Executive Summary

The sentiment of news drives a population's actions and reactions. This is no different when considering how financial news affects investors' behaviors. This study focuses on computing news sentiment scores, readability index scores, and daily news volume to construct stock portfolios with the end goal of outperforming the average of the S&P 500 stock index. The project was broken down into four main phases: data collection, data processing, portfolio construction, and results.

The data collection process began by gathering news articles from a worldwide database called GDELT 2.0 GKG Database. The columns were reduced to preserve only the necessary data including the datetime stamp, website, and URL. Additional filtering was performed to retain three websites. The criteria for selecting these websites included a high volume of financial news, no paywall, and a consistent webpage format that allows for repeated web scraping. Yahoo Finance was used to collect the daily open and close prices for each individual stock in the S&P 500 for all of 2021.

Three web scraping templates were developed to handle URLs from each of the three filtered websites. The web scraped news contents were then processed through an NLP API transformer model from HuggingFace. The *distilroberta-finetuned-financial-news-sentiment-analysis model* is a financially tuned model that produces three sentiment scores: positive, neutral, and negative. The article text was then also passed through readability and string-matching packages. The *readability* package was used to gage the reading level of the article contents and the *fuzzywuzzy* package was used to determine which stocks were mentioned in each article. A volume count was then added as an importance factor to examine the quantity of news published per day about each stock. Final data cleaning was then performed to address weekend/holiday news and if a stock was not mentioned in any published news. Holiday and weekend news was reallocated to the most previous market day in order to select stocks for the next market day. If a stock had no news published about it on a certain day, then a decayed score was generated from prior days. The data processing phase resulted in 27 final features.

Portfolios were constructed using the features generated in the processing phase. Long-short portfolios were the type of portfolios constructed where the top ranked stocks were bought in a long position, and the lowest ranked stocks were bet against in a short position. The stocks were ranked by scaling and aggregating the sentiment scores, readability scores, and daily volume count to generate one overall score for each stock, per day. Three different variations of long-short portfolios were considered by separating stocks into different bin sizes of 3, 5, and 10. The top bin took a long position, the bottom bin took a short position, and the middle bin(s) were left alone.

All three portfolios produced positive returns of at least 1% and outperformed the average of the S&P 500 by over 2.5% during January of 2021. The 10-bin portfolio performed the best producing a monthly profit of 2.12%. Further investigation into the portfolio compositions shows that all the profits were driven by the short positions in the bottom bin producing returns around 3%. The top bin in long positions all had negative returns. This indicates that negative news sentiment affects stock price movement more than positive news. Riskier investors can take advantage of this by only taking short positions in the low ranked stocks with negative sentiment.

This project can be taken in many different directions. Developing the study to include additional years of data along adding supplementary websites would provide a more diverse and reliable track record of returns. Expanding the investment timeline from daily analysis to weekly or monthly investments would benefit long-term investors. The final product of this analysis could be distributed to everyday retail investors on the quest to outperform the market.

Table of Contents

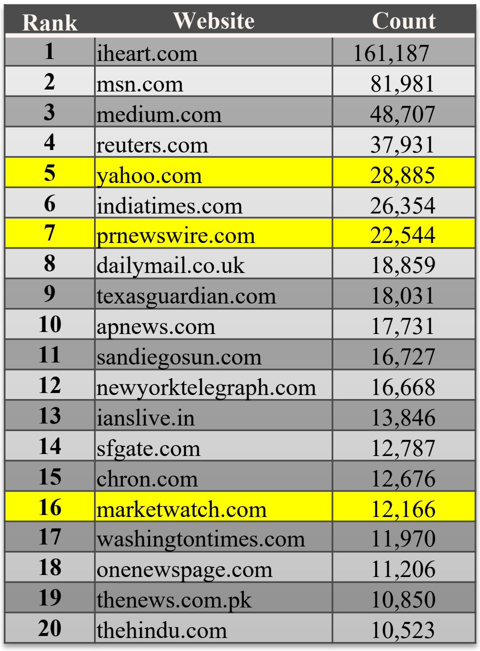
Introduction

This project is designed to explore the relevance of news articles and how it reflects in the movement of stock prices. The basis of ‘good’ news versus ‘bad’ news will be evaluated using sentiment analysis, a natural language processing (NLP) technique. Sentiment analysis produces scores to rate text as positive, negative, or neutral. Other characteristics of the text such as complex words and reading level will also be evaluated to determine the complexity of the news. The last aspect of news that will be considered is the volume of news published daily regarding each individual stock. News must be collected, audited, and then evaluated in comparison to daily stock returns to evaluate if sentiment affects prices. The results are then used to construct portfolios of stocks to perform again the average of the S&P 500 index.

Data Collection

The collection phase required three different types of data collected from three different sources. The main source was the GDELT 2.0 GKG Database that was used to assemble news articles. GDELT is a worldwide news database that releases event files in 15-minute increments every day of the week. The files originally contained 26 columns (~15GB) but were condensed to three columns (~15KB) to reduce storage issues. The three columns include the datetime stamp, website, and URL. Additional filtering was then performed on the website column to only analyze news from three websites: yahoo.com, marketwatch.com, and prnewswire.com. Below are the counts of each website for January 2021. The criteria for selecting these websites included having a high volume of financial news, no paywall to allow unlimited access, and a consistent webpage format to allow for repeated web scraping. A Python script was developed and executed on ACCRE to navigate and download the event data by adjusting the datetime stamp to the GDELT file links to collect all of the 15-minute files. The collected data spans the entire year of 2021 from January 1st to December 31st.

The second data source was Wikipedia that contained the S&P 500 index stock composition for 2021. Finally, the stock tickers gathered from Wikipedia were then passed into a web driver that navigated to Yahoo Finance to download the daily stock open and close prices for all of 2021. The last two data sources utilized web scraping as its collection method.



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

Four methodologies were used for data processing throughout the duration of the project. The first process composed the text contents of the news articles listed in the GDELT Database. The second phase passed the article text through two NLP processes to generate sentiment and readability scores. The third technique accumulated a daily volume count for articles published about each stock. The final step involved data cleaning to accommodate news published on the weekend/holidays along with handling days where a stock has no associated news published.

The first step in collecting the contents of the news articles was to create three web scraping templates for each of the three websites specified above. The three templates were developed using the *BeautifulSoup* package. A common issue when accessing news articles from a year prior was that some URLs were dead and no longer active. Specific error handling techniques were deployed to extract only the active URLs. If a URL was no longer active or inaccessible, this row of data was removed from the dataset.

After the news collection process was complete, the text needed to be passed through two NLP models/packages to be analyzed. The first model was an NLP application programming interface (API) deployed from *HuggingFace* used to generate sentiment scores for each article. *HuggingFace* is an open-source, artificial intelligence community specializing in NLP technologies. The model of choice from *HuggingFace* was the *mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis model* which is a "fine-tuned version of the *distilroberta-base* on the financial_phrasebank dataset.[1]" The financial specific language training was critical in the selection of this model. The *distileroberta-base model* consists of 6 layers, 768 dimension and 12 heads, totalizing 82M parameters and runs twice as fast as its predecessor the *RoBERTa-base model* which has a total of 125M parameters.[2] The *RoBERTa-base model* is a transformers model, that is self-supervised on raw English text which eliminated the process of human labeling. The model objective was to use Masked language modeling (MLM), which involves masking 15% of the words in a sentence and using the context of the sentence to predict those masked words.[3]

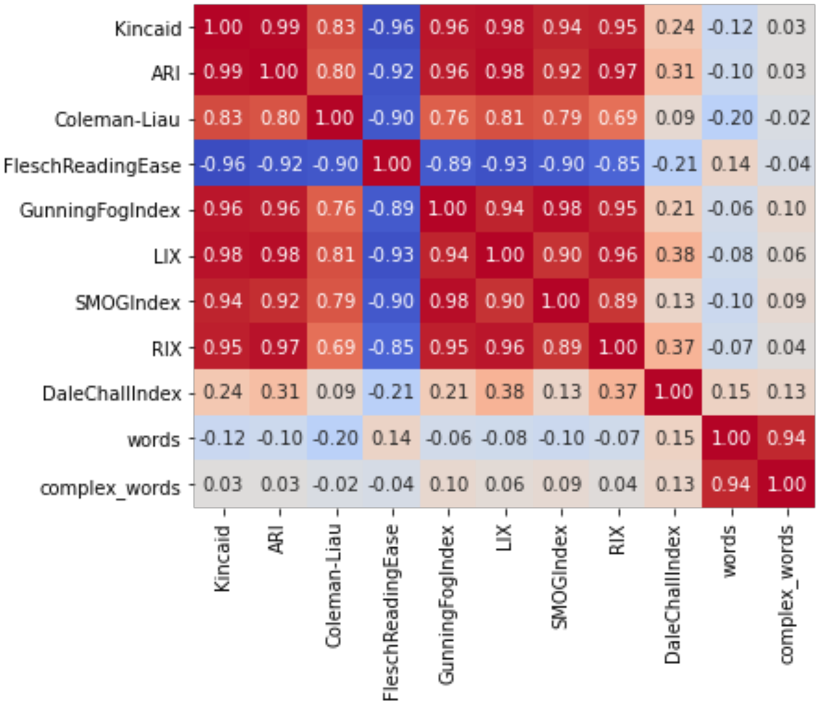
The *mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis model* works best when text is passed in on a sentence-by-sentence basis. This was achieved by scraping the news contents and splitting the texts at periods with a space afterwards. The output of the model produces three sentiment scores: positive, neutral, and negative. The combination of these three scores sums to one. A net sentiment scores was manually calculated by subtracting the negative sentiment score from the positive sentiment score. After computing the sentiment scores for each sentence, the scores were aggregated together to find summary scores, such as mean, median, minimum, and maximum, for each sentiment type of positive, neutral, negative, and net. In total, there are 16 summary sentiment scores for each article.

[1] *MRM8488/Distilroberta-finetuned-financial-news-sentiment-analysis · hugging face*. mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis · Hugging Face. (n.d.). Retrieved May 7, 2022, from https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

[2] *Distilroberta-base · hugging face*. distilroberta-base · Hugging Face. (n.d.). Retrieved May 7, 2022, from https://huggingface.co/distilroberta-base

[3] *Roberta-base · hugging face*. roberta-base · Hugging Face. (n.d.). Retrieved May 7, 2022, from https://huggingface.co/roberta-base

The second NLP process utilized the *readability* package to produce readability index scores. The package consists of traditional readability measures focused on basic characteristics of raw text. The results are essentially linear regressions based on basic characteristics such as words, syllables, and sentences.[4] The types of output that were collected from this package were the *readability grades* and the *sentence info* portion of the results. The entirety of the *readability grades* was selected for analysis including: Kincaid, ARI, Coleman-Liau, Flesch Reading Ease, Gunning Fog Index, LIX, SMOG Index, RIX, and Dale Chall Index. Only two measurements, word count and complex word count, were selected from the *sentence info* portion of the results. A correlation matrix was produced to examine the relationship among scores to potentially reduce the number of features. The matrix showed the readability scores were highly, positively correlated, with exception of Flesch Reading Ease which was highly, negatively correlated because its readability scale works in the opposite direction of the other scores. Dale Chall Index and both word counts had limited correlation with the other readability scores. These four results were removed because the other readability scores were assumed to provide adequate insight of the reading level. Since the readability scores were highly correlated, each score was scaled and added together to find an average readability score to reduce the feature count.



	Kincaid	ARI	Coleman-Liau	FleschReadingEase	GunningFogIndex	LIX	SMOGIndex	RIX	DaleChallIndex	words	complex_words
Kincaid	1.00	0.99	0.83	-0.96	0.96	0.98	0.94	0.95	0.24	-0.12	0.03
ARI	0.99	1.00	0.80	-0.92	0.96	0.98	0.92	0.97	0.31	-0.10	0.03
Coleman-Liau	0.83	0.80	1.00	-0.90	0.76	0.81	0.79	0.69	0.09	-0.20	-0.02
FleschReadingEase	-0.96	-0.92	-0.90	1.00	-0.89	-0.93	-0.90	-0.85	-0.21	0.14	-0.04
GunningFogIndex	0.96	0.96	0.76	-0.89	1.00	0.94	0.98	0.95	0.21	-0.06	0.10
LIX	0.98	0.98	0.81	-0.93	0.94	1.00	0.90	0.96	0.38	-0.08	0.06
SMOGIndex	0.94	0.92	0.79	-0.90	0.98	0.90	1.00	0.89	0.13	-0.10	0.09
RIX	0.95	0.97	0.69	-0.85	0.95	0.96	0.89	1.00	0.37	-0.07	0.04
DaleChallIndex	0.24	0.31	0.09	-0.21	0.21	0.38	0.13	0.37	1.00	0.15	0.13
words	-0.12	-0.10	-0.20	0.14	-0.06	-0.08	-0.10	-0.07	0.15	1.00	0.94
complex_words	0.03	0.03	-0.02	-0.04	0.10	0.06	0.09	0.04	0.13	0.94	1.00

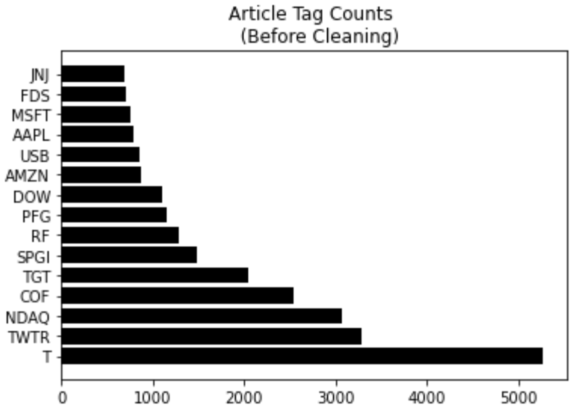
After collecting both the sentiment and readability scores, each article needed to be scanned to detect which of the S&P 500 stocks were mentioned in the contents of the article. The *fuzzywuzzy* package was imported to tag each article. *Fuzzywuzzy* is a string-matching package that applies Levenshtein Distance to differentiate distances between different character sequences.[5] The *fuzzywuzzy* extraction of string matches was applied to compare article contents and company names. The matching function used was the *token set ratio.* The *token set ratio* tokenizes the strings while converting all letters to lower case

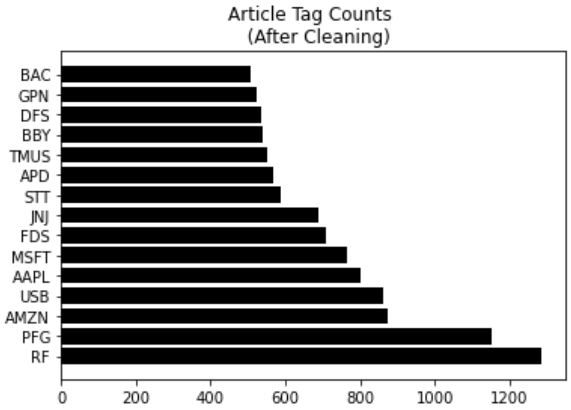
[4] *Readability*. PyPI. (n.d.). Retrieved May 7, 2022, from https://pypi.org/project/readability/

[5] *Fuzzywuzzy*. PyPI. (n.d.). Retrieved May 7, 2022, from https://pypi.org/project/fuzzywuzzy/

and removing punctuation. Common tokens are then removed, and a ratio is calculated producing a rating from 0 to 100. A filter was applied to only display tags with a rating of 90 or above. The *token set ratio* was selected because it is most useful when comparing sets of strings that greatly differ in length, which was present when comparing one or two word company names to paragraphs of news text.[6] This process produced a list of stock tickers for each article except for articles that did not result in any stock tags. The rows of data with no tags were dropped from the dataset. Each stock tag, for each URL article needed to be separated to its own individual row for future portfolio construction. The final output was a row of data for each stock mentioned within each article along with the associated sentiment and readability scores for the article.

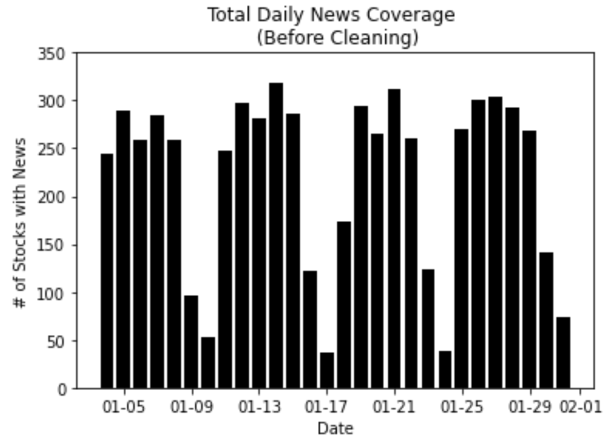
After assigning scores and tags to each article, a daily volume count of articles published about each stock was calculated. The volume count is a representation of the importance and popularity of news regarding that stock on a specific day. A stock's daily news features should be evaluated differently when it has a volume count of 15 compared to a volume count of three. Below is the breakdown of the top 15 most popular stocks that were tagged among articles in January 2021. Some tickers needed to be removed from the dataset because of inaccurate tagging. Companies such as Twitter and Target had high counts due to authors of articles citing their Twitter handle and mentioning "target" in relation to a benchmark goal for a stock.



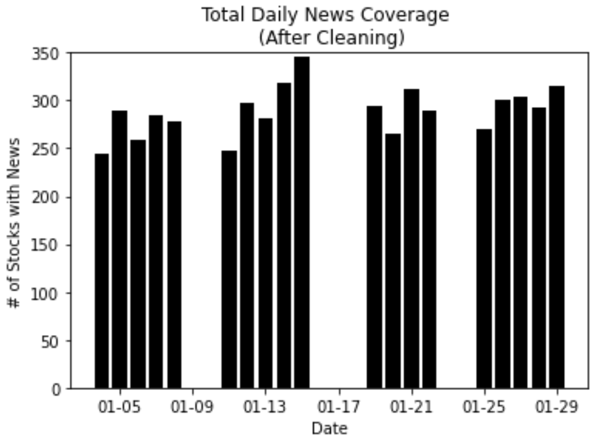


The last step of the data processing phase was finalized data cleaning. There were two parts of data cleaning: first for aligning news and market data and second for missing news. News articles are published every day of the week, but financial markets are only open on weekdays. This causes discrepancies when using weekend/holiday news to select the next day's stock movements. Weekend news was reassigned to the previous Friday to select stocks for the market opening on Monday. The same was applied to holidays where news published on the holiday was assigned to the previous market day news to select stocks for the next market day. The graphs below display the before and after allocations of news articles for January 2021. The "After Cleaning" graphs represent the reallocated weekend and holiday news, therefore have days with missing news. The first graph shows the coverage of news regarding the number of stocks that had at least one news article published. The second graph presents the total news volume that was analyzed on a daily basis.

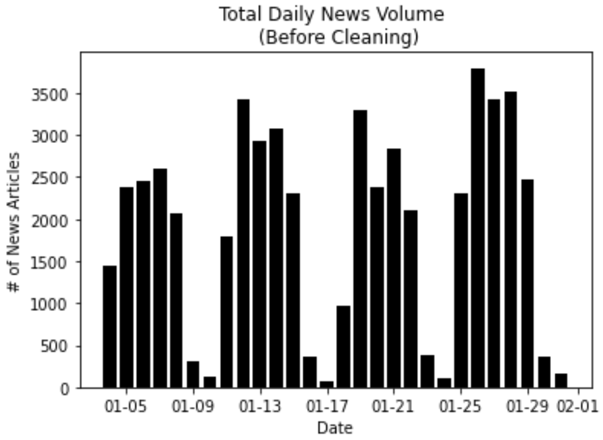
[6] Wong, J. (2020, November 8). *String matching with FuzzyWuzzy*. Medium. Retrieved May 7, 2022, from https://towardsdatascience.com/string-matching-with-fuzzywuzzy-e982c61f8a84



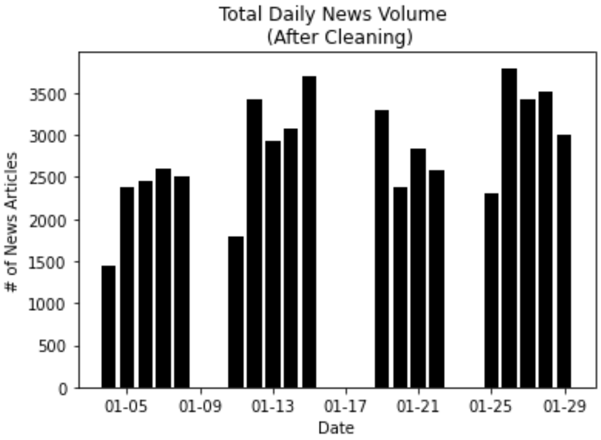
Total Daily News Coverage (Before Cleaning)



Total Daily News Coverage (After Cleaning)



Total Daily News Volume (Before Cleaning)



Total Daily News Volume (After Cleaning)

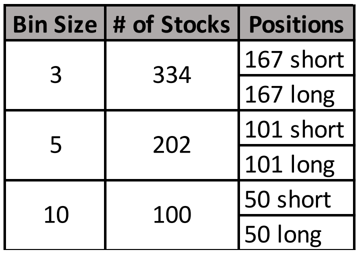
Finally, the data was grouped by date and stock to create one row of summary scores per stock per day. Another issue occurred where certain stocks had no news published about them on a specific day. A decayed version of previous days' news was applied to impute the scores for the day with no news. This was done by allocating 70%, 20%, 10% weights to the lagging three days with the most weight going to yesterday and least weight going to the most distant day. This type of decaying method lessens the effect of news as times moves on.

The entire data processing phase was run on Amazon Web Services Elastic Compute Cloud (AWS EC2). This produced 27 final variables listed below:

- Date
- Tags
- Ticker
- Volume
- Positive mean
- Positive median
- Positive minimum
- Positive maximum
- Negative mean
- Negative median
- Negative minimum
- Negative maximum
- Neutral mean
- Neutral median
- Neutral minimum
- Neutral maximum
- Net mean
- Net median
- Net minimum
- Net maximum
- Read score
- Same day stock raw return
- Next day stock raw return
- Same day S&P 500 raw return
- Next day S&P 500 raw return
- Same day stock relative return
- Next day stock relative return

Portfolio Construction

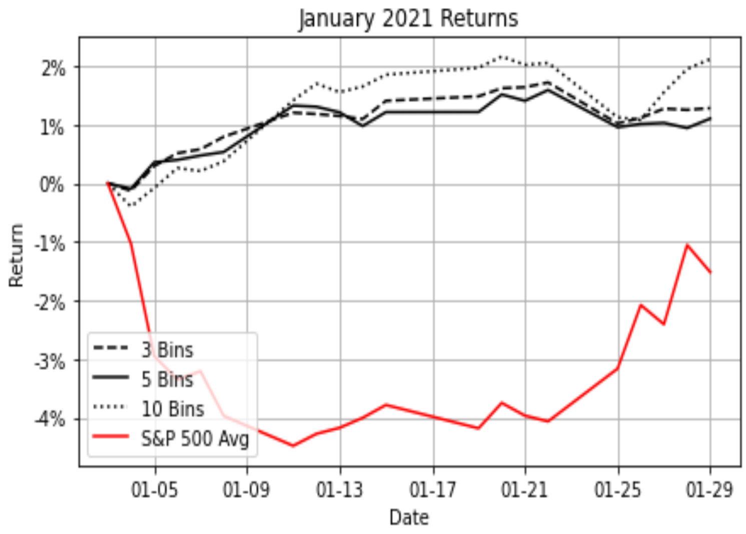
The final part of the project involves portfolio construction. A portfolio is a collection of stocks that a person wants to invest in. The type of portfolio that was constructed was a long-short portfolio. This means that half of the portfolio will take a long position, while the other half will take a short position. A long position implies that an investor is optimistic and buys a stock they think will go up in value. A short position is the opposite where an investor is pessimistic about a stock and bets against in hopes that it will decrease in price. The list of S&P 500 index stocks must be ranked and then split into sections called "bins" to establish the long and short positions. The list of stocks was ranked by scaling and aggregating sentiment scores, readability score, and volume count. This created an aggregated final score that allowed the stocks to be ranked from highest to lowest. The number of bins was the tuning parameter used to vary approaches of splitting the stocks. Three different numbers of bins (3, 5, and 10) were applied when separating the stocks. After splitting the stocks into bins, the top bin took a long position, the bottom bin took a short position, and the middle bin(s) took no position. Below details the number of stocks selected and the number of stocks per position for each bin size out of the total 504 stocks in the S&P 500 index.

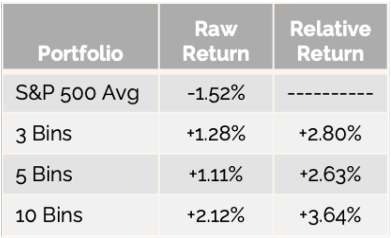


Bin Size	# of Stocks	Positions
3	334	167 short
		167 long
5	202	101 short
		101 long
10	100	50 short
		50 long

Results

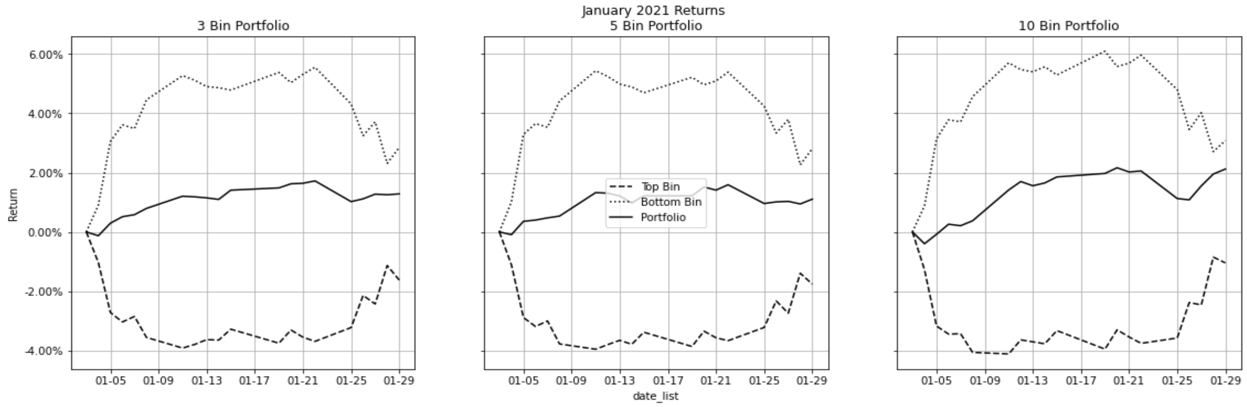
The performance baseline used in comparison with the three different long-short portfolios was the equally weighted average return of the 504 stocks in the S&P 500 index. Notice this is a different return than the S&P 500 index as that return is weighted differently based on the market cap size of companies. The graph and chart below indicate that all three portfolios outperformed the average of the S&P 500 index by at least 2.5% with the 10-bin portfolio performing the best with a return of 2.12%. The three portfolios also all produced positive returns where the benchmark had a negative return of -1.52%.

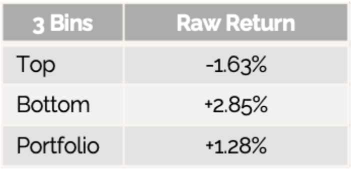




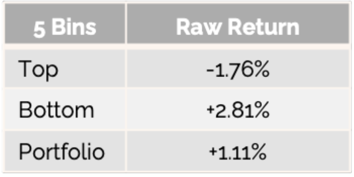
Portfolio	Raw Return	Relative Return
S&P 500 Avg	-1.52%	----------
3 Bins	+1.28%	+2.80%
5 Bins	+1.11%	+2.63%
10 Bins	+2.12%	+3.64%

The next step was to analyze the composition and performance of each portfolio. Further investigation showed that all the profits were driven by the short positions in each portfolio with returns of at least 2.8%. All the long positions produced negative returns. For a riskier investor, creating a portfolio with a heavier concentration on the short positions could produce higher returns. One concern regarding the daily analysis was the transaction costs associated with the number of stocks entering and exiting bins. The average daily turnover for the top bin was 43.5% while the bottom bin was 30.5% which implies that this strategy would still be profitable when considering transaction costs.

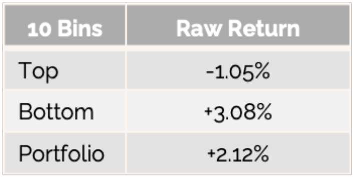




3 Bins	Raw Return
Top	-1.63%
Bottom	+2.85%
Portfolio	+1.28%



5 Bins	Raw Return
Top	-1.76%
Bottom	+2.81%
Portfolio	+1.11%



10 Bins	Raw Return
Top	-1.05%
Bottom	+3.08%
Portfolio	+2.12%

Conclusion

Although this analysis produced encouraging results, there is room for additional improvement. First, gathering additional years of data along with selecting more news sources when filtering on websites would increase both the quantity and diversity of data to provide more trustworthy results. Another improvement would be to apply GPU computing to the scripts ran on AWS EC2. Running the scripts on GPUs or with parallel processing would cut run time and allow for quicker decision making.

The analysis of the feature and prediction horizon could be adjusted to examine which timeframe produces the best results. The study used today's news to select tomorrow's stocks. Instead, investigating weekly, monthly, or quarterly timelines could produce different results with more significance. This could be more beneficial for long-term investors rather than short-term investors.

After these options are explored and evaluated, a final product can be produced. This product would resemble a daily newsletter that includes the ranking of stocks, the various bin portfolio compositions along with the historical performances of both the bin portfolios and the long and short positions within the portfolios. This product could then be marketed and sold to everyday retail investors with the goal of outperforming the average S&P 500 index return.