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 Event Database

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

Yahoo Finance

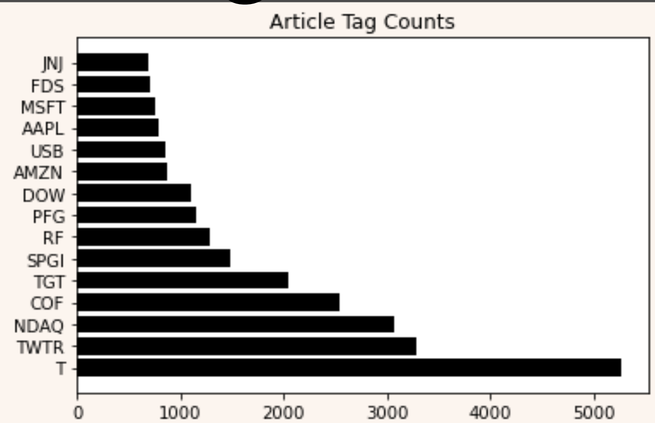
- 2021 S&P 500 individual stock daily returns

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

Beautiful Soup

Roberta Financial sentiment, tagging, and readability



Web scrape — NLP — Volume — Cleaning

HUGGING FACE

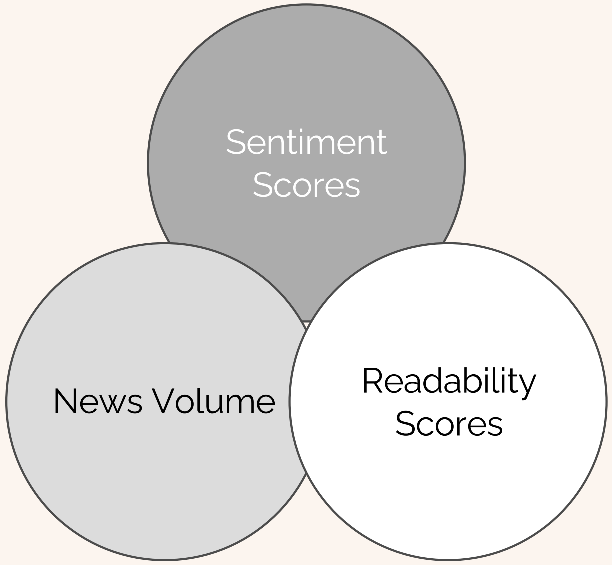
Article tagging

Weekends, holidays, no news etc.

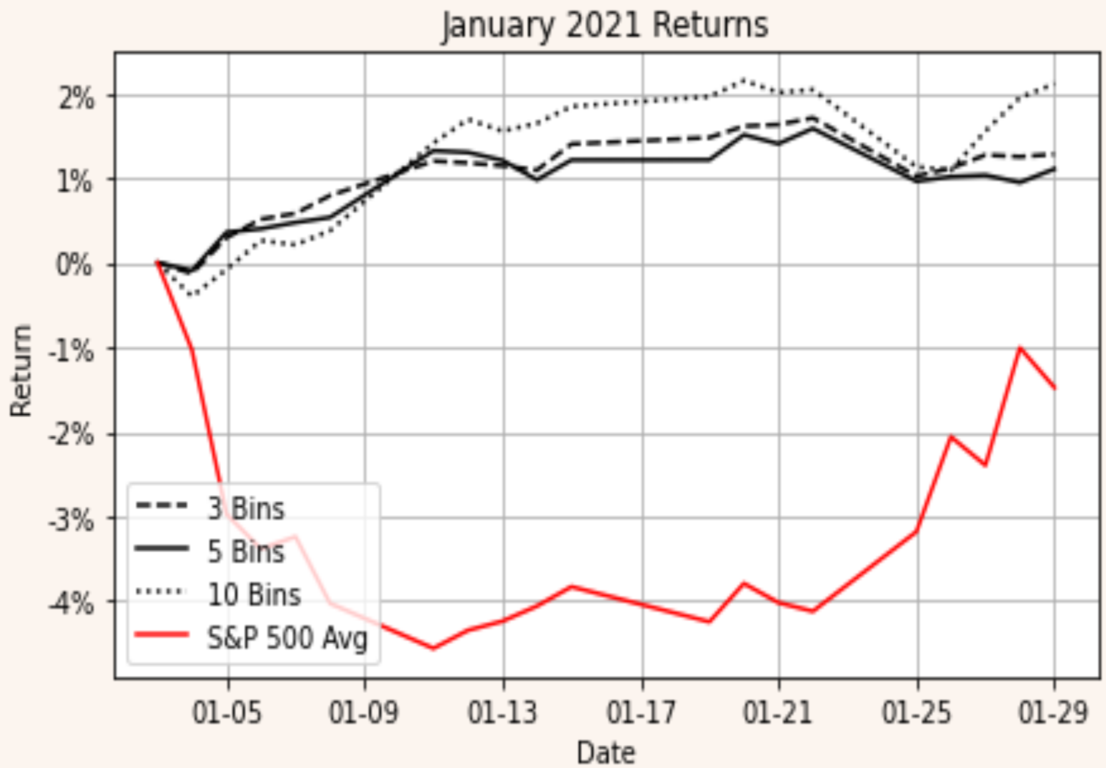
Portfolio Construction

Long-Short Portfolio

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

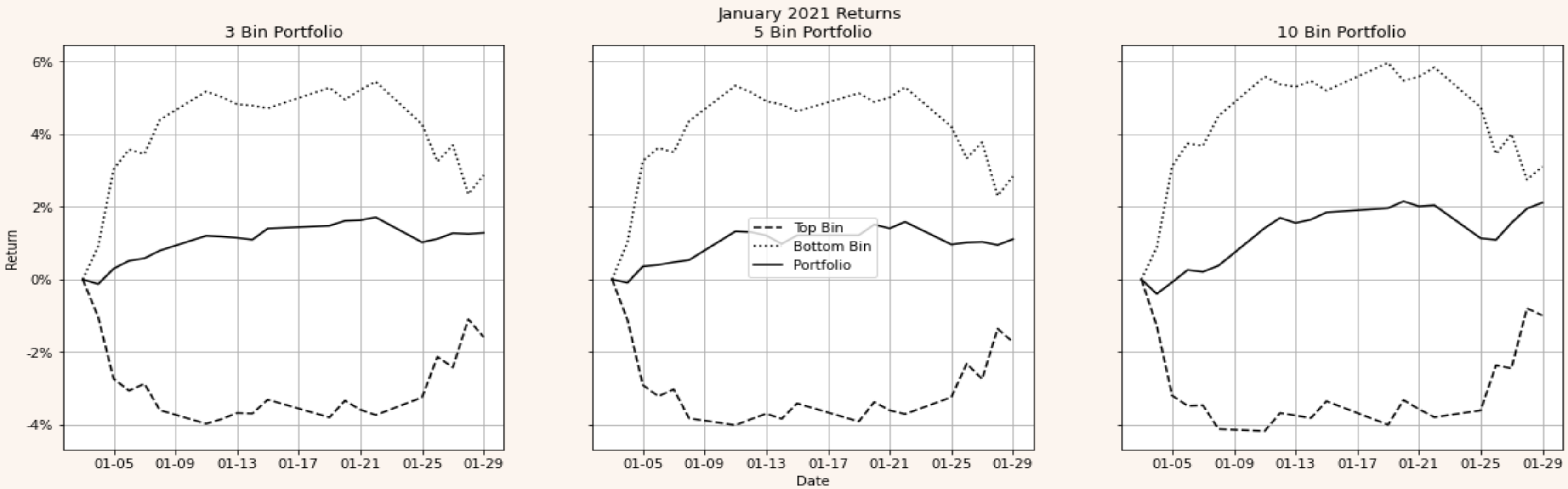


Results



Portfolio	Raw Return	Relative Return
S&P 500 Avg	-1.48%	----------
3 Bins	+1.28%	+2.76%
5 Bins	+1.11%	+2.59%
10 Bins	+2.12%	+3.60%

Results



3 Bins	Raw Return
Top	-1.59%
Bottom	+2.87%
Portfolio	+1.28%

5 Bins	Raw Return
Top	-1.73%
Bottom	+2.84%
Portfolio	+1.11%

10 Bins	Raw Return
Top	-0.99%
Bottom	+3.11%
Portfolio	+2.12%

Thank you!

Check back for final results!