learning algorithms such as logistic regression and decision trees, and also will include some state-of-the-art models from deep learning. If we have extra bandwidth, we will also make a web UI to make predictions on new data (tools may include javascript, html, bootstrap, exact tools TBD).

Data:

Data for sarcasm detection
https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

Data for sentiment analysis (back-up plan)
https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews

We expect these two tasks to be very similar.

Expected Outcome

Expected outcome of our group project is to categorize each news headline into the categories defined (positive vs negative, sarcastic or not) based on the route we take during the project phase.

Evaluation metrics

We will evaluate the outcome by metrics such as accuracy, precision, recall and F1 score and any other metrics that we find suitable during our project.

2. **Which programming language do you plan to use?**

   Python

3. **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.**

   - Preprocess data (stemming, lemmatization etc.) - 10 hrs
   - Explore the correct models/tools to use - 10 hrs
   - Build and train the model - 15 hrs
   - Potential class imbalance of data labels - 10 hrs
   - Test data with actual dataset & test with different parameters/models - 10 hrs
   - Create visualization demonstrating the results - 10 hrs
   - Create an API such that it can take in some unseen data and make a prediction (15 hrs, if time permits)
   - Documenting work and writing reports - (15 hrs)