Deriving Stock Alpha From News

Ty Painter
Master's Capstone Project
Vanderbilt Data Science Institute

Framework

01

Data Collection

Where does the data come from?

02

Data Processing

What to do with the data?

03

Portfolio Construction

How is the data used to select stocks?

04

Results

How did the portfolio perform?

Data Collection

GDELT 2.0 GKG Database

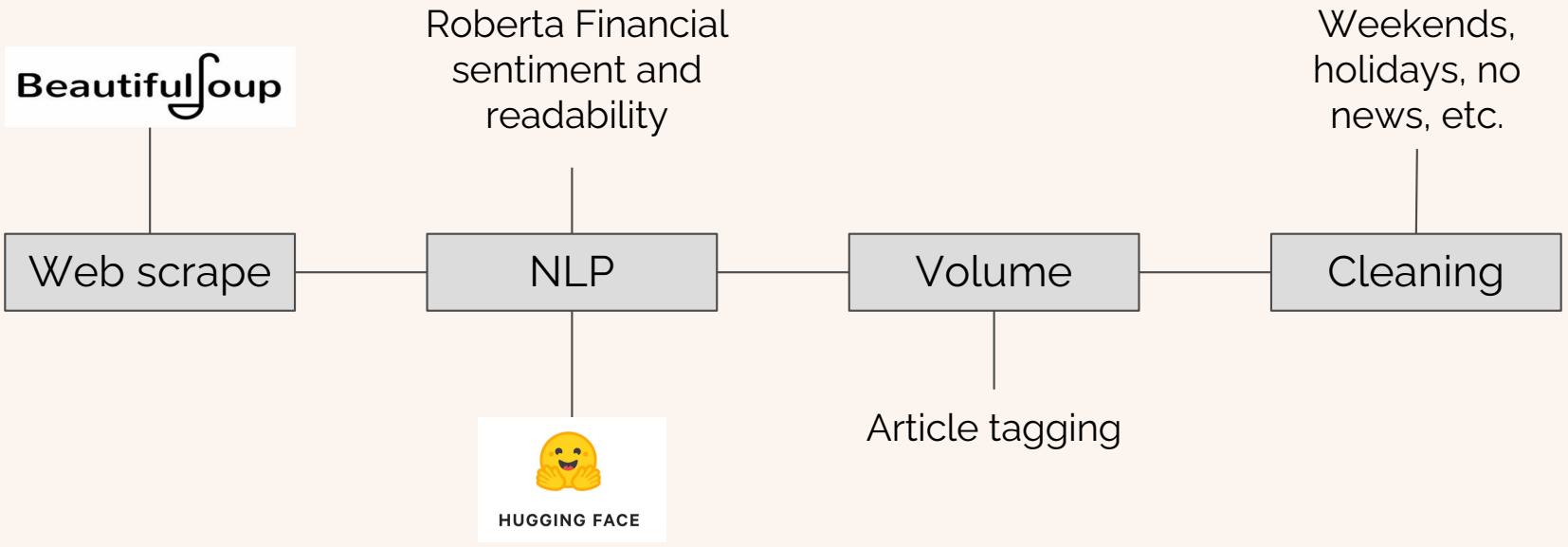
- Files produced every 15 minutes
- 135K records per day, 4.1M per month
- 26 columns → 3 columns
- Filtered based on website

Yahoo Finance

- 2021 S&P 500 individual stock daily returns

Rank	Website	Count
1	iheart.com	161,187
2	msn.com	81,981
3	medium.com	48,707
4	reuters.com	37,931
5	yahoo.com	28,885
6	indiatimes.com	26,354
7	prnewswire.com	22,544
8	dailymail.co.uk	18,859
9	texasguardian.com	18,031
10	apnews.com	17,731
11	sandiegosun.com	16,727
12	newyorktelegraph.com	16,668
13	ianslive.in	13,846
14	sfgate.com	12,787
15	chron.com	12,676
16	marketwatch.com	12,166
17	washingtontimes.com	11,970
18	onenewspage.com	11,206
19	thenews.com.pk	10,850
20	thehindu.com	10,523

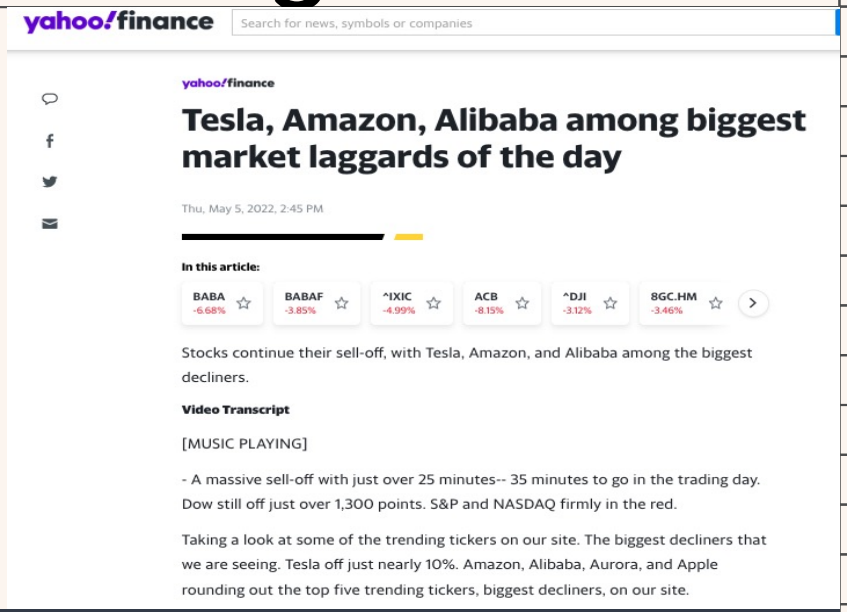
Data Processing



Data Processing

Web Scraping

- Daily open & close prices
  - *Selenium, ChromeDriverManager*
- News articles
  - *BeautifulSoup*
  - *~65K articles per month*
  - *Example Link*




```
'Stocks continue their sell-off, with Tesla, Amazon, and Alibaba among the biggest decliners.Video Transcript[MUSIC PLAYING]- A massive sell-off with just over 25 minutes-- 35 minutes to go in the trading day. Dow still off just over 1,300 points. S&P and NASDAQ firmly in the red.Taking a look at some of the trending tickers on our site. The biggest decliners that we are seeing. Tesla off just nearly 10%. Amazon, Alibaba, Aurora, and Apple rounding out the top five trending tickers, biggest decliners, on our site.'
```

Data Processing

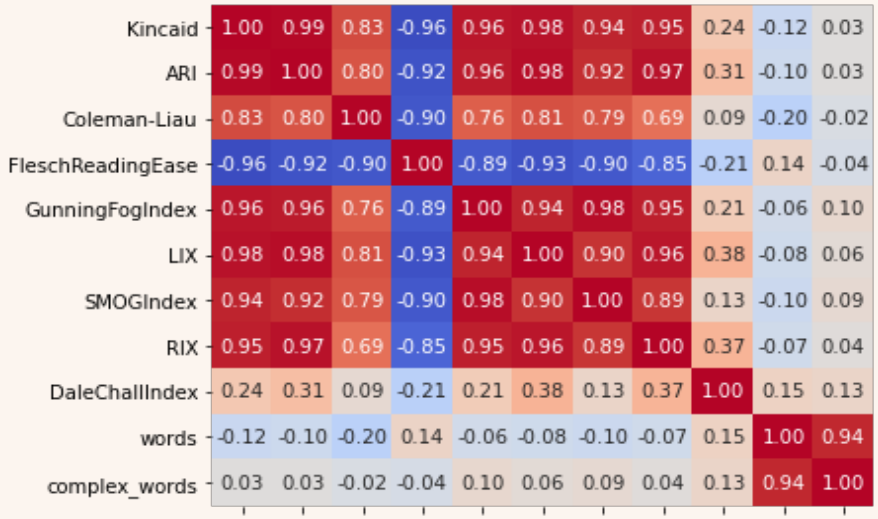
NLP

- Sentiment Analysis
  - HuggingFace → *distilroberta-financial-news-sentiment-analysis model*

sentence	negative	neutral	positive	net
Stocks continue their sell-off, with Tesla, Amazon, and Alibaba among the biggest decliners. Video Transcript[MUSIC PLAYING]- A massive sell-off with just over 25 minutes-- 35 minutes to go in the trading day	0.99225	0.00714	0.00061	-0.99164
Dow still off just over 1,300 points	0.99659	0.00246	0.00095	-0.99564
S&P and NASDAQ firmly in the red.Taking a look at some of the trending tickers on our site	0.00008	0.99981	0.00011	0.00002
The biggest decliners that we are seeing	0.00014	0.99979	0.00007	-0.00007
Tesla off just nearly 10%	0.99761	0.00185	0.00055	-0.99706
Amazon, Alibaba, Aurora, and Apple rounding out the top five trending tickers, biggest decliners, on our site.	0.00007	0.99987	0.00006	-0.00001

Data Processing

- [ ] NLP
  - Readability



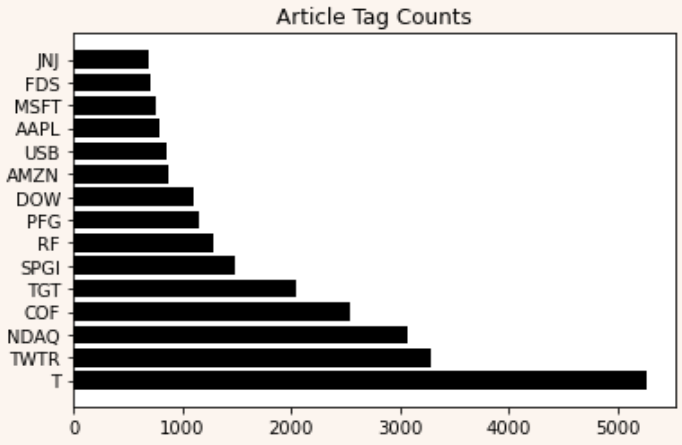

```
 Kincaid        ARI  Coleman-Liau  FleschReadingEase  GunningFogIndex
9.101261  11.293636     12.007044          62.419079        13.373497

      LIX  SMOGIndex       RIX  DaleChallIndex  words  complex_words
47.299289  12.114654  5.307692       11.179015    238             36
```

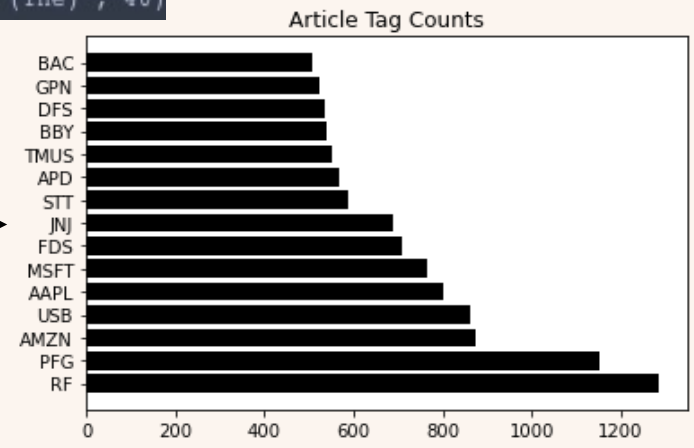
Data Processing

Volume

- Importance factor
- Fuzzywuzzy → token set ratio

```
('Amazon', 100),
('Dow', 100),
('Nasdaq', 100),
('Tesla', 100),
('AT&T', 67),
('S&P Global', 46),
('Snap-on', 44),
('Hartford (The)', 40)
```

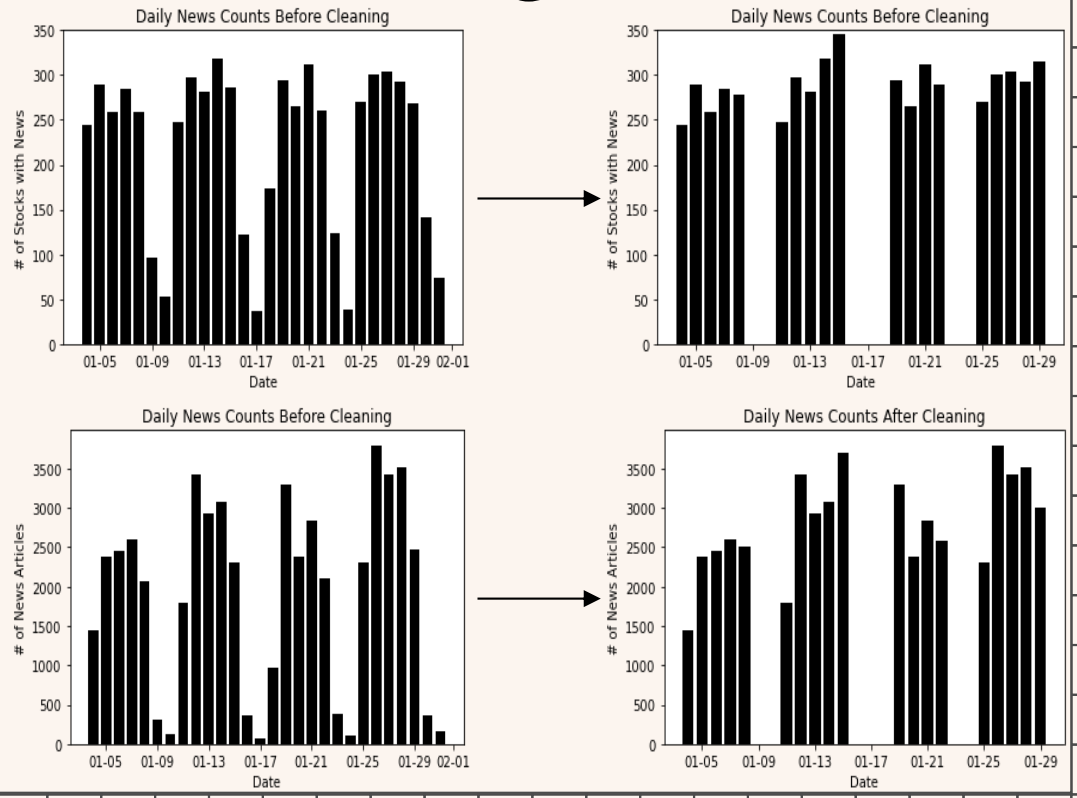




Data Processing

Data Cleaning

- Weekends & Holidays



Data Processing

Data Cleaning

- No news days
  - Decayed lagging days
  - 70% → 1 day prior
  - 20% → 2 days prior
  - 10% → 3 days prior

date	tags	ticker	volume	pos_mean	pos_median	pos_min	pos_max
2021-01-26 00:00:00	YUM	YUM	6.00	0.36	0.33	0.00	0.85
2021-01-27 00:00:00	YUM	YUM	1.00	0.53	0.83	0.00	1.00
2021-01-28 00:00:00	YUM	YUM	2.00	0.22	0.00	0.00	1.00
2021-01-29 00:00:00	N/A	YUM	0.00	N/A	N/A	N/A	N/A

date	tags	ticker	volume	pos_mean	pos_median	pos_min	pos_max
2021-01-26 00:00:00	YUM	YUM	6.00	0.36	0.33	0.00	0.85
2021-01-27 00:00:00	YUM	YUM	1.00	0.53	0.83	0.00	1.00
2021-01-28 00:00:00	YUM	YUM	2.00	0.22	0.00	0.00	1.00
2021-01-29 00:00:00	YUM	YUM	2.20	0.30	0.20	0.00	0.99

Data Processing

- [ ] Finalized Data Columns

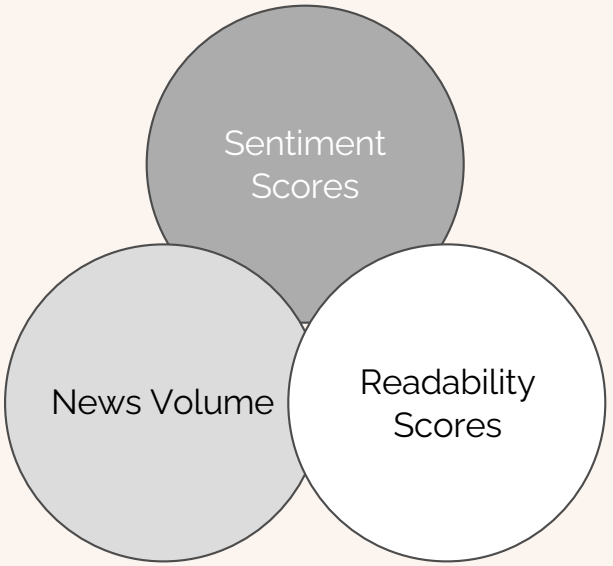
Descriptions	News		Financials
date	volume	pos_mean	same_day_raw
tags	pos_median	pos_min	next_day_raw
ticker	pos_max	neg_mean	sp_avg_return
	neg_median	neg_min	next_day_sp
	neg_max	neu_mean	same_day_relative
	neu_median	neu_min	next_day_relative
	neu_max	net_mean	
	net_median	net_min	
	net_max	read_score	

Portfolio Construction

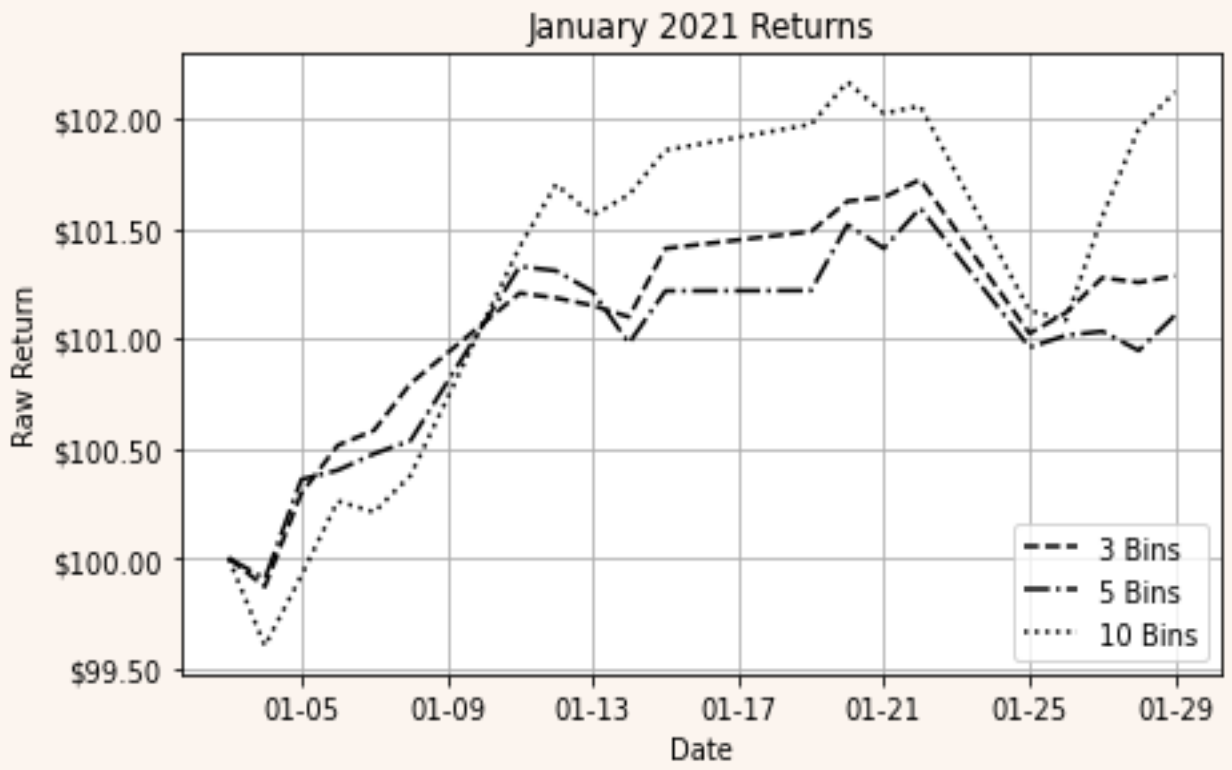
Long-Short Portfolio

- Ranked based on scaled, aggregate feature scores
- 3, 5, 10 bins
- Equally-weighted

Bin Size	# of Stocks	Positions
3	334	167 short
		167 long
5	202	101 short
		101 long
10	100	50 short
		50 long

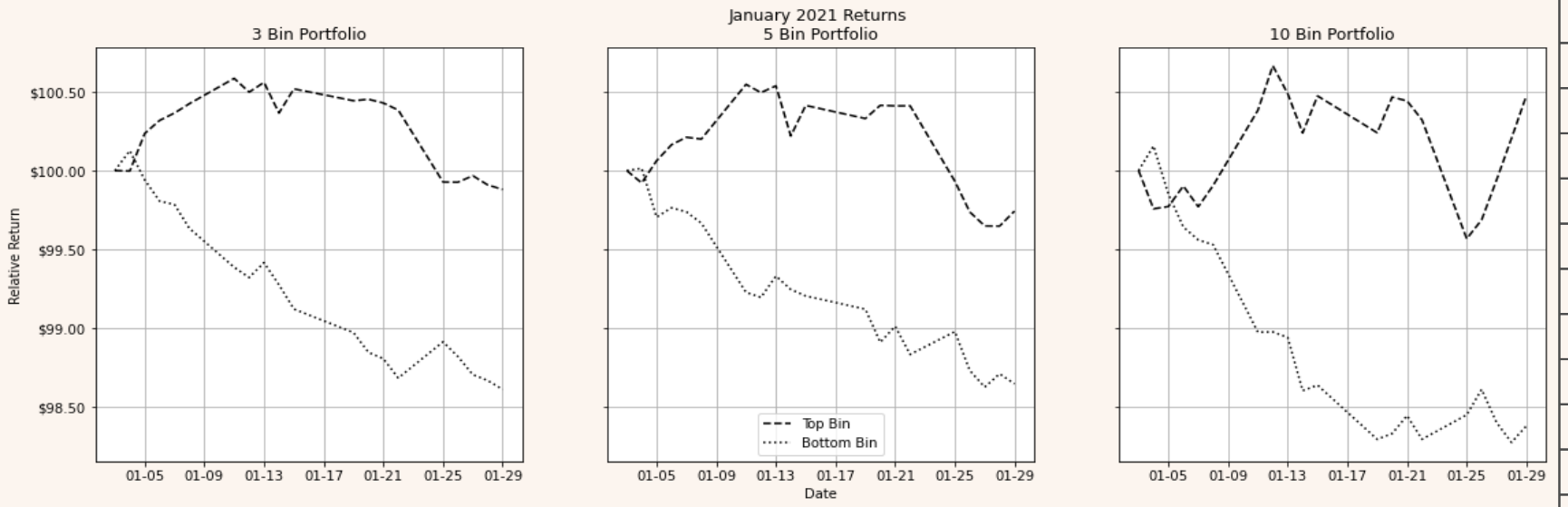


Results



Portfolio	Raw Return
3 Bins	+1.28%
5 Bins	+1.11%
10 Bins	+2.12%

Results



3 Bins	Relative Return
Top	-0.12%
Bottom	-1.39%

5 Bins	Relative Return
Top	-0.26%
Bottom	-1.35%

10 Bins	Relative Return
Top	+0.48%
Bottom	-1.62%

Thank you!

Acknowledgements
Andrew Chin
Michael Li
Dr. Jesse Blocher

Check back for final results!

Contact Information:
Ty Painter
tydpaint@gmail.com