INFM 603
Angela Tseng
Yogesh Boricha
Abdul Shaik
Sadaf Davre

Project Prototype

The prototype can be accessed from this folder under the Project Prototype Colab Notebook. All other necessary files to run the program are contained in the folder. Since the detailed description, the project has been scaled down to focus on the programming and machine learning components instead of database or website design. So, as per feedback, we have eliminated MySQL database and website from our project. Now, the project centers largely around executing and displaying results of the machine learning code in a Colab notebook. The notebook contains code that requests top hashtags from Twitter, parses for news regarding the hashtags, and assigns a 0 or 1 label to a news source depending on its reliability according to trained data. A label of 0 implies that the source may not be reliable, while a label of 1 suggests that the source is worth looking at. We are currently working on improving the prototype by assigning a percentage that indicates the level of authenticity rather than a simple 0 and 1 label, along with presenting it in a more user-friendly manner.

The first cell uses Twitter's elevated access status and various keys to fetch the top 50 Twitter trending topics in an area. The area is determined by the WOEID, which can easily be looked up to match the location of interest. These trends are stored into a dictionary named "top_trends." The user can browse the trends to find something they are interested in, or simply proceed to the next few blocks of installing packages. Using the pygooglenews, beautifulsoup4, and newspaper3k packages, the program reads in search results and stores the links into a list named "links". With the list created, it iterates through the entries and stores the title, link, content, and date of the article fetched into a csv file. If any of these are absent or unreadable by the program, "unusable article found, aborting parse attempt and continuing to next article" is printed to the console. This information is stored in a csv file named "data", that will be used later on. After transforming and cleaning the data for the machine learning code, the pandas library is used to read in the trainer files "True.csv" and "False.csv" to train the algorithm. Each csv file is a dataset with 10,000 new articles organized into columns "title", "text", "subject", and "date". A new column named "label" is created to store whether the source is reliable or not. The code then cleans up the training data by dropping missing data, parsing for stopwords, and splitting the words accordingly. After the processing, the sklearn package is used to train and classify the data before the accuracy score of the trained model is printed. The pickle package is then used to load a model and vectorizer, tokenize the data, and predict whether the article is reliable or not. The resulting output is a list of links with accompanying 0's and 1's on the left.

Running the prototype takes some time since certain packages must be installed and data is trained to assign a label to the fetched news sources. However, due to the iterative nature of Colab notebooks, simply running the cells one by one and waiting for cells to complete

running is enough to receive results. To change the location of trends fetched, simply change the WOEID in the first code block. This parses for links for that particular topic. Beyond that, each block after should be run chronologically. The last block of code assigns a 0 or 1 to the left of a linked article, allowing the user to scroll through the list in the console to select an article of their choosing. More detailed instructions and explanation of the implementation can be found in the Colab notebook itself.