NYU

Sentiment & Capital: The Data Behind Financial Movements

Group 14
12/10/24

Satyanarayna Chillale	sc9960
Srivats Poddar	sp7811
Rahul Raman	rr4549
Varun Tirthani	vyt2006

Abstract

In today's age where financial indicators are worth their weight in gold, we are experimenting with legacy data in the form of bitcoin and personal loan data as well as tweets involving publicly-traded companies to test and understand their impact on the market prices of stocks and their level of interdependence with each other

Platforms: Google Dataproc, Zeppelin

Motivation

- **Who are the users of this analytic?** Hedge Funds, Portfolio Managers, Equity Researchers
- **Who will benefit from this analytic?** Anyone enthusiastic about capital markets and curious to know which factors can actually significantly move markets
- **Why is this analytic important?** Those familiar with finance would be vaguely aware of the factors that could potentially affect market prices but might not be aware of their magnitude and direction and hence assigning estimate numerical values to this solidifies our understanding of the impact

Goodness

What steps were taken to assess the ‘goodness’ of the analytic?

- Our data sources were spread across different time intervals so we ensured to perform our analysis and derived insights over a time interval common across datasets (2015-2018).

- We did correlation of stock prices and bitcoin prices and noticed that there was a strong correlation with ETF’s that had BTC in their portfolio.

NYU

Data Sources

Dataset Name	Description	Size (MB)
Stock Market	Nasdaq and NYSE – 2012 - 2018 (on day basis)	809.2
Twitter Tweets	Tweets – 2015 - 2020	772.34
Bitcoin Market	Cryptocurrencies – 2012 - 2018 (on minute basis)	348.52
Personal Loans	Loan applications – Source: Investors – 2007-2018	648.05

Data Sample - Stock Market

```
cleanedData.show(10)
```

date	stockName	open	close	low	high	volume	Symbol	Name	Country	Sector	Industry
2013-01-16	ABBV	29.308	30.127	29.206	30.127	1.947899E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-17	ABBV	30.308	30.875	30.308	31.113	1.7017768E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-18	ABBV	30.882	31.637	30.551	31.637	2.1049603E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-22	ABBV	31.346	30.875	30.457	31.577	1.864969E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-23	ABBV	31.291	32.044	30.942	32.076	1.2954549E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-24	ABBV	32.036	31.756	31.56	32.655	1.3362402E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-25	ABBV	31.806	31.874	31.059	32.076	1.0544411E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-28	ABBV	31.951	31.272	31.12	32.102	8422649.0	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-29	ABBV	31.12	31.63	30.73	31.756	1.0426869E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...
2013-01-30	ABBV	31.205	31.31	31.205	31.814	1.279017E7	ABBV	AbbVie Inc. Commo...	United States	Health Care	Biotechnology: Ph...

only showing top 10 rows

```
root
 |-- date: date (nullable = true)
 |-- stockName: string (nullable = true)
 |-- open: double (nullable = true)
 |-- close: double (nullable = true)
 |-- low: double (nullable = true)
 |-- high: double (nullable = true)
 |-- volume: double (nullable = true)
 |-- Symbol: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- Sector: string (nullable = true)
 |-- Industry: string (nullable = true)
```

NYU

Data Sample - Twitter Tweets

```
|-- tweet_id: string (nullable = true)
|-- writer: string (nullable = true)
|-- post_date: string (nullable = true)
|-- body: string (nullable = true)
|-- comment_num: string (nullable = true)
|-- retweet_num: string (nullable = true)
|-- like_num: string (nullable = true)
```

tweet_id	writer	post_date	body	comment_num	retweet_num	like_num
550441509175443456	VisualStockRSRC	1420070457	lx21 made $10,008...	0	0	1
550441672312512512	KeralaGuy77	1420070496	Insanity of today...	0	0	0
550441732014223360	DozenStocks	1420070510	S&P100 #Stocks Pe...	0	0	0
550442977802207232	ShowDreamCar	1420070807	$GM $TSLA: Volksw...	0	0	1
550443807834402816	i_Know_First	1420071005	Swing Trading: Up...	0	0	1
550443808606126081	aaplstocknews	1420071005	Swing Trading: Up...	0	0	1
550443809700851716	iknowfirst	1420071005	Swing Trading: Up...	0	0	1
550443857142611968	Cprediction	1420071016	Swing Trading: Up...	0	0	1
550443857595600896	iknowfirst_br	1420071017	Swing Trading: Up...	0	0	1
550443857692078081	Gold_prediction	1420071017	Swing Trading: Up...	0	0	1
550443858010861568	IKFResearch	1420071017	Swing Trading: Up...	0	0	1
550444112328261632	GetAOM	1420071077	$UNP $ORCL $QCOM ...	0	0	0
550444969924653056	AppleNewsAAPL	1420071282	$AAPL Apple goes ...	0	0	1
550444970738335744	espositooooo	1420071282	“@WSJ: Apple is b...	0	0	0

NYU

Data Sample - Bitcoin Market

	Date	High	Low	Open	Close	Volume	Average Price	5Day_SMA	28Day_SMA
	01/01/2012	5	4.58	4.58	5	20.1	4.86	4.86	4.86
	02/01/2012	5	5	5	5	19.05	5	4.93	4.93
	03/01/2012	5.32	5	5	5.29	88.04	5.2	5.02	5.02
	04/01/2012	5.57	4.93	5.29	5.57	107.23	5.36	5.11	5.11
	05/01/2012	6.65	5.57	5.57	6.65	94.8	6.29	5.34	5.34
	06/01/2012	6.9	6	6.65	6	33.88	6.3	5.63	5.5
	07/01/2012	6.8	6	6	6.8	0.3	6.53	5.94	5.65
	08/01/2012	7	6.8	6.8	7	5	6.93	6.28	5.81
	09/01/2012	7	6.23	7	6.3	66.87	6.51	6.51	5.89
	10/01/2012	7.14	6.24	6.3	7.14	62.29	6.84	6.62	5.98
	11/01/2012	7.33	6.25	7.14	7	105.36	6.86	6.74	6.06
	12/01/2012	7.38	6.51	7	6.51	82.3	6.8	6.79	6.12
	13/01/2012	7.36	6.51	6.51	6.6	48.97	6.82	6.77	6.18
	14/01/2012	6.6	6.3	6.6	6.3	16.84	6.4	6.74	6.19

```
finalDf.printSchema()

root
 |-- date: date (nullable = true)
 |-- High: double (nullable = true)
 |-- Low: double (nullable = true)
 |-- Open: double (nullable = true)
 |-- Close: double (nullable = true)
 |-- Volume: double (nullable = true)
 |-- AveragePrice: double (nullable = true)
 |-- 5Day_SMA: double (nullable = true)
 |-- 28Day_SMA: double (nullable = true)
```

NYU

Data Sample - Personal Loans

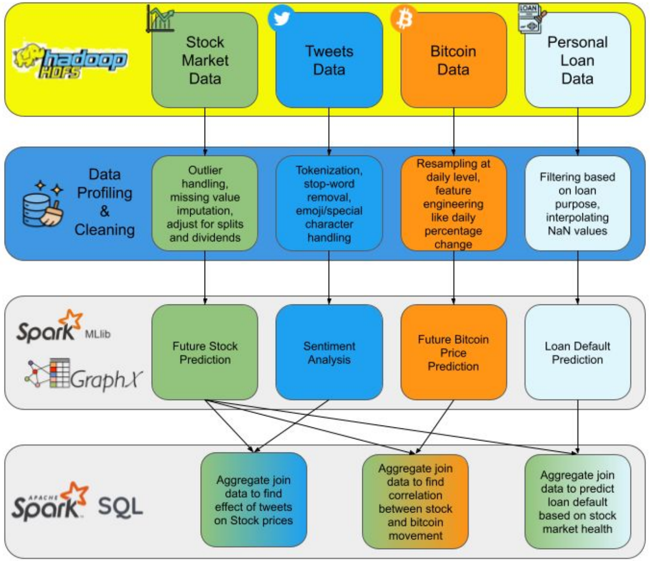
Amount Requested	Application Date	Loan Title	Risk_Score	Zip Code	State	emp_length_year	Debt-To-Income Ratio (%)
1000.0	2007-05-26	Wedding Covered b...	693.0	481xx	NM	4	10.0
1000.0	2007-05-26	Consolidating Debt	703.0	010xx	MA	1	10.0
11000.0	2007-05-27	Want to consolida...	715.0	212xx	MD	1	10.0
6000.0	2007-05-27	waksman	698.0	017xx	MA	1	38.0
1500.0	2007-05-27	mdrigo	509.0	209xx	MD	1	9.0
15000.0	2007-05-27	Trinfiniti	645.0	105xx	NY	3	0.0
10000.0	2007-05-27	NOTIFYi Inc	693.0	210xx	MD	1	10.0
3900.0	2007-05-27	For Justin.	700.0	469xx	IN	2	10.0
3000.0	2007-05-28	title?	694.0	808xx	CO	4	10.0
2500.0	2007-05-28	timgerst	573.0	407xx	KY	4	11.0
3900.0	2007-05-28	need to consolidate	710.0	705xx	LA	10	10.0
1000.0	2007-05-28	sixstrings	680.0	424xx	KY	1	10.0
3000.0	2007-05-28	bmoore5110	688.0	190xx	PA	1	10.0
1500.0	2007-05-28	MHarkins	704.0	189xx	PA	3	10.0
1000.0	2007-05-28	Moving	694.0	354xx	AL	1	10.0
8000.0	2007-05-28	Recent College Gr...	708.0	374xx	TN	1	10.0
12000.0	2007-05-29	FoundersCafe.com	685.0	770xx	TX	3	10.0
1000.0	2007-05-29	UChicago2004	698.0	207xx	MD	3	10.0
15000.0	2007-05-29	Cancer is Killing...	680.0	432xx	OH	1	10.0
5000.0	2007-05-29	2006-2007 College...	680.0	011xx	MA	1	10.0

only showing top 20 rows

Dataset Name	Rows	Columns	Size
accepted 2007 to 2018-Q4.csv.gz	2,260,701	151	392.6 MB
rejected 2007 to 2018-Q4.csv.gz	27,648,741	9	255.5 MB

NYU

Design Diagram



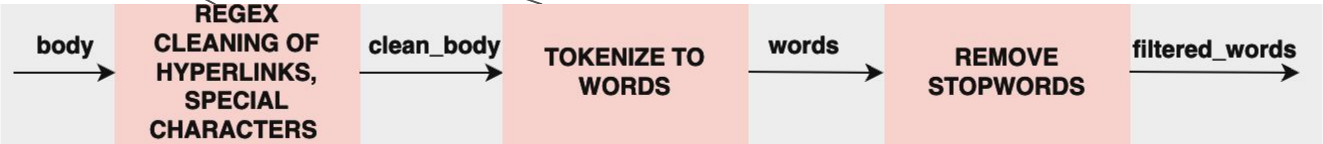
- Analyze **Effect of Tweets** on:
  - Bitcoin
  - Stock Market
- Analyze **Effect of Stock Market** on:
  - Bitcoin
- Analyze **Effect of general market trend** (Stock + Bitcoin) on:
  - Rejection of personal loans
- MLLib – **LR models** for predicting stocks
- Graphx – **shock propagation** between companies across sectors.

NYU

Code Challenge - Processing Tweets

```
import org.apache.spark.ml.feature.Tokenizer

val cleanDF = adjustedDateDF.withColumn("clean_body",
  regexp_replace(col("body"), "(@\\w+|http\\S+|[^a-zA-Z\\s$])", ""))

val tokenizer = new Tokenizer().setInputCol("clean_body").setOutputCol("words")

val tokenizedDF = tokenizer.transform(cleanDF)

val normalizedDF = tokenizedDF.withColumn("words",
  expr("transform(words, word -> lower(word))")
)

z.show(normalizedDF)
```

body	clean_body	words	filtered_words
Trailing Stop taken out on my $GOOGL #trade taking my locked in profits, Read My Full Trade Diary for ALL to view http://ow.ly/GDYGz	Trailing Stop taken out on my $GOOGL trade taking my locked in profits Read My Full Trade Diary for ALL to view	WrappedArray(trailing, stop, taken, out, on, my, $googl, trade, taking, my, locked, in, profits, read, my, full, trade, diary, for, all, to, view)	WrappedArray(trailing, stop, taken, $googl, trade, taking, locked, profits, read, full, trade, diary, view)



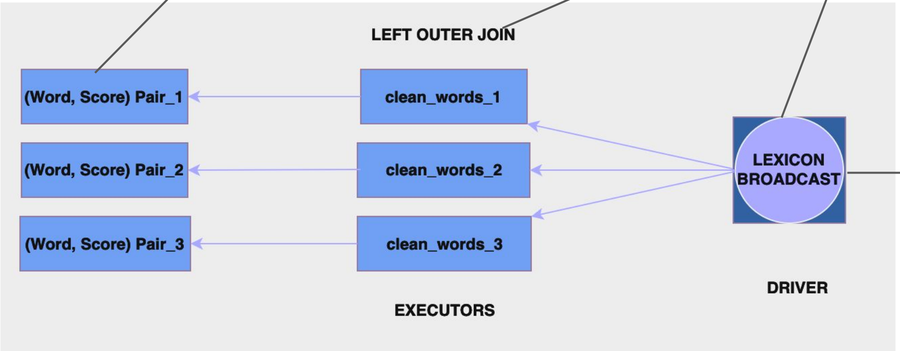

```
import org.apache.spark.ml.feature.StopWordsRemover

val remover = new StopWordsRemover().setInputCol("words").setOutputCol("filtered_words")

val filteredDF = remover.transform(tokenizedDF)

z.show(filteredDF)
```

NYU

Code Challenge - Lexicon Based Sentiment

filtered_words	clean_words	afinn_score
WrappedArray(, apple, sued, falsely, advertising, storage, capacity, iphones, , $aapl)	sued	-2
WrappedArray(, apple, sued, falsely, advertising, storage,	falsely	-2

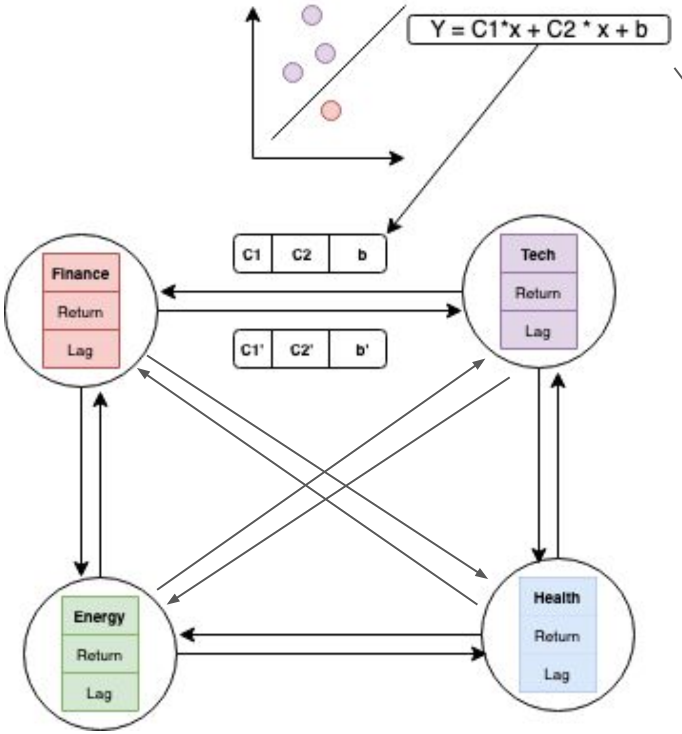
```
val broadcastAfinnDF = broadcast(afinnDF)

val enrichedDF = convertedDF
  .join(
    afinnDF.withColumnRenamed("score", "afinn_score"),
    convertedDF("clean_words") === afinnDF("word"),
    "left_outer"
  )
  .na.fill(0, Seq("afinn_score")) // Fill null scores with 0

z.show(enrichedDF)
```



word	score
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
abduction	-2
abductions	-2
abhor	-3
abhorred	-3

NYU

Code Challenge - Sector Shock Propogation




```scala
val resultGraph = graph.pregel(Seq[Double](),
maxIterations)(

  // Update Vertex:
  (id, oldValue, newMessages) => {
    val updatedValues = oldValue.map { case (sector,
avgReturn, avgLag1, avgLag2) =>
      val updatedAvgReturn: Double =
newMessages.sum.toDouble / newMessages.size.toDouble
      (sector, updatedAvgReturn, avgLag1, avgLag2)
    }
    updatedValues
  },

  // Send Message
  triplet => {
    val messages = triplet.srcAttr.map { case (_,
avgReturn, avgLag1, _) =>
      val (ce1, ce2, b) = triplet.attr
      avgReturn * ce1 + avgLag1 * ce2 + b
    }

    Iterator((triplet.dstId, messages))
  },

  // Merge Messages
  (msg1, msg2) => msg1 ++ msg2
)
```

Code Challenge - Moving Average and Correlation Analysis

```
val window5 = Window.orderBy("date").rowsBetween(-4, 0)
val window28 = Window.orderBy("date").rowsBetween(-27, 0)

val finalDf = withAveragePrice
    .withColumn("5Day_SMA", avg("AveragePrice").over(window5))
    .withColumn("28Day_SMA", avg("AveragePrice").over(window28))

z.show(finalDf)
```

```
val periods = Array(1, 3, 7, 14)
var returnsDF = mergedDF
for (period <- periods) {
  val futureClose = lead(col("Close"), period).over(Window.orderBy("date"))
  returnsDF = returnsDF.withColumn(
    s"${period}day_future_return",
    ((futureClose - col("Close")) / col("Close") * 100)
  )
}
```

Results

```
Price Correlations:
Positive: 0.565
Neutral: 0.627
Negative: 0.547

Return Correlations:

Positive Count:
1-day: -0.024
3-day: -0.047
7-day: -0.070
14-day: -0.091

Neutral Count:
1-day: -0.005
3-day: -0.011
7-day: -0.018
14-day: -0.023

Negative Count:
1-day: -0.002
3-day: -0.024
7-day: -0.042
14-day: -0.050

Total Positive Posts: 6436.0
Total Neutral Posts: 5254.0
Total Negative Posts: 1668.0
```

BTC vs Twitter Sentiment

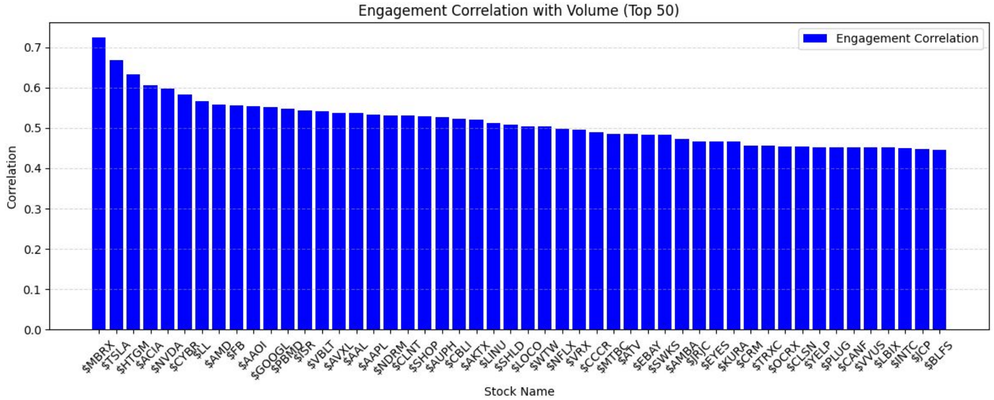
```
Top positively correlated stocks with BTC:
            correlation
arkk      0.950033
pypl      0.947104
ifly      0.945113
arkw      0.941155
race      0.924074
hcm       0.917740
sq        0.917648
arkq      0.913846
snc       0.913083
htht      0.911794

Top negatively correlated stocks with BTC:
             correlation
rely      -0.759078
pti       -0.773643
nndm      -0.775933
wmih      -0.784681
aemd      -0.801042
dhx       -0.805537
vtgn      -0.809197
oasm      -0.814145
chad      -0.823643
...
median_correlation: 0.4203
std_correlation: 0.4349
positive_correlations: 4000.0000
negative_correlations: 1643.0000
```

BTC vs Stocks

```
Top 50 nodes by degree count:
Rank | Node | Degree
------------------------------
   1 | spsm             | 409
   2 | pkw              | 393
   3 | jhmm             | 388
   4 | ptmc             | 386
   5 | wbid             | 348
   6 | dwas             | 329
   7 | csf              | 327
   8 | cfa              | 325
   9 | wbib             | 324
  10 | jhml             | 303
  11 | cfo              | 298
  12 | wbic             | 283
  13 | gbci             | 277
  14 | wbia             | 277
  15 | ptlc             | 263
  16 | pho              | 259
  17 | dhvw             | 257
  18 | ewmc             | 250
  19 | snv              | 242
  20 | jpus             | 241
```

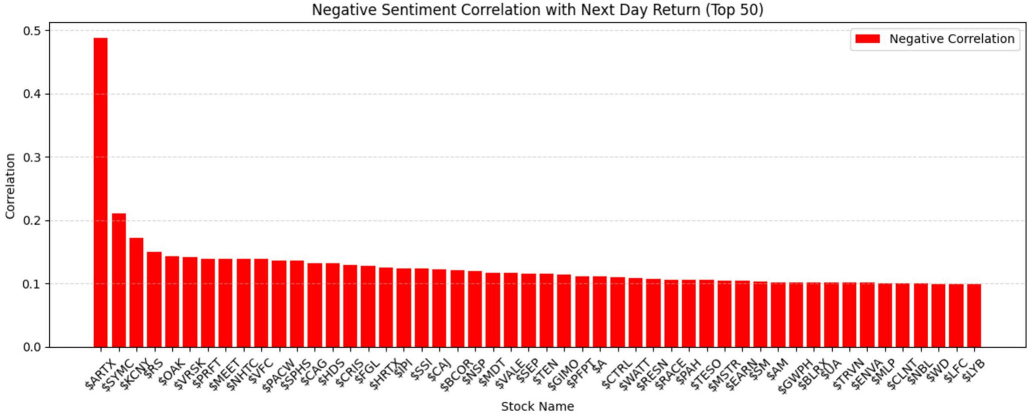
Most correlated stocks

Results



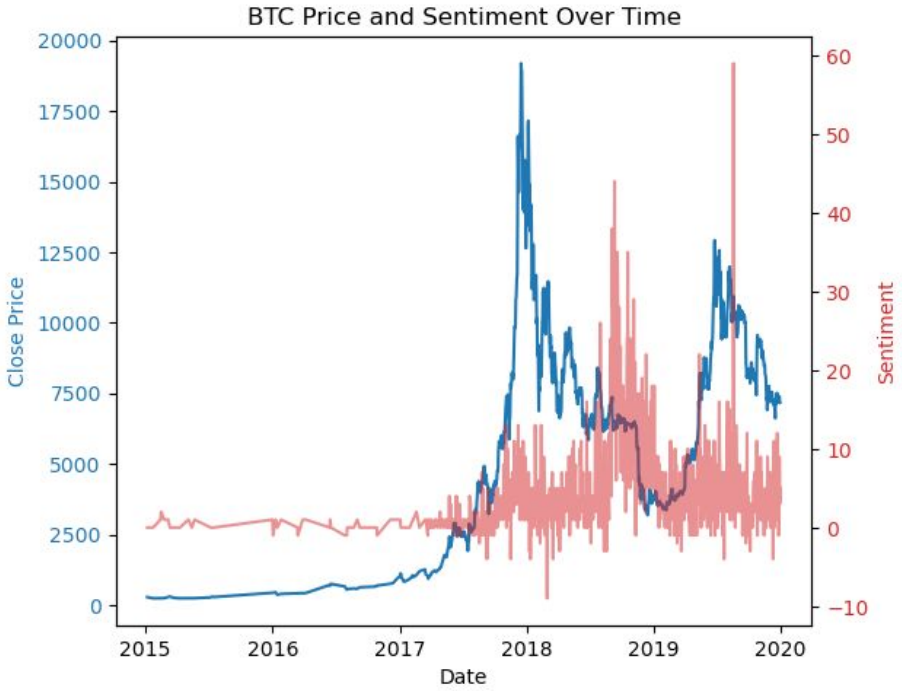
The graph on the left is correlation between volume traded vs the number of tweets for each stock. This graph shows the top 50 correlation stocks.

The graph on the right is correlation between next day return vs the negative sentiment on twitter for each stock. This graph shows the top 50 correlation stocks.

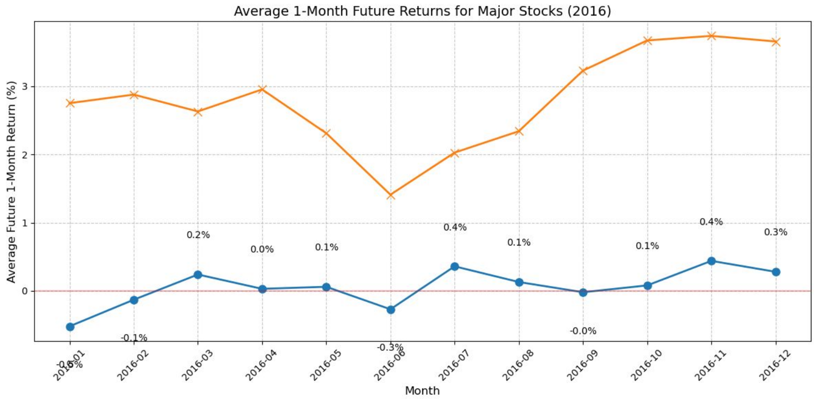


NYU

Results



BTC Price and Sentiment Over Time
Close Price
Sentiment
Date
20000
17500
15000
12500
10000
7500
5000
2500
0
60
50
40
30
20
10
0
−10
2015
2016
2017
2018
2019
2020



Average 1-Month Future Returns for Major Stocks (2016)
Average Future 1-Month Return (%)
Month
0
1
2
3
-0.1%
0.2%
0.0%
0.1%
-0.3%
0.4%
0.1%
-0.0%
0.1%
0.4%
0.3%
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12

Other Experiments

- Linear Regression Models to predict stock based on Returns and Volume.
- GraphX to predict the Ripple Effect of a sector affecting across multiple sectors.
- Analyze the datasets separately.

Obstacles

- LR models are **not ideal** for **time-series data**.
  - Spark Scala ML – no support for time series models (ARIMA).
  - Spark time-series – **Deprecated**, Open source library

- **Limited** Correlation statistics.
  - Spark Scala Stats – does not provide **hypothesis tests** like Granger causality and models like VAR.

- Multiple plots (line + scatter) – difficult to build this visualization using Zeppelin.

Acknowledgements

- Thank you to the NYU HPC for making great guides for how to use the HPC on their Google Sites webpage, and for providing the Spark cluster to conduct these analytics.

- Thank you to the Kaggle Collaborators to open source their data and share on Kaggle.

- Lastly, thank you to Professor Yang for all the support that you have provided us throughout the semester, and analyzing and approving our project idea!

References

- Kaggle Datasets
  - Stock Market
  - Bitcoin
  - Twitter
  - Personal loans
- Spark Scala
  - MLLib
  - GraphX
- Tutorial
  - Pregel API

NYU

Thank You