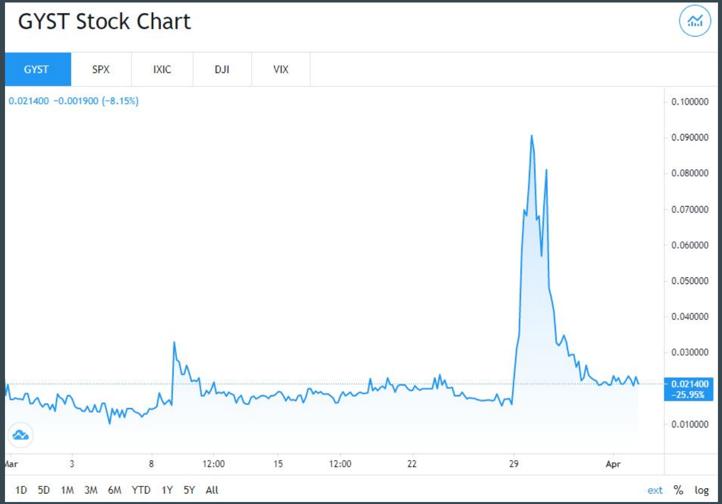
Stock Market and Social Media

Matt Gilgo, George Padavick

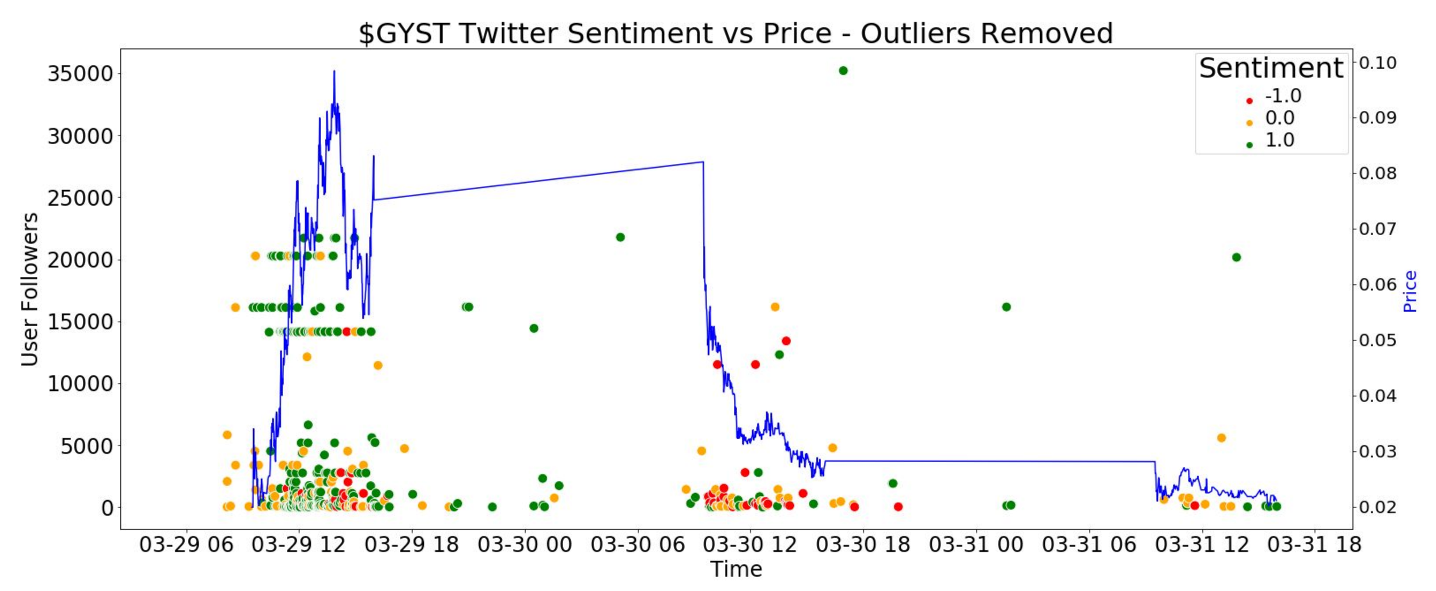
Motivation/Goals



- COVID driven market downturn and lack of activities to spend money on turned many new investors to the market (2020: 500%+ transactional value in Robinhood YoY)
- Major increase in new/unskilled investors has led to a surge in predatory market manipulation through means of social media
- Goals were to map potential pump and dumps amongst social media influencers and quantify impact to those falling into the scheme

Results/Analysis

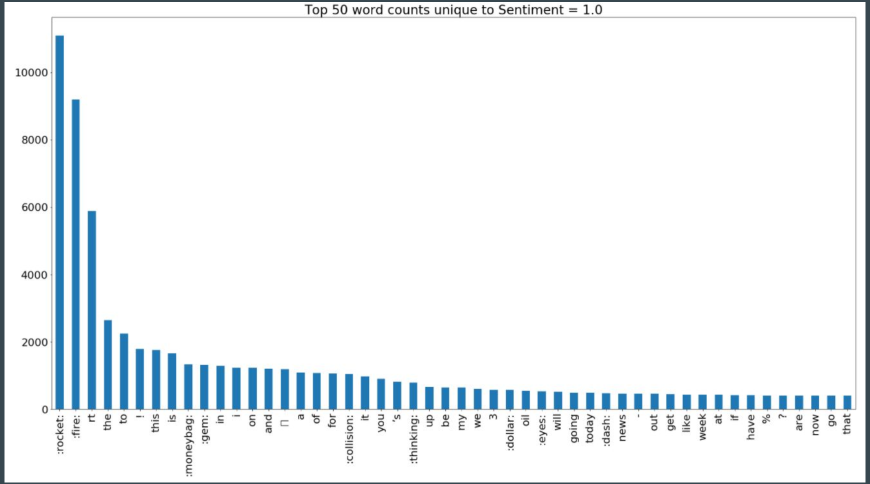
- 10000+ Tweets scored for Sentiment, Known Pumper Profiles, Buy/Sell Region, and Price Inflection Point across 10+ examples of known pumps



Correlation between price and sentiment found in labeled data

Results/Analysis

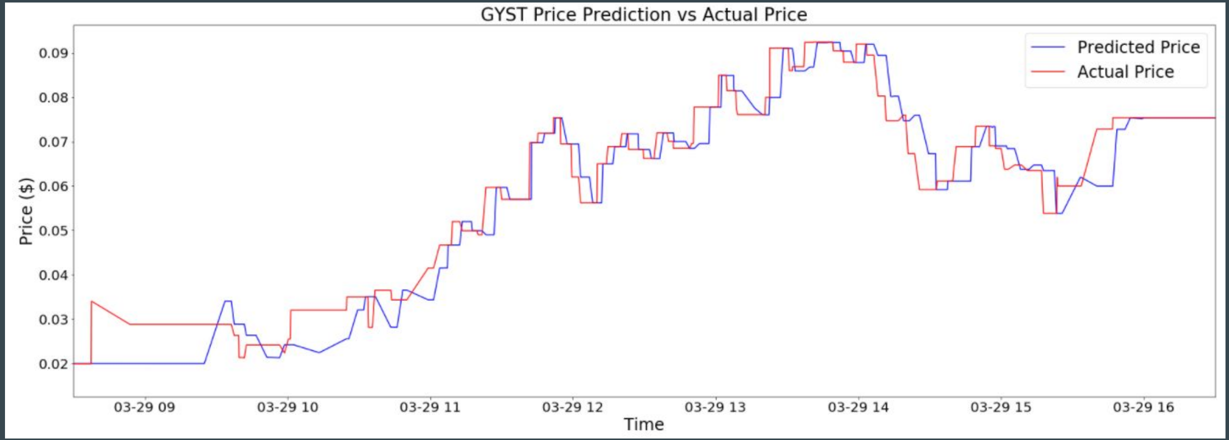
- Determined high use words for each sentiment classification
  - Positive sentiment: 🚀 , 🔥 , 💰
  - Neutral sentiment: “news”, “today”, “highest”
  - Negative sentiment: “dump”, “offering”, “sell”
- Incorporated top words into vocabulary for Countvectorizer and TF-IDF



Results/Analysis

- Generated sentiment classification models based on TF-IDF features - test results provided below
- Applied classification model to remaining unscored tweets
- Initial price model does not show predictive capability but demonstrates importance of sentiment in price prediction

	Test Accuracy	Cross Validation Accuracy
Logistic Regression	87%	88%
Random Forest	87%	88%
KNN	75%	87%
SVM	87%	87%

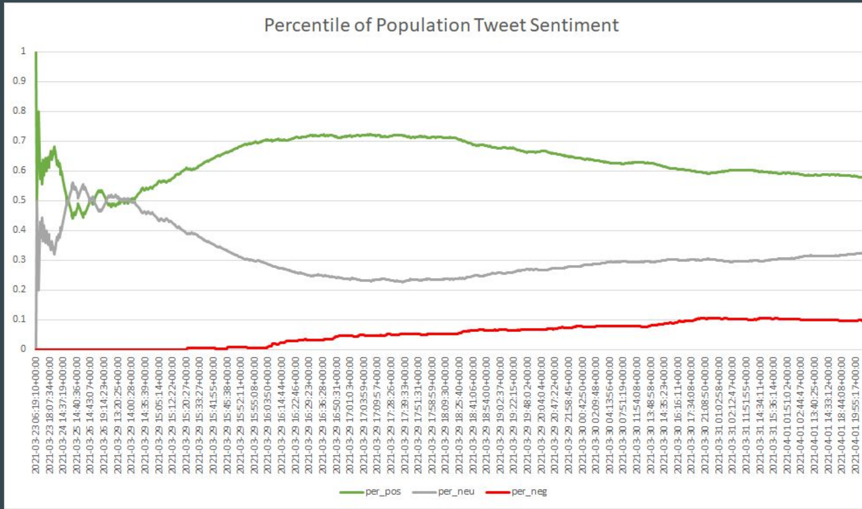


Challenges

- Tweepy API rates and tweet limit hindered collection of tweets
  - Determined pulling trailing 7 days on a weekly basis would be most effective strategy
- Stocks like GME, AMC, and KODK off the table due to not being recent enough
  - Found recent, short term pump and dumps like GYST, EEENF, etc. to pull tweets from
- "Fintwit" lingo is different then the lingo used in traditional sentiment analysis
  - Supervised Learning through manually scored tweets to ensure sentiment scored correctly
- Majority of tweets about stocks are positive, limiting negative sentiment examples
  - Even split between positive, neutral, and negative tweets when developing models

Limitations/Nexts Steps

- Limited datasets and time prevented long term studies on sentiment as well as price models
- Next steps would be further data collection, feature extraction (example to right), and development of price based models
- For further research, the dataset used in our sentiment analysis as been uploaded to Kaggle for others in the community to do studies with
    - https://www.kaggle.com/mattgilgo/stock-related-tweet-sentiment

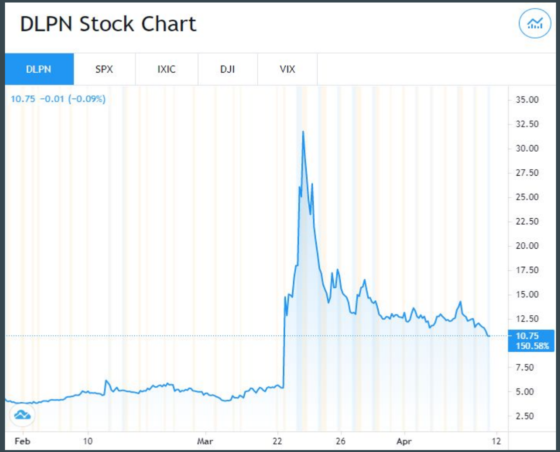


Thank you!

Appendix/Backup Slides

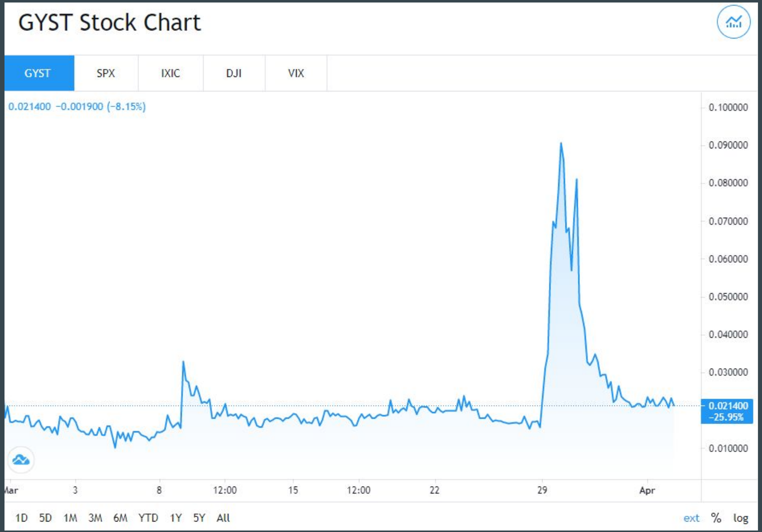
Data Collection

- Stocks used for investigation:
  - DLPN - (Digital Media & NFTs)
  - EEENF - (Australian Oil Driller)
  - GYST - (Precious Metal Mining Developer)
  - 10 others
- Tweepy (Twitter API) used for pulling tweets tagged with specified stock tickers
- Yahoo Finance API used for pulling interval stock price/volume data



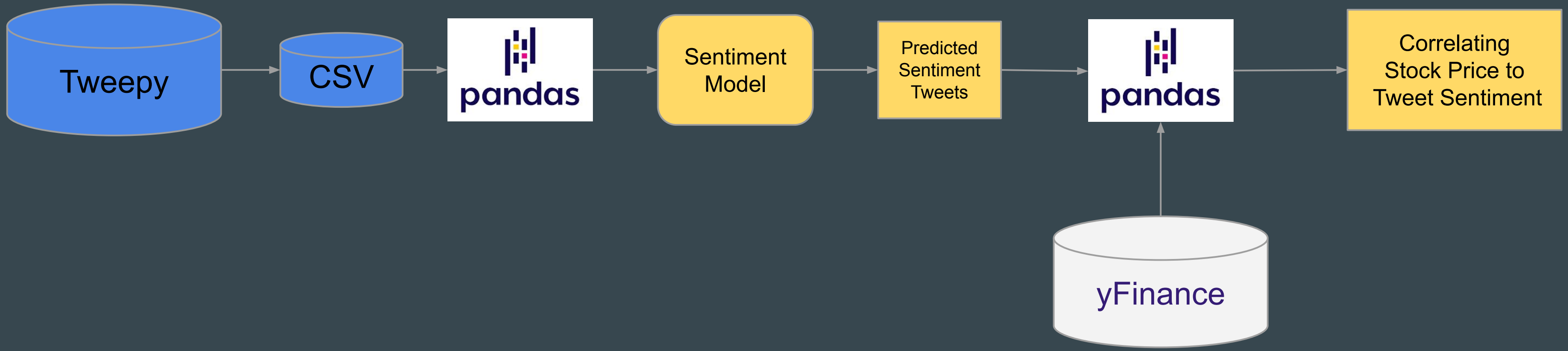
Or Data Collection

- Stocks used for investigation:
    - DLPN - (Digital Media & NFTs)
    - EEENF - (Australian Oil Driller)
    - GYST - (Precious Metal Mining Developer)
    - 10 others
- Tweepy (Twitter API) used for pulling tweets tagged with specified stock tickers
- Yahoo Finance API used for pulling interval stock price/volume data



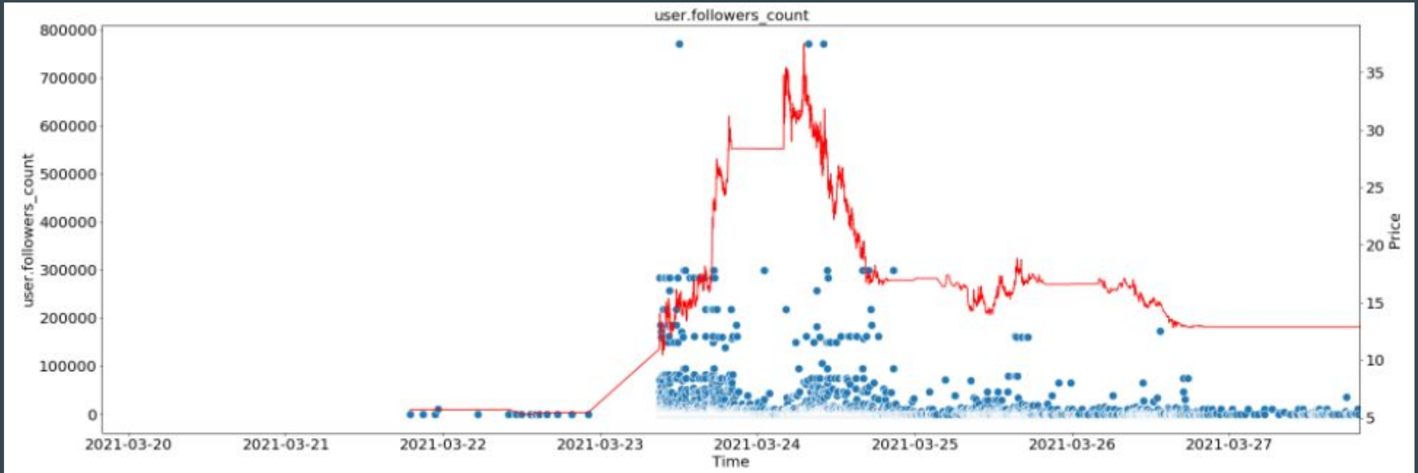
Data Progression

- Data issues
- Performed investigation into relationship between available data using $DLPN as an example case



Data Progression/Cleaning

- Developed scripts for pulling data using Tweepy
- Performed investigation into relationship between available data using $DLPN as an example case



Feature Extraction

Example below: Seeking various methods of capturing change in market behavior

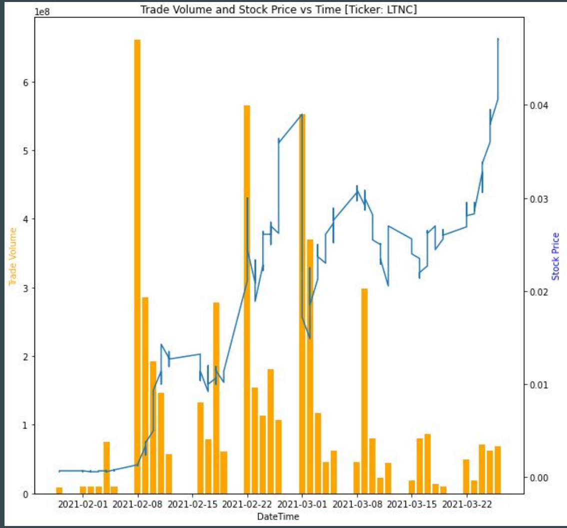


Fig 1: Total Volume vs Time

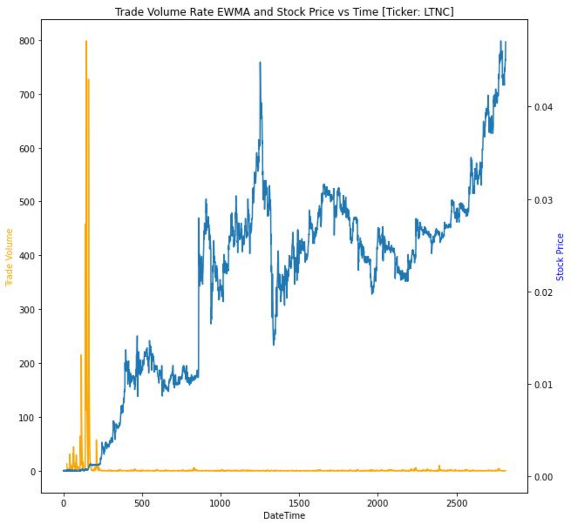


Fig 2: Total Volume Rate EWMA vs Time

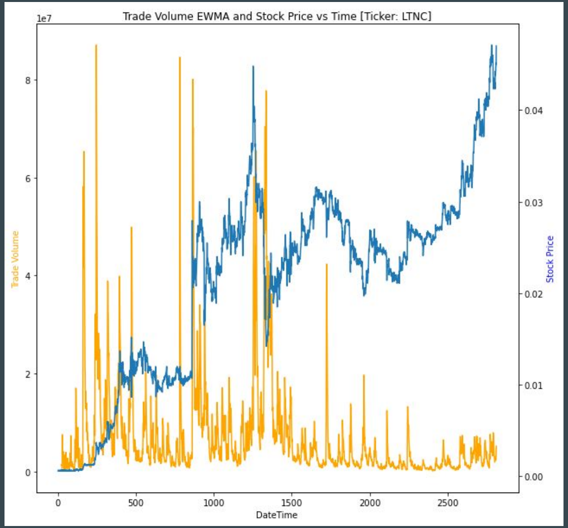


Fig 3: Total Volume EWMA vs Time