Report

July 28, 2021

1 COVID-19 Retweet Prediction

1.1 Team Members

1. Aaron Khoo
2. Calvin Yusnoveri
3. Amarjyot Kaur Narula
4. Joseph Chng

1.2 Task Description

1.3 Dataset Description

Data downloaded from here: https://data.gesis.org/tweetscov19/#dataset

Data should be stored in `./data`

Coulmn Names: 1. Tweet Id: Long. - NOT USED 2. Username: String. Encrypted for privacy issues. - NOT USED

3. Timestamp: Format ("EEE MMM dd HH:mm:ss Z yyyy"). ISOString => integer (e.g. 23517957).
4. #Followers: Integer.
5. #Friends: Integer.
6. #Retweets: Integer.
7. #Favorites: Integer.
8. Entities: String. For each entity, we aggregated the original text, the annotated entity and the produced score from FEL library. Each entity is separated from another entity by char ";". Also, each entity is separated by char ":" in order to store "original_text:annotated_entity:score;". If FEL did not find any entities, we have stored "null;".
9. Sentiment: String. SentiStrength produces a score for positive (1 to 5) and negative (-1 to -5) sentiment. We splitted these two numbers by whitespace char " ". Positive sentiment was stored first and then negative sentiment (i.e. "2 -1").
10. Mentions: String. If the tweet contains mentions, we remove the char "@" and concatenate the mentions with whitespace char " ". If no mentions appear, we have stored"null;".
11. Hashtags: String. If the tweet contains hashtags, we remove the char "#" and concatenate the hashtags with whitespace char " ". If no hashtags appear, we have stored"null;".

12. URLs: String: If the tweet contains URLs, we concatenate the URLs using ":-:". If no URLs appear, we have stored "null;"

1.4 Preprocessing

Clean Data (Final structure/form of data before it is fed into the model): 1. Tweet Id: Long. - NOT USED 2. Username: String. Encrypted for privacy issues. - NOT USED

3. Timestamp: Format ("EEE MMM dd HH:mm:ss Z yyyy"). ISOString => integer (e.g. 23517957).
4. #Followers: Integer.
5. #Friends: Integer.
6. #Retweets: Integer.
7. #Favorites: Integer.
8. Entities: String. For each entity, we aggregated the original text, the annotated entity and the produced score from FEL library. Each entity is separated from another entity by char ";". Also, each entity is separated by char ":" in order to store "original_text:annotated_entity:score;". If FEL did not find any entities, we have stored "null;".
9. Sentiment: String. SentiStrength produces a score for positive (1 to 5) and negative (-1 to -5) sentiment. We splitted these two numbers by whitespace char " ". Positive sentiment was stored first and then negative sentiment (i.e. "2 -1").
10. Mentions: String. If the tweet contains mentions, we remove the char "@" and concatenate the mentions with whitespace char " ". If no mentions appear, we have stored"null;".
11. Hashtags: String. If the tweet contains hashtags, we remove the char "#" and concatenate the hashtags with whitespace char " ". If no hashtags appear, we have stored"null;".
12. URLs: String: If the tweet contains URLs, we concatenate the URLs using ":-:". If no URLs appear, we have stored "null;"

1.5 Model Architecture

We use ensemble methods (insert image): 1. 0-classifier (print out layer) 2. regression model (print out layer)

1.6 Results

Loss curve image.

Accuracy on train and test set.

Validation images.

1.7 Discussion

comparing with state of the art.

possible issues

1.8 GUI

Step-by-step Usage: 1. run xxx 2. click button 3. done

1.9 Sources

1. Source code: https://github.com/arglux/50021-ai-project
2. Report:
3. Reference papers:

[ ]: