**Summary FGH/GWU Datathon January 14th-17th, 2022**

By Jake Lieberfarb

In collaboration with the [GWU Data Science Association](https://datasci.columbian.gwu.edu/data-science-association) and the [American Statistical Association](https://www.amstat.org/), a datathon was created utilizing Twitter’s [rest API](https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet) **to** understand and predict engagement. The data sample contained 90 columns with each row of data describing a single tweet such as [**user data**](https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user) (name, screen name, follower count, verified, etc.), **tweet information** (tweet text, created at, etc.), and **engagement data** (retweet count, favorite count, reply count). The contestant’s code, video presentation, and description of their project are linked below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Submission Summary** | | | |
| **1** | **2** | **3** | **4** |
| [Github](https://github.com/hum1995/GWU-Datathon) | [Github](https://github.com/Mooozer/GWUDatathon/blob/main/2022GWUDatathon_MuzheGuo.ipynb) | [Github](https://github.com/xiangli2pro/GWU_Datathon_2022) | [Github](https://github.com/RmmLeo/2022-GWU-Datathon.git) |
| [Video Presentation](https://youtu.be/jlaKCt3Bc0c) | [Video Presentation](https://drive.google.com/file/d/1JHGARSgxnm2h3xS7W3Js-I4grFR0FzST/view?usp=sharing) | [Video Presentation](https://drive.google.com/file/d/1_gck0DtHVBHyAjeM3WF0HU6K7DlB1_7Z/view?usp=sharing) | [Video Presentation](https://drive.google.com/file/d/17jy6E2Ep9grVCLK2aoITfDGFqu5fWuVq/view?usp=sharing) |
| * Built multinominal model to predicted retweet counts based on the content of the tweet('text'), the identity of the account('decription') and other characteristics of the account.   + Retweets were broken down into three categories:   **Low(<=5), Medium(6-100),**  **High(>=100)**   * Model accuracy prediction results:   Low- 80.03%  Medium- 67.74%  High- 84.71% | * Classified retweets into five groups: very low, low, moderate, high, and very high * Constructed five different models to evaluate engagement * Accuracy prediction of models   **Naïve** **Bayes**- 40%  **Logit**- 43%  **BERT**- 45%  **DNN -** 45%  **BERT + DNN**- 51% | * Categorized tweets into seven categories (anger, joy, sadness, anticipation, trust, surprise, and disgust) to understand engagement from political tweets through [NRC word-emotion association lexicon](https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm) * Politicians had greatest engagement with tweets categorized as ‘anger’ | * Utilized s[entiment analysis](https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6) to categorize tweets   + **+ 1 for positive**   + **0 for neutral**   + **-1 negative** * Fed tweets into neural network to find probability of tweet being categorized into three groups |

**Take away:**

How can we implement these models in our current analysis of tweets