Sentiment Analysis and Text Classification to Predict S&P 500 Behavior

Aditya Bindra
University of Virginia
Charlottesville, Virginia
ab4es@virginia.edu

Paul Cherian
University of Virginia
Charlottesville, Virginia
pmc6dn@virginia.edu

Ben Haines
University of Virginia
Charlottesville, Virginia
bmh5wx@virginia.edu

ABSTRACT

This paper describes our system for using sentiment analysis and text classification on the minutes of the Federal Open Market Committee (FOMC) to predict the behavior of the S&P 500. This paper summarizes existing work in the field and describes our novel implementation. We evaluate our predictions using Modern Portfolio Theory, the financial industry standard, and propose a few potential areas for future development.

1. INTRODUCTION

The challenge our project addresses is how to use publicly available text-based information sources produced by the United States government to predict the behaviors of financial markets. In particular, we attempt to apply sentiment analysis and text classification techniques to the minutes of the FOMC in order to identify correlations with the behavior of individual sectors of the S&P 500.

There are a number of incentives that attract research attention to the area of predicting publicly traded markets. Conspicuously, many of these incentives involve the prospect of economic gain. In theory, the ability to predict the behavior of a particular component of the market would allow the individual possessing such knowledge to make significant profits. It is already the case that algorithms attempting to determine stock prices play a large role in the market. The introduction of high-powered computing to the realm of stock market trading created an entire industry based on algorithmic trading. In 2012 roughly 50% of US stock trades were made by high frequency traders using computers to automatically detect trends. The competition for any possible advantage is so large that significant capital is spent placing computer systems close to the stock market in order to shave milliseconds off of the network latency. Some of these systems already use text mining techniques to analyze news reports and make decisions faster than humans are capable of doing.

Beyond being able to make incredibly fast micro trades, there is also appeal in being able to reliably predict the direction a market is headed over a longer period of time. For the average trader who doesn't have access to ultra-fast computers and proprietary algorithms, having a reliable indicator of larger trends may be even more desirable. Finally, there is the academic incentive to potentially gain a better understanding of the forces that drive the market.

After settling on the general idea of sentiment analysis applied to financial markets the decision still remained to be made about exactly how we would approach the problem. In particular we needed to decide what our data source would be and what we would be attempting to predict. The decision to track the sectors of the S&P 500 was relatively intuitive. This market provides an overall measure of market performance broken down into specific categories. If we had instead attempted to track performance of individual stocks, obtaining data would have been more complicated. It is unlikely that any given sentence in a corpus will be relevant to a specific company, but the chance of a sentence being relevant to the housing sector as a whole is much larger.

The decision of what corpus to mine for sentiments provided more viable options. Much of the existing work in this area uses social media or news reports. The primary advantage to these sources is that they are easily accessible and in very large quantities. In order to differentiate ourselves from this work we instead chose to use official releases from the Federal Open Market Committee. There are advantages and disadvantages that came with this decision. The FOMC is responsible for deciding the monetary policy so their minutes are likely more relevant than tweets. On the other hand, there is a natural limit to the amount of data we can collect, the reports are only eight times per year and have only been made available online for the past five years. Another advantage to using a source such as twitter is that an argument can be made that analysis is revealing hidden, but widespread emotions that may actually be driving the market. In contrast, the majority of the sentiment likely to be discovered in an official government source is probably explicit. Much of the benefit then is simply the ability to process the information quickly. This is of limited benefit because the documents are short enough to be read by humans in a reasonable amount of time but in theory the system could be expanded to take input from a wider variety of government sources.

2. RELATED WORK

As mentioned previously, there is a significant amount of work within the area of applying sentiment analysis to fi-

PLDI '13 June 16–19, 2013, Seattle, WA, USA

DOI: 10.475/123_4

nancial markets. Due to the financial incentives much of this work is actually done by private companies. For example, Bloomberg Labs both maintains a very large news corpus and operates data science labs that work on problems of text mining and sentiment analysis. There are also a number of academic works on the subject. We will discuss two works that give an overview of the state of the art for this type of problem. A publication that gives a summary of standard techniques is the 2009 paper *Sentiment Analysis in Financial News*[1] by Azar which discusses both the problems of training a sentiment classifier as well as attempting to use it to predict future returns. The paper is quite comprehensive but differs in many ways to our work. The paper focuses on predicting returns for specific stocks based on news releases about the relevant companies. The author then analyzes four different classifiers based on decision trees, SVM, Naive Bayes, and word counting in their ability to detect sentiment. The classifiers are trained directly by the market, using stock returns for a given day as labels for the news articles released the day previous. When used to create portfolio strategies each classifier simply recommends to buy when news is good and sell when it is bad. The approach here is fundamentally different from what we propose. Azar combines all of the news related to a company on a particular day into a single document and then classifies this as positive or negative. In contrast we begin with a single document from which we extract relevant portions for each of our areas of interest. We then apply a sentence level sentiment classifier and use an algorithm to aggregate these individual sentiments into a single measure. Our classifier is based on neural networks rather than any of the four learning techniques used by Azar. Finally, because we are not making day by day decisions to buy or sell particular stocks our technique for deriving a strategy is more involved. One conclusion made by Azar that is relevant to our work is that it is possible to develop a strategy that yields significant returns. However, this was only the case when the actor was allowed to trade on information before it was made public, returns were outweighed by trading costs when trading was limited to after the news had been made public.

The second work which developed a very different strategy is the 2010 paper *Twitter Mood Predicts the Stock Market* by Bollen et al[4]. Rather than use a two dimensional classification of sentiment from negative to positive the authors break down sentiment into six components such as "Happy", "Calm", or "Alert." They then create time series data by tracking these components as they occur in twitter data. They used neural networks to incorporate different combinations of these components into a standard predictive model based on past stock prices. This paper in particular sparked interest in sentiment analysis because it claimed to be able to predict whether the Dow Jones Industrial Average would close up or down on a given day from two to six days ahead of time with 86.7% accuracy, although some have disputed these claims.

In summary the primary differences between our proposed solution and the majority of existing work are as follows. First, we focus on tracking broad economic sectors rather than individual stocks. However, we also predict at a finer level that simply tracking the index as a whole. Second, rather than making daily decisions on whether to buy or sell, we recommend longer term strategies. Finally, we investigate a new source of information as opposed to existing work which looks at social media and new articles. The primary contribution of our work is the combination of sentiment analysis and text classification in order to extract information about multiple sectors from a single document.

3. SOLUTION

Our solution has three components. First we need a method for determining sentiment measures for sentences. Next we need a way to identify which sentences within a given document are relevant to which sectors of the S&P 500. Finally, we need an algorithm to combine these two measures in order to create a market strategy. We discuss each of these components below.

3.1 Sentiment Analysis

We decided to use the Stanford Sentiment Analysis[8] system for identifying sentiment within sentences. We did consider developing our own system but it seemed unlikely that the resulting performance would justify the effort expended. Stanford's analyzer has the benefits of being an established and trusted system, having an accessible Java API, and providing sentence level sentiment classifications. These were all features that appealed to us. One downside to this system is that classifications are limited to five classes. It would have been useful to be able to access some continuous value representing confidence of each class or intensity of sentiment on a negative to positive scale.

3.1.1 Examples

We begin by showing some examples of classifications. Successful Classifications:

1. Positive: Payroll employment continued to move up, and the unemployment rate, while still elevated, declined a little further.
2. Very Positive: In recent months, the production of motor vehicles continued to rise appreciably in response to both higher vehicle sales and dealers' additions to relatively low levels of inventories; output gains in other industries also were solid and widespread.
3. Positive: Labor market conditions improved in recent months.
4. Negative: Moreover, sales of new and existing homes edged down, on net, in recent months.
5. Neutral: However, the staff's medium-term projection for real GDP growth in the April forecast was little changed from the one presented in March.

Unsuccessful Classifications:

1. Positive: With respect to credit to households, developments over the intermeeting period were mixed.
2. Negative: By unanimous vote, the Committee ratified the Desk's domestic transactions over the intermeeting period.
3. Negative: The unemployment rate declined to 8.2 percent in March.

4. Negative: Some indicators of job openings and firms' hiring plans improved.

5. Neutral: Overall, the level of activity in the sector remained depressed.

From the results above we can see that there were correct and incorrect classifications for each sentiment class. There were some limitations to performing sentiment analysis in the way we did. Clearly some results are classified in ways that are counterintuitive. We suspect there are multiple factors that contributed to the poor performance in these cases. We are applying a general model to a very specific domain for which it may not have adequate training. As indicated in the paper by Azar discussed earlier, models trained on a specific domain, for example movie reviews, are not likely to generalize well to financial text. In fact, the models Azar trained on a specific domain performed as well as humans in that area but did no better than random guessing in a different area. We attempted to counteract this by training the model ourselves on a domain specific dataset but we were not successful. The only sentiment labeled data we were able to find was the Loughran-McDonald dictionary. Unfortunately this is a only a dictionary and does not contain labeled sentences. The resulting model performed much worse that the default one.

The second problem is that sentiment within this domain can be very ambiguous and depend heavily on context. Any particular piece of news, for example the price of oil rising, could be considered a positive or negative thing from the perspective of producers or consumers of oil respectively. Finally, we frequently encountered sentences such as "Bank credit quality was increasing, although commercial lending remained relatively weak" where a positive first half of the sentence is counteracted by a negative second half and it is unclear which is more important.

To some degree the fact the classification system gives sentiments on a scale rather than a binary measure allowed us to "weight" the confidence of a particular classification. However, there were plenty of cases of sentences being marked "highly negative" when a human observer would immediately have labeled it "highly positive". Additionally, although the scale from very negative to very positive is useful, it is not as useful as a pure numerical indicator of sentiment would have been for our purposes.

The graph below shows the results of testing the sentiment classifier on a set of hand labeled sentences.

Table 1: Binary Classification

		Predicted Sentiment		
		Positive	Negative	Total
True Sentiment	Positive	9	18	27
	Negative	0	33	33
	Total	9	51	60

For this two class classification problem the classifier achieved an overall accuracy of 70%. It was able to correctly identify 100% of the negative documents but only identified one third of the positive documents correctly. When a third neutral class was added the accuracy rate dropped to approximately 47% with the majority of the new errors resulting from misclassifying neutral sentences as negative. Although this performance may not seem great, the type of error that is made is not that harmful. We will ultimately avoid trading on sectors which have determined to be negative which means that a tendency to misclassify things as negative translates into a very low risk strategy. It would be much worse to mistakenly identify negative documents as positive.

Overall, this particular step of the pipeline represents an area of potentially significant future improvement. In particular, taking inspiration from the papers mentioned in the relevant work section, training the classifier using actual economic data as labels is likely to improve performance.

3.2 Classification

Our aim was to determine the most likely sectors that each assessment in the FOMC report would have impact on. To simplify the method, we considered each sentence in the FOMC Report to be indicative of an assessment of the financial condition by the members of the committee. To identify the sectors we trained a text classifier on corpus consisting of financial news and the ran the classifier on the federal report minutes.

3.2.1 Data

We scraped around 180 articles for each of the 11 sectors from the Bloomberg site to be used as a training corpus for text classification. Bloomberg is one of the most reputed agencies for financial news and offered an easy search functionality to search news by sector. The articles also had a significant amount of sector specific financial terms, which would help in predicting topics for FOMC reports. We used the Beautiful Soup library in Python to scrape web pages related to each sector.

3.2.2 Methodology

Since we had to determine a sentence level classification for the federal reports, we decided to use each sentence of the Bloomberg articles as a separate document for the purpose of training. After dividing the training corpus by sentences, our corpus size was around 51k documents. Each sentence of the article was labeled with the sector to which the article belonged. We decided to use the softmax classifier/ max entropy classifier to classify the text, because it gives better accuracy in attributing probabilities to different classes. This was important to us, because we are weighing the sentiment values of each sentence with the probabilities of each class for that sentence.

The softmax function is given by

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

Intuitively, the primary topic of a sentence of the Fed Report can be attributed to a particular sector, if it discusses metrics or indicators that are relevant to that sector. Therefore, we thought it would make more sense to use just the noun phrases(eg: consumer price inflation, housing index etc..) in a sentence to classify the text. Just using the word tokens would remove the meanings associated with sector specific terms.

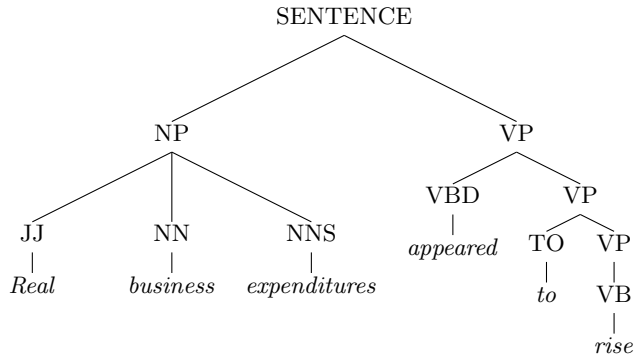
We used the Stanford NLP Column Data Classifier[6] which can use a variety of methods including max entropy, naive bayes and SVM to do supervised learning. The Stanford NLP classifier allows the users to specify particular feature set for the documents, instead of using the whole text as the feature set. Therefore we could use just the noun phrases as

Table 2: Test Scores

Method	Recall	Precision	F1-Score
Words	0.19	0.14	0.161
Words(Stemmed)	0.23	0.15	0.182
NP	0.30	0.24	0.266
NP(Stemmed)	0.35	0.32	0.334

the features for the training. For the training purpose the documents had to be formatted to a tab delimited format ,with the first column being the class and the rest ,features. We used a Probabilistic Context Free Grammar Parser in the Stanford NLP package to build a tree representing the lexical information for each sentence. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely grammatical structure of the sentence

For example, parse tree for *Real business expenditures on equipment and software appeared to rise.*



The parser we used was an already trained English parser model provided by Stanford NLP. Once the tree was created, we extracted the the named nouns in the noun phrases. The words we are interested in in the above sentence would be *real business expenditures*, *equipment* and *the first quarter.*

To test our intuition that noun phrases could be a better predictor of text classes, we ran 10 fold cross validation using the classifier on all words and just the noun phrases. We also ran the test for stemmed words to see if that improved the results. Text for the all the tests were subjected to a common pre-processing pipeline:

1. Conversion to lowercase letters
2. Removal of stop words and punctuations

The stemmed tokens scored consistently better than the raw words in classification tests. The tests also proved that noun phrases were better features for classification. The overall precision and recall metrics for our best classifier were low. But when metrics are compared across the different classes, we find that certain classes like Industrials, Health Care, Technology and Real Estate have F1 scores closer to 0.5 and sectors like Consumer Discretionary, Consumer Staples, and Utilities had F1 scores below 0.2. The FOMC reports usually talk about the the macro economic situation prevailing in the country. Therefore the topics they covers are usually sectors like industrials, real estate and technology, which incidentally are classes for which our classifier is much more accurate.

3.3 Combining Sentiment Analysis and Classification

As a reminder, our novel approach to predicting future performance of the S&P 500 is to combine our sentiment analysis model and trained sentence classifier. While many have attempted to predict the future returns of the S&P 500, we felt that by breaking down this task to a more granular level, we should be able to observe superior returns. That is, by examining the sentiment of each of the individual sectors of the S&P 500, we expect to see greater prediction power, which would ultimately be reflected in superior returns of that of this benchmark index.

In essence, we sought to find the sentiment of each sector for each FOMC Minutes so that we could predict future performance of these individual sectors. To do this we calculated the sentiment and classification of each sentence of each FOMC Minutes. Then, by multiplying a sentence's sentiment against its respective sector classifications, we had a sentiment value for each of the sectors for each of the sentences. These values were then averaged across each FOMC Minutes to then provide an average sentiment per sector. Then, each of the sector's excess sentiment over the average of that document was calculated. This allowed us to compare the extent to which a particular sector's sentiment was either more positive or more negative than its counterparts within this document.

More concretely, for every sentence S_x of a FOMC Minutes $D = \{S_1, \ldots, S_n\}$, we calculated its sentiment s where $s \in [-1, 1]$ and its classifications $c = \{c_1, \ldots, c_{11}\}$ where $c_i \in [0, 1]$ and $\sum_{i=1}^{11} c_i = 1$. Then, we stored the sentiment of these sector classification values for this sentence S as $v = \{sc_1, \ldots, sc_{11}\}$ or $v = s\sum_{i=1}^{11} c_i$. We then stored the average v across all sentences in our document as $v_{avg} = \frac{v}{n}$ where n is the number of sentences in our document D. Finally, we calculated the excess sentiment of each sector, v_e, against the average, v_{avg}, as $v_e = v_i - v_{avg}$. This process was carried out for all of the FOMC Minutes in our corpus.

Below are examples of such calculations for two sentences in the January 2011 FOMC Minutes.

1. S_x: There were no open market operations in foreign currencies for the System's account over the intermeeting period.

$$s = -0.5 \text{ [negative]} \tag{1}$$

$$c = \begin{cases} c_{\text{financials}} = 19.6\% \\ c_{\text{consumer\_staples}} = 3.5\% \\ c_{\text{materials}} = 8.0\% \\ c_{\text{health\_care}} = 6.8\% \\ c_{\text{consumer\_discretionary}} = 4.3\% \\ c_{\text{technology}} = 14.2\% \\ c_{\text{financial\_services}} = 11.2\% \\ c_{\text{utilities}} = 7.3\% \\ c_{\text{industrials}} = 11.6\% \\ c_{\text{energy}} = 4.4\% \\ c_{\text{real\_estate}} = 9.1\% \end{cases} \tag{2}$$

2. S_{x+1}: The rise in imports reflected an increase in the value of imported petroleum products, mostly explained by higher prices, and of capital goods, which was sup-

ported importantly by a jump in computers.

$$s = 0.5 \text{ [positive]} \tag{3}$$

$$c = \begin{cases} c_{\text{financials}} = 3.2\% \\ c_{\text{consumer\_staples}} = 0.7\% \\ c_{\text{materials}} = 13.8\% \\ c_{\text{health\_care}} = 2.2\% \\ c_{\text{consumer\_discretionary}} = 8.7\% \\ c_{\text{technology}} = 4.1\% \\ c_{\text{financial\_services}} = 6.5\% \\ c_{\text{utilities}} = 1.6\% \\ c_{\text{industrials}} = 7.9\% \\ c_{\text{energy}} = 4.7\% \\ c_{\text{real\_estate}} = 46.7\% \end{cases} \tag{4}$$

$$\vdots$$

$$\text{jan\_2011}_{v_e} = \begin{cases} c_{\text{financials}} = -0.0031 \\ c_{\text{consumer\_staples}} = 0.0112 \\ c_{\text{materials}} = -0.0042 \\ c_{\text{health\_care}} = -0.0040 \\ c_{\text{consumer\_discretionary}} = -0.0181 \\ c_{\text{technology}} = 0.0081 \\ c_{\text{financial\_services}} = 0.0043 \\ c_{\text{utilities}} = -0.0003 \\ c_{\text{industrials}} = 0.005 \\ c_{\text{energy}} = 0.0072 \\ c_{\text{real\_estate}} = -0.0067 \end{cases} \tag{5}$$

Ultimately, we will use the values in v_e for each FOMC Minutes to construct our trading strategy.

3.4 Quantitative Trading Strategy

In order to establish that the returns we are able to garner are more the result of our sentiment analysis model and sentence classifier rather than that of a creative portfolio selection and weighting schema, we employed a fairly simple smart-beta strategy.

3.4.1 Smart-beta Strategy

At its core, smart beta attempts to improve returns, reduce risk, and diversify an investor's portfolio by leveraging systemic investment factors[2]. In other words, smart-beta strategies attempt to construct novel portfolios that employ different portfolio selection and weighting techniques in order to generate greater returns than those of its benchmark index while hopefully taking on lesser risk and diversifying further. In our case, our strategy is beta-smart in that it examines our benchmark, the S&P 500, for each of its individual 11 sectors. Doing so allows us to construct our portfolio at a more granular level and focus on specific sectors which we expect to outperform the market as whole. This qualifies as our attempt at generating excess returns while reducing risk as we would be exposed to less of the market systemic risks and noise.

We should take some time to mention specifically why we have chosen to develop a smart-beta strategy rather then a long-only, active investing strategy. There has been a historic rise in smart-beta strategies over the last decade. Blackrock, a global investment management company with 4.6 trillion USD under management,[3] has attributed this to three core drivers[2]:

1. Smart-beta investing, a passive investing approach, requires small management fees while delivering greater transparency and consistency than their active investing counterparts.

2. Many who are attracted to the benefits of passive investing are not entirely comfortable parting with additional returns that is expected of active investing; smart-beta strategies serve as a middle ground between the two.

3. The advancement of technology has allowed for smart-beta strategies, which are heavily rules-based, to be easily automated.

Clearly the allure for such strategies is growing and we would like to address that need in our exploration; however, we are especially interested in the notion that smart-beta strategies are known to particularly transparent and based heavily on well-defined rules. While most active investors will invest in companies through an bottoms-up or top-down investment process that is extremely difficult to automate, smart-beta strategies are inherently easy to automate. Consequently, an automated system that trades off of a simple set of rules will allow for more emphasis to be placed on our sentiment analysis model and sentence classifier. We feel that it is the best strategy as it provides scientific rigor in proving our novel approach to investing against the S&P 500.

3.4.2 Portfolio Position Selection and Weighting

At this point, we are assuming we will have calculated v_e for a given FOMC Minutes. Then, if any value in v_e is positive, add its value to pos_T. Next, for all these positive values i in v_e, long the corresponding S&P 500 sector Electronically Trade Fund (ETF) until the next FOMC Minutes is released with the weight $w = \frac{v_e[1]}{pos_T}$. In essence, this strategy is investing in the sectors that have a greater sentiment than that of the average among all sectors for that FOMC Minutes and weighting it relative to the extent by which it is greater than that of the average. Again, the intent behind such a simple strategy is to ensure the emphasis of this exploration is on that of our sentiment analysis model and sentence classifier.

4. EVALUATION

4.1 Backtest Details

In order to evaluate the performance of our exploration, we performed a backtest from January 2011 to March 2016, the time frame for which the FOMC Minutes has published its minutes online in a .txt format. This backtest occurred over 39 FOMC Minutes and compared the returns of the S&P 500 against that of our strategy, which invested in the S&P 500 sectors through sector SPDRs (ETFs that track the S&P 500 sectors). It should be noted at this time that these SPDRs are not available for short-selling and our strategy was long-only as a result. Additionally, since the Real Estate and Financial Services SPDRs had only been available since May 2015, we ignored them for the purposes of this backtest. Lastly, this backtest did not include trading costs and all returns are gross returns.

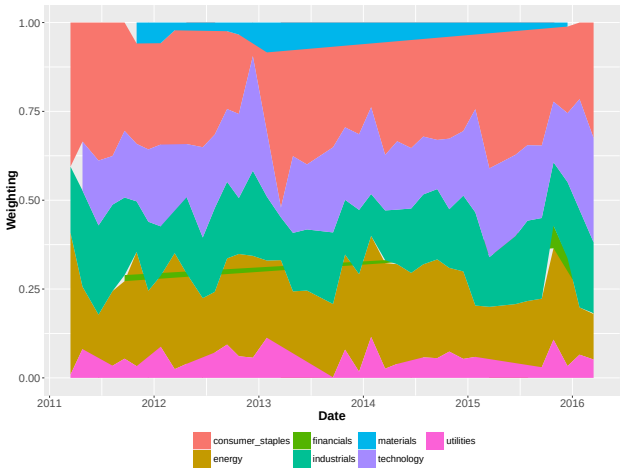


Figure 1: Portfolio Selection and Weightings over Backtest

4.2 Portfolio Selection and Weighting

Figure 1 shows our portfolio selection and weighting over the length of our 5-year backtest. We see that for the most part consumer staples, technology, industrials, and energy comprised most of each the portfolio for each FOMC Minutes. We attribute this to the FOMC Minutes focusing on sectors that affect the US economy as a whole. Consumer staples can be tied to consumer spending and therefore the FOMC focuses some time to discuss consumer spending in each report, for instance. Real Estate is another example of such sectors; however, since it had only gone live as a publicly traded SPDR in May 2015, it was not included in this backtest.

4.3 Modern Portfolio Theory

Modern Portfolio Theory (MPT) is a mathematical framework to analyze a position within a portfolio by the impact of its expected return and risk on the portfolio as a whole. This framework was introduced by Harry Markowitz in his 1952 essay *Portfolio Selection*[7]. MPT now serves as the industry standard when analyzing the results of any trading strategy and we will apply it to our strategy accordingly.

4.3.1 Normal Distribution of Returns

A normal distribution of returns is particularly important as it will allow for a more accurate prediction of what returns are to be expected and serves as the basis for MPT[7]. Examining at the first four moments of a return distribution – mean, variance, skewness, and kurtosis – together will help in establishing a normal return distribution.

Table 3: Moments of Return Distribution over Backtest

	Strategy	S&P 500
Geometric Mean	0.0123	0.0115
Variance	0.0686	0.0685
Skewness	-0.0739	-0.3436
Kurtosis	2.9127	3.5134

With kurtosis values fairly close to 3 and skewness values close to 0 for the strategy and the S&P 500, it can be argued that both have adequately normal distributions for the purposes of MPT.

4.3.2 Returns through Backtest

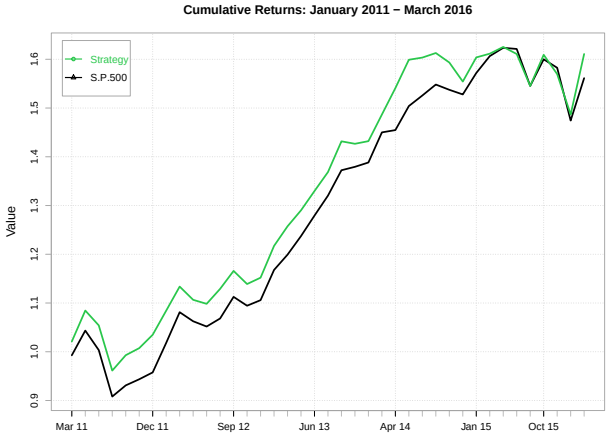


Figure 2: Cumulative Returns over Backtest

This strategy employs a smart beta strategy in that it relies on an alternate index construction – both in position selection and weighting – to generate additional alpha against the S&P 500. Whenever possible, performance should be analyzed relative to the S&P 500. As such, while Figure 2 shows that this strategy enjoys cumulative returns of 162.35% relative to the S&P 500s 157.82%, Figure 3 illuminates these returns more clearly. By plotting this strategy's returns relative to those of its benchmark, a relative return of 133.44% is apparent.

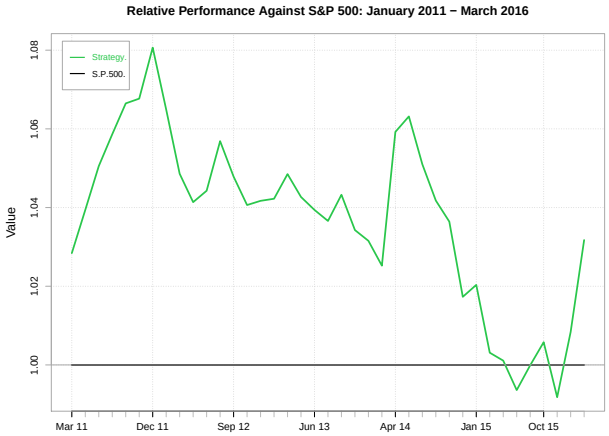


Figure 3: Relative Returns over Backtest

4.3.3 Capital Asset Pricing Model

MPT Capital Asset Pricing Model (CAPM), first developed by Craig W. French in 1962[5], provides an addition on top of the MPT framework with which an investor can

quantitatively justify passive or active investing. Table 4 shows that the strategy generates a Risk Premium of 0.30% over this backtest by remaining nearly market-neutral to its benchmark as evidenced by its 0.9405 Beta. Ultimately, this also generates an Alpha of 0.06% over the S&P 500. CAPM, a single-factor model based on risk, demonstrates that the strategy does not take on significant risk to capture additional returns over both its risk-free asset and benchmark.

Table 4: Strategy CAPM Metrics

	Strategy
Risk Premium	0.0030
Alpha	0.0006
Beta	0.9405
Bull Beta(β^+)	0.9855
Bear Beta(β^-)	0.9088
Timing Ratio	1.0844
Annualized Information Ratio	0.1393

5. CONCLUSION AND FUTURE WORK

In conclusion, we were successful in our attempt to use our analysis to create a smart-beta strategy that maintained small gains over the S&P 500. The primary limitations we encountered were difficulties with training the sentiment classifier and data scarcity for many of the sectors within our chosen corpus. These limitations suggest areas for future improvement. First, we anticipate significant gains if the sentiment classifier can be trained on a labeled corpus specifically constructed of financial text. This would however require a long time to build. The fact that the data we mined was sparse and came in low quantities means that the usefulness of an automated method for processing it is also limited because it is feasible for humans to simply read the documents. For this reason it is important to test the hypothesis that this approach will generalize to more diverse data sources. Finally it would also be interesting to see how alternative strategies for generating portfolios perform.

6. REFERENCES

[1] P. D. Azar. Sentiment analysis in financial news, 2009.

[2] Blackrock. Smart beta: Defining the opportunity and solutions, February 2015.

[3] Blackrock. Who we are, December 2015.

[4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.

[5] C. W. French. The treynor capital asset pricing model. *Journal of Investment Management*, 1(2):60–72, 1962.

[6] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[7] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[8] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October 2013. Association for Computational Linguistics.