Sentiment-Based Stock Market Prediction

By Gauri Narayan & Raefah Wahid

Goal

Our goal for this project is to evaluate the different models that can be used to tackle the problem of sentiment-based stock market prediction. We will begin with a simple Naive Bayes model for sentiment prediction as a baseline, and then follow it with a continuous Dirichlet Process Mixture Model for topic-based sentiment prediction. We will then use Vector Autoregression to evaluate how well these models' outputs work in forecasting stock market closing prices.

Data: Stocks

	Date	Open	High	Low	Close	Adj Close	Volume
0	2018-12-31	1510.800049	1520.76001	1487.0	1501.969971	1501.969971	6954500
1	2019-01-02	1465.199951	1553.359985	1460.930054	1539.130005	1539.130005	7983100
2	2019-01-03	1520.01001	1538.0	1497.109985	1500.280029	1500.280029	6975600
3	2019-01-04	1530.0	1594.0	1518.310059	1575.390015	1575.390015	9182600
4	2019-01-07	1602.310059	1634.560059	1589.189941	1629.51001	1629.51001	7993200
5	2019-01-08	1664.689941	1676.609985	1616.609985	1656.579956	1656.579956	8881400
6	2019-01-09	1652.97998	1667.800049	1641.400024	1659.420044	1659.420044	6348800
7	2019-01-10	1641.01001	1663.25	1621.619995	1656.219971	1656.219971	6507700
8	2019-01-11	1640.550049	1660.290039	1636.219971	1640.560059	1640.560059	4686200
9	2019-01-14	1615.0	1648.199951	1595.150024	1617.209961	1617.209961	6005900
10	2019-01-15	1632.0	1675.160034	1626.01001	1674.560059	1674.560059	5998500

Data: Tweets

	Time	Text
0	2018-12-31 23:59:56+00:00	dang amazon. i talking customer service the phone they transfer to somewhere else. i waited minutes just hanging up. they pı
1	2020-11-25 17:06:25+00:00	learn the new cloud like way manage premise databases this depth article rds on vmware cloud solution architect blogger sat
2	2018-12-31 23:59:53+00:00	chacousa amazonpay i ordered pair chaco's sunday paid via amazon pay never got a confirmation shipment email idk to get
3	2018-12-31 23:59:43+00:00	check loiygit amazon music
4	2018-12-31 23:59:41+00:00	head banging doll [clean] kakicchysmusic mp downloads amazon prime
5	2018-12-31 23:59:35+00:00	kris sacrebleu i out luck. i wanted see roman j israel esq movie denzel you either subscribe to starz amazon. if i wait years i ca
6	2018-12-31 23:59:32+00:00	elon musk, amazon bezos fed powell land blunders bright spot list foxbusiness
7	2018-12-31 23:59:28+00:00	chipsandgist disney. apple. amazon. might be same it opens new lane. but definitely disney. these three the ones the deep pc
8	2018-12-31 23:59:25+00:00	i want give shoutout rippinhv real quick. this dude given mad amazon gift cards it's absolutely incredible. go show some love.
9	2018-12-31 23:59:21+00:00	amount books i buy makes amazon suggestions embarrassing
10	2018-12-31 23:59:14+00:00	amazon gift card giveaway ....ends in hours..............

Modeling: Naive Bayes

Modeling: Naive Bayes

$$\hat{y} = \frac{p(S_k) \cdot \prod_{i=1}^{n} p(x_i \mid S_k)}{\prod_{i=1}^{n} p(x_i)}$$

Sentiment
[67.04713531202752, 72.79504476422076]
[64.00006532285462, 84.31429844722402]
[65.39425411784755, 41.31187851313434]
[8.880502774415412, 7.690295085054137]
[33.09958337103018, 16.334416620769655]
[40.00814028825592, 25.074408905036936]
[30.313795039874066, 31.156117024623768]
[38.226439802744444, 30.108422772040566]
[43.56410663649677, 48.24644724010975]
[20.069533587621482, 24.011507977056585]
[19.477051769546197, 20.888468286185866]

Modeling: Vector Autoregression (VAR)

	curr_close	prev_close	pos_sentiment	neg_sentiment
0	1539.130005	1501.969971	46.400108	34.220366
1	1500.280029	1539.130005	138.404674	97.169707
2	1575.390015	1500.280029	98.321253	90.279607
3	1629.510010	1575.390015	55.811772	41.735355
4	1656.579956	1629.510010	29.608060	21.908681

$$y_t = \theta_1 x_{t-\text{lag}} + \theta_2 y_{t-\text{lag}} + b$$

Modeling: Vector Autoregression (VAR)

```
Amazon with lag = 1 and positive sentiment:
[iteration 0050] loss: 401709.9062
[iteration 0100] loss: 148309.6406
[iteration 0150] loss: 144385.6250
[iteration 0200] loss: 143009.9844
[iteration 0250] loss: 142243.5938
[iteration 0300] loss: 141869.2188
[iteration 0350] loss: 141706.6406
[iteration 0400] loss: 141643.0000
[iteration 0450] loss: 141619.8750
[iteration 0500] loss: 141611.5938
[iteration 0550] loss: 141608.0156
[iteration 0600] loss: 141605.7188
[iteration 0650] loss: 141603.7188
[iteration 0700] loss: 141601.7188
[iteration 0750] loss: 141599.5625
[iteration 0800] loss: 141597.2812
[iteration 0850] loss: 141594.9844
[iteration 0900] loss: 141592.5781
[iteration 0950] loss: 141590.0625
[iteration 1000] loss: 141587.4531
[iteration 1050] loss: 141584.6562
[iteration 1100] loss: 141581.7812
[iteration 1150] loss: 141578.8906
[iteration 1200] loss: 141575.8125
[iteration 1250] loss: 141572.6719
[iteration 1300] loss: 141569.3594
[iteration 1350] loss: 141565.9844
[iteration 1400] loss: 141562.5000
[iteration 1450] loss: 141558.8281
[iteration 1500] loss: 141555.0781
Learned parameters:
weight [[0.99737793 0.11895032]]
bias [1.0641915]
```

```
Amazon with lag = 1 and negative sentiment:
[iteration 0050] loss: 6927855.0000
[iteration 0100] loss: 272449.4062
[iteration 0150] loss: 253917.0938
[iteration 0200] loss: 235364.7344
[iteration 0250] loss: 217731.5625
[iteration 0300] loss: 201788.7188
[iteration 0350] loss: 187959.4531
[iteration 0400] loss: 176381.9688
[iteration 0450] loss: 166991.7656
[iteration 0500] loss: 159594.8906
[iteration 0550] loss: 153925.8125
[iteration 0600] loss: 149693.1719
[iteration 0650] loss: 146612.1094
[iteration 0700] loss: 144423.9375
[iteration 0750] loss: 142907.2031
[iteration 0800] loss: 141880.6250
[iteration 0850] loss: 141202.2031
[iteration 0900] loss: 140764.3438
[iteration 0950] loss: 140488.2500
[iteration 1000] loss: 140318.2969
[iteration 1050] loss: 140216.0625
[iteration 1100] loss: 140155.8438
[iteration 1150] loss: 140121.2031
[iteration 1200] loss: 140101.5781
[iteration 1250] loss: 140090.6875
[iteration 1300] loss: 140084.4844
[iteration 1350] loss: 140080.9531
[iteration 1400] loss: 140078.8594
[iteration 1450] loss: 140077.4688
[iteration 1500] loss: 140076.4531
Learned parameters:
weight [[0.9971941  0.17243722]]
bias [1.358112]
```

Modeling: Vector Autoregression (VAR)

```
Amazon with lag = 3 and positive sentiment:
[iteration 0050] loss: 702268.7500
[iteration 0100] loss: 396466.5625
[iteration 0150] loss: 394371.7812
[iteration 0200] loss: 393508.0312
[iteration 0250] loss: 392991.9375
[iteration 0300] loss: 392716.2500
[iteration 0350] loss: 392581.1250
[iteration 0400] loss: 392517.7500
[iteration 0450] loss: 392486.4375
[iteration 0500] loss: 392467.7812
[iteration 0550] loss: 392453.1875
[iteration 0600] loss: 392439.5938
[iteration 0650] loss: 392425.6875
[iteration 0700] loss: 392411.1562
[iteration 0750] loss: 392395.9375
[iteration 0800] loss: 392380.0938
[iteration 0850] loss: 392363.4062
[iteration 0900] loss: 392346.0625
[iteration 0950] loss: 392327.8438
[iteration 1000] loss: 392309.0312
[iteration 1050] loss: 392289.3750
[iteration 1100] loss: 392268.9375
[iteration 1150] loss: 392247.8750
[iteration 1200] loss: 392225.9375
[iteration 1250] loss: 392203.2500
[iteration 1300] loss: 392179.6562
[iteration 1350] loss: 392155.3438
[iteration 1400] loss: 392130.2812
[iteration 1450] loss: 392104.4062
[iteration 1500] loss: 392077.5000
Learned parameters:
weight [[0.99826187 0.11148553]]
bias [1.7639569]
```

```
Amazon with lag = 3 and negative sentiment:
[iteration 0050] loss: 5900669.0000
[iteration 0100] loss: 526575.9375
[iteration 0150] loss: 492494.8438
[iteration 0200] loss: 472000.3125
[iteration 0250] loss: 453515.9375
[iteration 0300] loss: 437771.4375
[iteration 0350] loss: 424994.7188
[iteration 0400] loss: 415055.6250
[iteration 0450] loss: 407614.7500
[iteration 0500] loss: 402239.1562
[iteration 0550] loss: 398484.4375
[iteration 0600] loss: 395945.7500
[iteration 0650] loss: 394282.4688
[iteration 0700] loss: 393225.4375
[iteration 0750] loss: 392573.5312
[iteration 0800] loss: 392182.5625
[iteration 0850] loss: 391954.4062
[iteration 0900] loss: 391824.1250
[iteration 0950] loss: 391750.9688
[iteration 1000] loss: 391709.8125
[iteration 1050] loss: 391686.3125
[iteration 1100] loss: 391671.7188
[iteration 1150] loss: 391661.7500
[iteration 1200] loss: 391653.9062
[iteration 1250] loss: 391647.0625
[iteration 1300] loss: 391640.5625
[iteration 1350] loss: 391633.8438
[iteration 1400] loss: 391627.1562
[iteration 1450] loss: 391620.1562
[iteration 1500] loss: 391613.1562
Learned parameters:
weight [[0.9989431  0.13550514]]
bias [1.2014734]
```

Modeling: Vector Autoregression (VAR)

```
Amazon with lag = 5 and positive sentiment:
[iteration 0050] loss: 867087.8750
[iteration 0100] loss: 614626.3125
[iteration 0150] loss: 612007.3750
[iteration 0200] loss: 611329.8125
[iteration 0250] loss: 610919.8750
[iteration 0300] loss: 610692.4375
[iteration 0350] loss: 610572.8125
[iteration 0400] loss: 610508.5000
[iteration 0450] loss: 610468.8750
[iteration 0500] loss: 610438.7500
[iteration 0550] loss: 610410.8750
[iteration 0600] loss: 610383.1875
[iteration 0650] loss: 610354.0625
[iteration 0700] loss: 610323.7500
[iteration 0750] loss: 610291.8125
[iteration 0800] loss: 610258.3750
[iteration 0850] loss: 610223.1875
[iteration 0900] loss: 610186.8750
[iteration 0950] loss: 610148.6875
[iteration 1000] loss: 610109.0625
[iteration 1050] loss: 610067.8750
[iteration 1100] loss: 610024.9375
[iteration 1150] loss: 609980.5625
[iteration 1200] loss: 609934.7500
[iteration 1250] loss: 609887.0000
[iteration 1300] loss: 609837.4375
[iteration 1350] loss: 609786.5000
[iteration 1400] loss: 609733.6250
[iteration 1450] loss: 609679.2500
[iteration 1500] loss: 609622.9375
Learned parameters:
weight [[0.9956019  0.25620225]]
bias [2.0007343]
```

```
Amazon with lag = 5 and negative sentiment:
[iteration 0050] loss: 2735914.7500
[iteration 0100] loss: 764938.4375
[iteration 0150] loss: 715818.8125
[iteration 0200] loss: 689414.1875
[iteration 0250] loss: 666847.3750
[iteration 0300] loss: 648767.6250
[iteration 0350] loss: 635069.1250
[iteration 0400] loss: 625189.0000
[iteration 0450] loss: 618377.3125
[iteration 0500] loss: 613876.0000
[iteration 0550] loss: 611018.6250
[iteration 0600] loss: 609272.7500
[iteration 0650] loss: 608245.2500
[iteration 0700] loss: 607660.3750
[iteration 0750] loss: 607337.6250
[iteration 0800] loss: 607163.2500
[iteration 0850] loss: 607069.3750
[iteration 0900] loss: 607017.5000
[iteration 0950] loss: 606986.5000
[iteration 1000] loss: 606965.2500
[iteration 1050] loss: 606948.1250
[iteration 1100] loss: 606932.8750
[iteration 1150] loss: 606917.7500
[iteration 1200] loss: 606902.5625
[iteration 1250] loss: 606887.0625
[iteration 1300] loss: 606870.9375
[iteration 1350] loss: 606854.4375
[iteration 1400] loss: 606837.1250
[iteration 1450] loss: 606819.3750
[iteration 1500] loss: 606800.9375
Learned parameters:
weight [[0.99651444 0.29964444]]
bias [2.2148945]
```

Inference: Granger Causality at lag = 1

Positive sentiment:

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.3122  , p=0.5768  , df_denom=248, df_num=1
ssr based chi2 test:   chi2=0.3160  , p=0.5740  , df=1
likelihood ratio test: chi2=0.3158  , p=0.5741  , df=1
parameter F test:         F=0.3122  , p=0.5768  , df_denom=248, df_num=1
```

Negative sentiment:

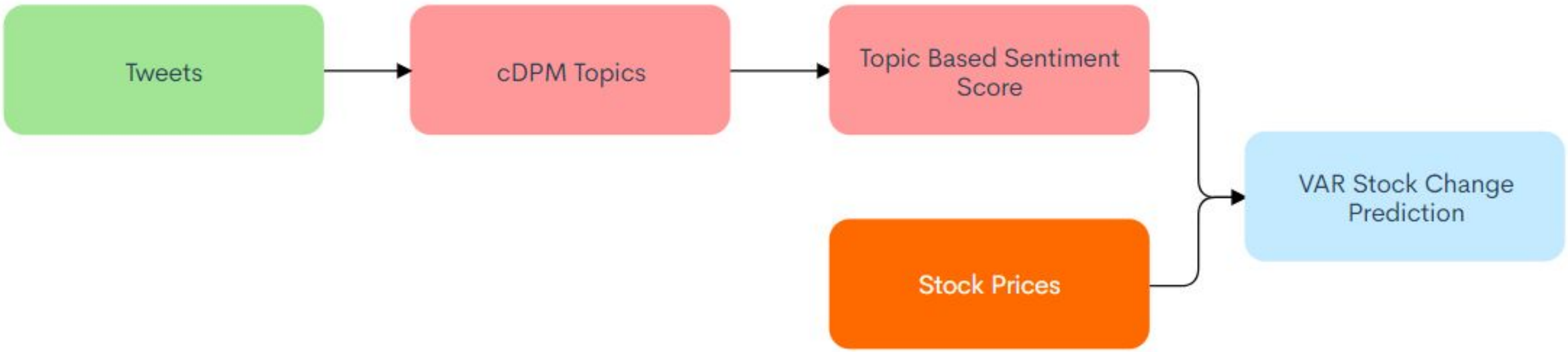
```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=1.8893  , p=0.1705  , df_denom=248, df_num=1
ssr based chi2 test:   chi2=1.9121  , p=0.1667  , df=1
likelihood ratio test: chi2=1.9049  , p=0.1675  , df=1
parameter F test:         F=1.8893  , p=0.1705  , df_denom=248, df_num=1
```

Evaluation: Mean Squared Error (MSE)

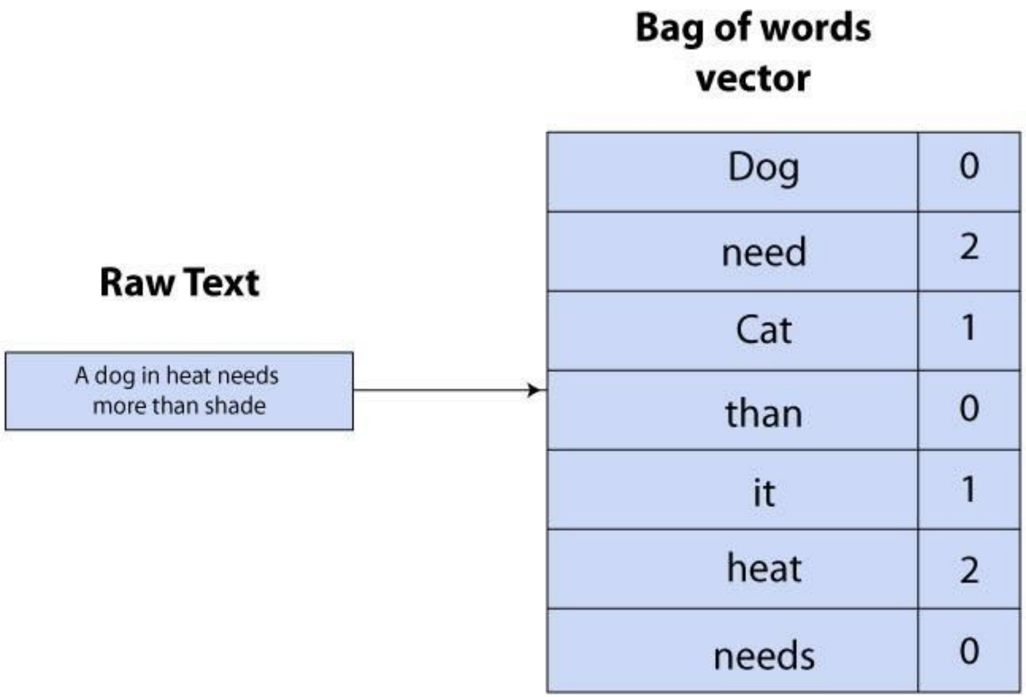
```
With lag = 1 :
The MSE of Amazon is 379.8229203350952 using positive sentiment as a parameter.
The MSE of Amazon is 367.1057093060366 using negative sentiment as a parameter.

With lag = 3 :
The MSE of Amazon is 924.2155785446265 using positive sentiment as a parameter.
The MSE of Amazon is 904.0989611903235 using negative sentiment as a parameter.

With lag = 5 :
The MSE of Amazon is 1387.7142995677307 using positive sentiment as a parameter.
The MSE of Amazon is 1364.1880103962494 using negative sentiment as a parameter.
```

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Modeling: Continuous DPM



Modeling: Continuous DPM



Modeling: Continuous DPM (base layer)

$$q(\beta, \theta, z) = \Pi_{k=1}^{K-1} q(\beta_k) \Pi_{k=1}^{K} q_k(\theta_k) \Pi_{n=1}^{N} q_n(z)$$

With $k = 1, \ldots K$ being the topics, for each observation $x_1, \ldots, x_N \in \mathbb{R}^C$

$$x_i | z_i, \theta_i \sim Mult(\theta_{z_i})$$

$$z_i | \beta \sim Categorical(stickbreak(\beta))$$

$$\theta_i | G \sim G_0$$

$$G_0 | \tau \sim Dirichlet((\tau_1, \ldots, \tau_C) = (1/C, \ldots, 1/C)$$

$$\beta \sim Beta(1, \kappa = \alpha).$$

Our variational parameters will therefore be τ, ϕ, and κ. We will sample them from

$$\phi \sim Dirichlet(1/K, \ldots, 1/K)$$

$$\tau_k \sim Normal(0.5, 0.25)$$

$$\kappa \sim Unif(0, 2).$$

Modeling: Continuous DPM (sequential layer)

1) New topic:

$$\theta_i \sim Dirichlet(1/C, \ldots, 1/C)$$

2) New topic Linked to old topic:

$$\theta_i \sim Dirichlet(\tau_{prev})$$

3) Old topic

$$\theta_i = \theta_{prev}$$

Inference: Granger Causality

Positive sentiment:

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.3675  , p=0.5449  , df_denom=248, df_num=1
ssr based chi2 test:   chi2=0.3719  , p=0.5420  , df=1
likelihood ratio test: chi2=0.3717  , p=0.5421  , df=1
parameter F test:         F=0.3675  , p=0.5449  , df_denom=248, df_num=1
```

Negative sentiment:

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.0201  , p=0.8873  , df_denom=248, df_num=1
ssr based chi2 test:   chi2=0.0204  , p=0.8865  , df=1
likelihood ratio test: chi2=0.0204  , p=0.8865  , df=1
parameter F test:         F=0.0201  , p=0.8873  , df_denom=248, df_num=1
The MSE of Amazon is 329.7234794464337 when using positive sentiment as a parameter.
The MSE of Amazon is 329.7234794464337 when using negative sentiment as a parameter.
```

Evaluation: Mean Squared Error (MSE)

```
 Amazon with lag = 1 and positive sentiment:
Learned parameters:
weight [[1.000046   0.08225729]]
bias [0.6987214]

 Amazon with lag = 1 and negative sentiment:
Learned parameters:
weight [[0.99963987 0.19081306]]
bias [1.2685877]
```

```
The MSE of Amazon is 329.7234794464337 when using positive sentiment as a parameter.
The MSE of Amazon is 329.7234794464337 when using negative sentiment as a parameter.
```

Conclusion

- P-values suggest no correlation, but there is a large decrease in the p-values associated with negative sentiment compared to positive sentiment, suggesting that some correlation exists and could be better captured with more refined models.
- The Naive Bayes allows us a simple baseline. When we compare DPM to it, we see from the MSE that the improvement seems to be marginal.
- Since DPM is computationally heavy, we were only able to find topic-based sentiment over the month of January. With a larger dataset and further work, such as hyperparameter tuning, increasing the maximum number of topics, and scaling the prior values of our parameters, we could further improve the results produced by DPM.

References

[1] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. \textit{Journal of Computational Science}, 2(1), 1-8. doi:10.1016/j.jocs.2010.12.007

[2] Si, J., Mukherjee, A., Liu B., Li, Q., Li, H., & Deng, X. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. 10.13140/2.1.3604.7043.

[3] Sun, Y., Gupta, M., Tang, J., Zhao B., Han, J. (2010). Community Evolution Detection in Dynamic Heterogeneous Information Networks. 10.1145/1830252.1830270.

[4] The linear regression used in the VAR model was based on the Pyro module example (https://docs.pyro.ai/en/stable/_modules/pyro/nn/module.html).

[5] The DPM model was based on the stick-breaking formulation in the Pyro documentation example (https://pyro.ai/examples/dirichlet_process_mixture.html).

[6] Porter, Andrew. Yahoo-historical. https://github.com/AndrewRPorter/yahoo-historical