Problem Set #2
Twitter Sentiment Analysis of Publicly Traded Equity Securities

Dehao Wang, Jiaxin Zhang and Xiaoya Wang

- Data preparation
    - The 12 Equities we worked on are listed below.

ALGN	Align Technology, Inc.
AMP	Ameriprise Financial, Inc.
AON	Aon PLC
BRK.B	Berkshire Hathaway Inc. Class B
CMCSA	Comcast Corporation
CMI	Cummins Inc.
CNP	CenterPoint Energy Inc
DG	Dollar General Corp.
EXR	Extra Space Storage, Inc.
JKHY	Jack Henry & Associates, Inc.
JPM	JPMorgan Chase & Co.
LHX	L3Harris Technologies Inc

    - Our research's timeline is from 11/30/2020 9:30:00 AM to 12/8/2020 3:30:00 PM. We obtain the hourly stock data from Bloomberg, including the Last Price and the Volume. Then we calculate the log return, excess log return (benchmark - SPY), and volatility (rolling standard deviation) based on the original stock data. Furthermore, we retrieve the tweets using the Tweepy API by restricting the searching keywords to the company names or nicknames.

- Data preprocessing

  Preprocessing of tweets is performed by taking multiple steps.
    - Drop the duplicate tweets.
    - Change all letters to be lowercase ones.
    - Remove Twitter user name after '@'
    - Remove special characters
    - Remove HTML tags

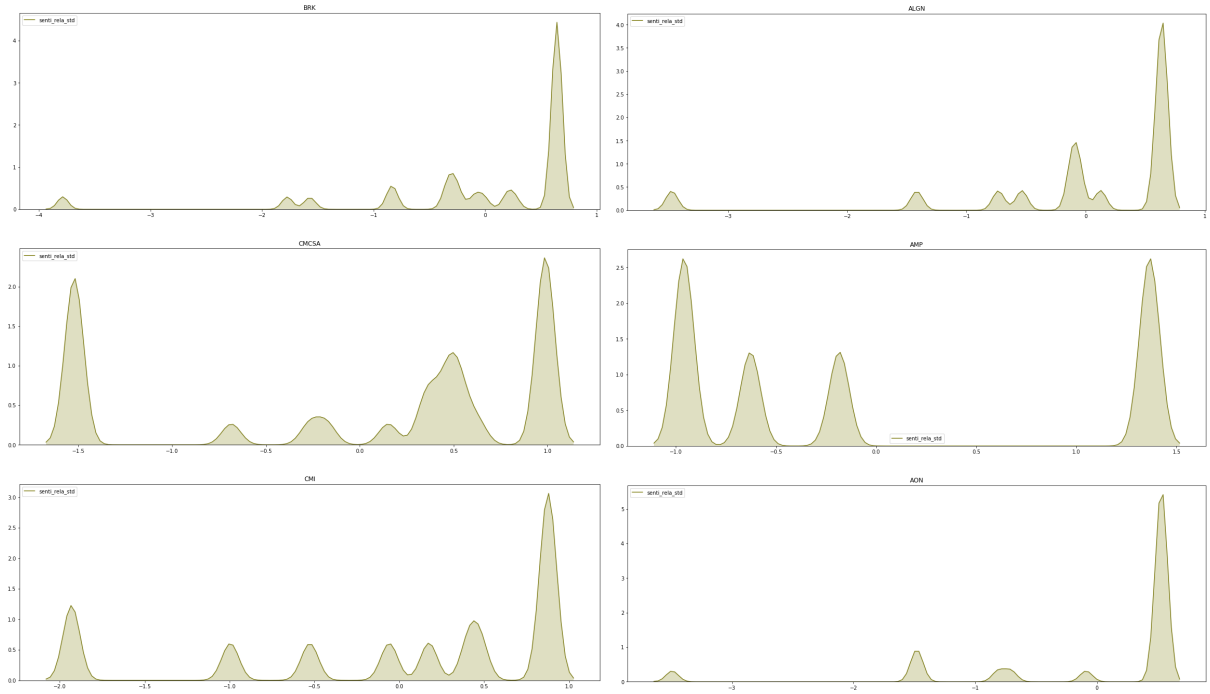
- Remove Emojis
- Remove '<.*?>' pattern texts
- Remove URL tags
- Remove numbers
- Remove stopwords
- Lemmatization

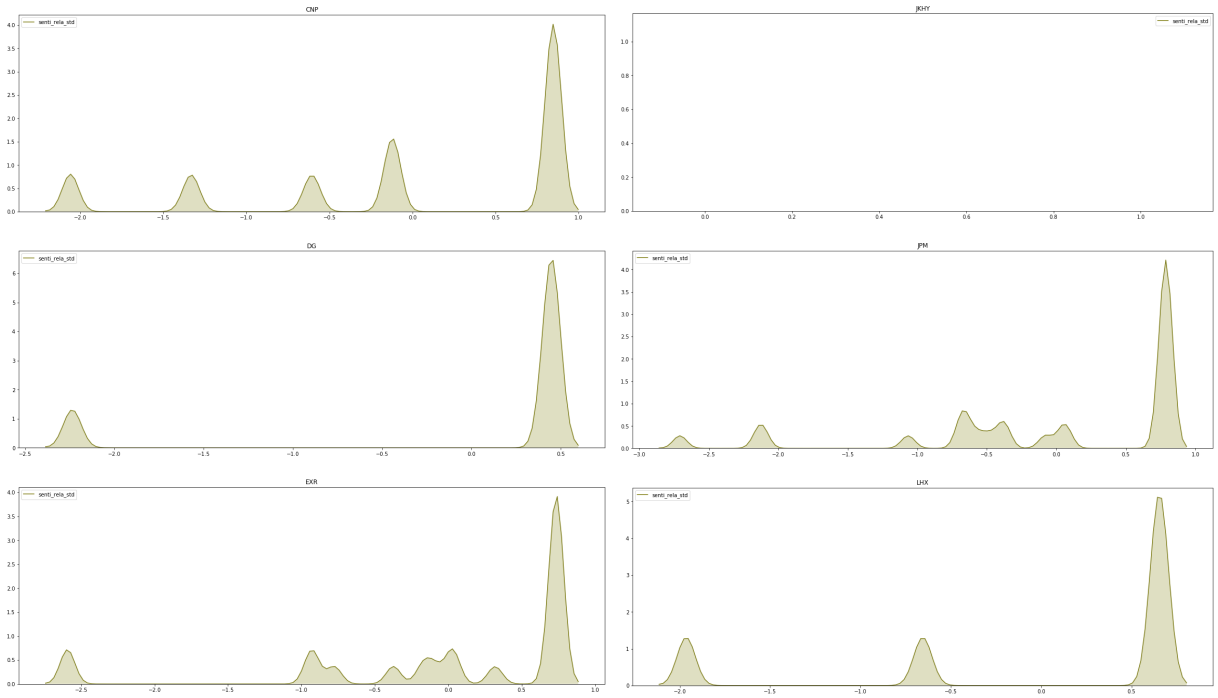
- Data analysis and result
  - First, we use Textblob, a python library for processing textual data, to classify the tweets. The Textblob could give us a sentiment polarity score after analyzing the tweets, and we classify the tweets based on the scores: if score > 0, positive; if score < 0, negative; if score = 0, neutral.

  - Next, we compute two variables— the absolute ($S_A(t)$) and the relative ($S_R(t)$) sentiment of each company in a given hour using the number of negative tweets ($B(t)$) and that of positive tweets ($G(t)$). The equations are shown below:

$$S_A(t) = G(t) - B(t) \quad S_R(t) = \frac{G(t)-B(t)}{G(t)+B(t)}$$

  - Then, we plot the distribution of relative sentiment from Twitter for the companies, and the Gaussian Kernel Density plots are shown below.





Due to the lack of Twitter data, some companies analyzed cannot depict significantly different distributions of sentiment. Also, JACK HENRY & ASSOCIATES, INC. does not have any available data.

- We use the Granger Causality test to determine if sentiment signal adds prediction power to various financial variables, i.e. volume, volatility, and return of each company. Although it is not a true Causality, we took 1 as max lag, which means we are testing if twitters in the last two hours have a significant effect on the financial variables considered. Test results for all companies in sample 18 are shown below. The pictures show the test results for them, and you can find the details from the ‘statistical analysis.ipynb’ file. Note that green icons in the tables indicate that the p-values are smaller than 0.1, and we have sufficient evidence to conclude that Twitter sentiment analytics is indeed relevant in the prediction of the financial variable. And the red icons mean that we cannot find a significant relationship between the Twitter sentiment analytics and the variable.

    - Excess log-return v.s. sentiment score

Equity	Level of statistical significance (p-value)			
	ssr F test	ssr chi2 test	likelihood ratio test	parameter F test
ALGN	0.6800	0.6449	0.6459	0.6800
AMP	0.5231	0.2250	0.2560	0.5231
AON	0.5520	0.5181	0.5199	0.5520
BRK.B	0.9553	0.9521	0.9521	0.9553
CMCSA	0.4669	0.4355	0.4379	0.4669
CMI	0.4567	0.3691	0.3770	0.4567
CNP	0.1555	0.0467	0.0701	0.1555
DG	0.9982	0.9976	0.9976	0.9982
EXR	0.1864	0.1379	0.1480	0.1864
JKHY	0.3874	0.2312	0.2509	0.3874
JPM	0.0878	0.0591	0.0674	0.0878
LHX	0.3863	0.1987	0.2233	0.3863

We can find that only the excess log-returns of JPM and CNP have significant relationships with Twitter's sentiment.

- Volatility v.s sentiment score

Equity	Level of statistical significance (p-value)			
	ssr F test	ssr chi2 test	likelihood ratio test	parameter F test
ALGN	0.5127	0.4626	0.4659	0.5127
AMP	0.4853	0.1795	0.2148	0.4853
AON	0.3720	0.3295	0.3342	0.3720
BRK.B	0.5713	0.5426	0.5440	0.5713
CMCSA	0.9047	0.8984	0.8984	0.9047
CMI	0.5594	0.4839	0.4883	0.5594
CNP	0.8981	0.8701	0.8702	0.8981
DG	0.8843	0.8465	0.8467	0.8843
EXR	0.7013	0.6737	0.6744	0.7013
JKHY	0.9703	0.9605	0.9605	0.9703
JPM	0.8562	0.8460	0.8460	0.8562
LHX	0.6197	0.4774	0.4851	0.6197

We cannot find any significant relationship between the volatility and the sentiment score for these 12 stocks.

- Volume v.s. sentiment score

Equity	Level of statistical significance (p-value)			
	ssr F test	ssr chi2 test	likelihood ratio test	parameter F test
ALGN	✗ 0.5801	✗ 0.5356	✗ 0.5377	✗ 0.5801
AMP	✗ 0.3619	✓ 0.0638	✗ 0.1059	✗ 0.3619
AON	✗ 0.5387	✗ 0.5040	✗ 0.5060	✗ 0.5387
BRK.B	✗ 0.4930	✗ 0.4602	✗ 0.4625	✗ 0.4930
CMCSA	✗ 0.2326	✗ 0.1968	✗ 0.2031	✗ 0.2326
CMI	✗ 0.1359	✓ 0.0612	✓ 0.0759	✗ 0.1395
CNP	✓ 0.0767	✓ 0.0089	✓ 0.0241	✓ 0.0767
DG	✗ 0.7028	✗ 0.6091	✗ 0.6120	✗ 0.7028
EXR	✗ 0.2995	✗ 0.2485	✗ 0.2558	✗ 0.2995
JKHY	✓ 0.0627	✓ 0.0025	✓ 0.0137	✓ 0.0627
JPM	✗ 0.1088	✓ 0.0773	✓ 0.0857	✗ 0.1088
LHX	✗ 0.1598	✓ 0.0226	✓ 0.0486	✗ 0.1598

There exist many equities whose volumes are relevant to the Twitter sentiment, including AMP, CMI, CNP, JKHY, JPM, and LHX.

- Conclusion
  - Twitter's sentiment scores Granger-cause excess of log-returns and volumes for a subset of companies.
  - Twitter's analytics showed a very weak relationship with volatility.
  - For further studies, we can try to achieve more Twitter data to better the model's performance. For example, the Twitter API now can only give us the data in the most recent 7 days, while if we have more historical data, we can increase the lags and build greater models. Also, the Twitter data are fetched only by using company names and nicknames now, and we can try some other relevant keywords such as industry names, product names, and relevant company names in the future.
  - Our test is an hourly test, while we may get a better result if we have more data to do the daily test.