**`Sentiment Analysis Domains:**

Social Media: Twitter, Facebook, Insta, etc.

Twitter: <https://www.kaggle.com/paoloripamonti/twitter-sentiment-analysis/notebook>

News:

Financial News Sentiment Analysis: <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news>

Commercials/Movies:

Songs:

49 song lyric txt files:

https://www.kaggle.com/paultimothymooney/poetry

Medical Dataset (ongoing Kaggle competition):

<https://www.kaggle.com/c/nbme-score-clinical-patient-notes/overview>

Mental Health in Tech Dataset (Kaggle)

<https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey>

<https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016>

Eating disorder tweets

<https://www.kaggle.com/datasets/jabenitez88/eating-disorders-tweets>

Natural disasters dataset:

<https://www.kaggle.com/c/nlp-getting-started/data?select=sample_submission.csv>

—----

Your proposal should be roughly **300 words** in total and contain the following:

1. **All group member’s names**

Seung Hyun (Amy) Ryu

Joshua Toth

Joseph Poirier

* 1. **Breakdown of what each member will be responsible for:**

EDA: Seung Hyun (Amy) Ryu

Timeseries EDA: Joshua Toth

Predictive modeling: Joseph Poirier

1. **Data source - if you are proposing your own dataset/ topic.** Mental Health in Tech <https://www.kaggle.com/datasets/osmi/mental-health-in-tech-2016> (also available for 2014 - 2018 timeseries). Wwith currently over 1400 responses for just the year 2016, the ongoing survey aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers.
2. **Project Plan**
   1. **Explain what you intend to study with your project.**

What best predicts an employee’s likelihood of discussing mental health issues/ effectively addressing mental health issues in the IT workplace? What are the resources that employwers can provide to improve their employee’s well being. Has the trend on mental health resource availability changed over time?

* 1. **What is the ultimate objective?** Predictive analytics on fields such as - “Would you have been willing to discuss a mental health issue with your direct supervisor(s)?”/ “Do you feel that being identified as a person with a mental health issue would hurt your career?” Ultimately identify the features with the biggest impact on mental health safety within a company. This could include features such as “Does your employer provide mental health benefits as part of healthcare coverage”, “Does your employer offer resources to learn more about mental health concerns and options for seeking help?”
  2. **What types of models are you considering?** Logistic regression or decision trees

1. **Why is this project interesting?**

This project could have significant social impact in a corporate setting for those who are struggling with mental health. Empoyers can use this data to drive work in raising awareness and improving conditions for those with mental health disorders in the IT workplace. Especially in the COVID remote work enviornment, mental health is increasingly an important aspect of worker well being and productivity. Companies have implemented various mental health education as well as insurance coverage measures but it’s still unclear what has proven the most effective. We hope that this project can use machine learning to quantitatively highlight the most effective policies.

1. **What challenges and obstacles might you anticipate with this project?**
2. Feature selection (There are a total of 63 questions in the survey)
3. Preprocessing (combination of binary, ordinal, datetime-formatted etc.)
4. Determining the best model to use
5. **TA name if you would like to request a TA** NA

Seung Hyun (Amy) Ryu, Joshua Toth, and Joseph Poirier

Project Proposal: Analyzing Mental Health in the Tech Workplace

For our project we plan to analyze the Mental Health in Tech Survey dataset (<https://www.kaggle.com/datasets/ekwiecinska96/mental-health-in-techology-survey-2014-and-2016?select=survey_2014.csv>). With currently over 1400 responses per year, this ongoing survey aims to measure attitudes towards mental health in the tech workplace, and examine the frequency of mental health disorders among tech workers. We plan to address questions such as what best predicts an tech employee’s willingness to discuss mental health issues, what resources can employers provide to improve their employees’ well being, and are trends in mental health occurence and resource availability improving or worsening over time? We hope to provide predictive analytics on fields within this dataset, such as responses to the questions, “Would you have been willing to discuss a mental health issue with your direct supervisor(s)?” and “Do you feel that being identified as a person with a mental health issue would hurt your career?” utilizing modeling methods such as logistic regression and decision trees.

This project could have significant impact in a corporate setting for those who are struggling with mental health and their employers, the latter of which can use results of our analyses for insight towards raising awareness and improving conditions for those with mental health disorders in the IT workplace. Especially for companies implementing remote work in light of COVID, mental health is increasingly an important aspect of worker well being and productivity. Companies have implemented various mental health education as well as insurance coverage measures but it’s still unclear what has proven the most effective. We hope that this project can use machine learning to quantitatively highlight the most effective policies.

At this time, we foresee three challenging aspects of this project. First, feature selection for our models will need to be carefully examined, as there are over 100 questions on these surveys. Second, data preprocessing and integration could be complex given the datasets contain binary, ordinal and time-formatted type data. We anticipate that free-response questions will require NLP techniques to format and extract relevant information. Lastly, since there are many possibilities regarding model development with this dataset, determining the best types of models to develop within the time frame of the project will involve some consideration.

We plan to divide the work for this project roughly as follows, however, we hope that all of us will ultimately become familiar with all aspects:

General preprocessing and EDA : Amy Ryu

Timeseries preprocessing and EDA/predictive modeling: Joshua Toth

Predictive modeling: Joseph Poirier

Requested TA: We are happy to work with any TA!

—-

**Topic Modeling for Research Articles: 24 Classes (Article Tags)**

<https://www.kaggle.com/datasets/abisheksudarshan/topic-modeling-for-research-articles/code?select=Train.csv>

Large dataset >>10K, part of a guided hackathon (there is some reference for us to go off of). Not commonly used so far (which is a good thing).

Can we accurately apply domain tags to science articles based on analysis of their abstracts?

**Sarcasm Prediction on Reddit: 32,332 comments, 50% sarcastic.**

<https://www.kaggle.com/datasets/danofer/sarcasm>

Large dataset >> 10K, but no other fields than the comment itself (although I don’t think that’s a problem?). Also four years old, meaning that some good solutions already exist. Seeing existing projects with >80% classification accuracy.

Can we use BERT/convolutions/something else to accurately interpret the sentiment of a comment as sarcastic or serious?

**NFL Dataset: Data from Every Play Since 2013**

<http://nflsavant.com/about.php>

Extremely large dataset, already preprocessed well. Lots of potential prediction areas. 39 Fields. Still some room for preprocessing if we want to separate out data from play descriptions.

What kind of play is most effective in a given situation?

Can we predict future performance of a specific type of player given past performance (lots of time-series potential)

Which players are most overvalued/undervalued based on their production?

**Positive/Neutral/Negative Sentiment Analysis on 943,672 Stock-Market Tweets**

<https://www.kaggle.com/datasets/sohelranaccselab/stock-market-tweets-data-sentiment-analysis>

Massive dataset, could lead to better/more interesting results than the other two. New, published January 2021. Looks like it hasn’t been used yet. If we posted a solution to this, we would be the first. Could adjust pretrained BERT model for this task.

Can we accurately predict sentiment of stock-market-related tweets?

**Microsoft News Recommendation Dataset**

<https://www.kaggle.com/datasets/arashnic/mind-news-dataset?datasetId=1049650>

Sample of 1M users who clicked on at least 6 Microsoft News articles in October – November of 2019. Biggest con: will be a complicated/involved project. Also will require cloud computing. Interesting questions though.

Given activity data for 1M users and two-thirds of the user’s clicks, can we build a recommender that predicts with high probability that a user will be interested in their remaining clicks?

Your proposal should be roughly **300 words** in total and contain the following: