CS 6220 Final Project Proposal

EDA and Sentiment Analysis on COVID-19 Tweets Data

Zhidong Qu, Yuxuan Xie, Suyue Jiang, Ruihan Wang

Introduction

During this unprecedented pandemic from Dec 2019 to now, we have been studying and working from home for more than 6 months. Social media started to become the only way that a lot of people actually interact with the outside world. Understanding how people feel, their sentiment shifts from the start of the pandemic to now, and the things that people actively discuss during this time could give us intuitions on how we might approach some of the problems when we faced the next pandemic. It is also a great way for us to study and understand people's sentiment and reactions when they are isolated from the real world with only the internet and social media available for them. Therefore, our team has decided to take a deeper look into the COVID-19 Tweets data and try to gather some intuitions and answer some of the problems discussed above.

Data Source

https://www.kaggle.com/gpreda/covid19-tweets

Proposed Analyses

We want to do two major analyses on the tweets dataset. Before diving into the detailed analyses, we would like to start with some simple data cleaning and visualization with the dataset. Then, we would start off gathering some general statistics and performing some EDAs with appropriate visualizations using pandas and Pyplot. Finally, we want to focus on analyzing the sentiment behind those tweets by training a simple sentiment prediction model with scikit-learn on the dataset which enables us to get the sentiment behind each tweet and perform some more interesting analyses. The detailed analyses that we would like to discover are listed below.

Exploratory Data Analysis

Data Cleaning:

- Visualize the columns with missing values
- Use different imputation methods to handle the missing values by column

Analyses on `hash_tags` column:

- Most trending (frequently tagged) `hash_tags` and how the trend changes over time

- Most frequently used `hash_tags` by celebrities/influencers (people with high `user_followers`) and how the trend changes over time
- Most used words of the most frequent hash_tags

Analyses on `text` column:

- Most frequently used words by all users and how the trend changes over time
- Most frequently used words by celebrities/influencers (people with high `user_followers`) and how the trend changes over time

Analyses on `is_retweet` column:

- Most retweeted tweets during the pandemic

Analyses on `user_location` column:

- Most active users categorized by geographic location during the pandemic after removing biases (potentially dividing the number active (users with a certain number of tweets per time period) of users during the pandemic by the number of total users given the geographic location)

Sentiment Analysis

Training:

We would most likely use the state-of-art RNN model to train our sentiment classifier. Alternative options like the multivariate naive bayes model might also be considered. The training will be done on the training set, prediction and analyses will be done on the test set. We may augment our dataset with some other twitter dataset during the COVID-19 with appropriate columns (e.g. *https://www.kaggle.com/surajkum1198/twitterdata*).

Analyses on predicted `sentiment`:

- People's sentiment shifts from COVID-19 start to present by month
- Most frequent words people use to express a sentiment level in a word cloud
- Most frequent hash_tags people use to express a sentiment level in a word cloud
- Sentiments of celebrities/influencers (people with high `user_followers`)
- Correlation between Hash-Tags and Sentiment

Emoji Sentiment Classifier:

We would also want to train a most recently proposed multi-class sentiment classifier that categorizes tweets to one single emoji and see how that plays out. What are some of the top frequent emojis that could express people's sentiment during this pandemic?
Pretrained Model: https://github.com/TetsumichiUmada/text2emoji

Final Deliverable

We will submit a complete jupyter notebook file that contains all EDAs listed above with appropriate plot/diagrams and text explanation for the discoveries we find. It would also include how we trained the sentiment classifier with RNN, hyperparameters of the model, train/test data split, training process, and the final prediction/analyses on the test set with text explanations. Finally, we will use the pre-trained text-to-emoji sentiment classifier obtained from the GitHub repository to categorize all tweets with emojis and include the relevant data and analyses in the notebook.