Sentiment Analysis to Predict S&P 500 Behavior

[Proposal]

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

1. PROBLEM INTRODUCTION

The challenge we have decided to address in our project is how to use publicly available text based information sources produced by the United States government in order to predict the behaviors of financial markets. In particular, our hope is to apply sentiment analysis techniques to the minutes of the Federal Open Market Committee (FMOC) in order to identify correlations with the behaviour of the S&P 500. The FOMC meets eight times yearly and publishes approximately twenty pages of minutes for each meeting. These minutes discuss the state of the economy and detail a plan for the monetary policy in the time until the next meeting. The frequency of their publication, their scope, and their source of a key economic organization leads us to believe that these documents in particular would be a good source of potential insight.

Financial markets are an area where standard predictive techniques have often been unsuccessful at achieving accurate results. Stock prices in particular are notorious for being difficult to forecast. The ability to make accurate predictions about economic conditions provides an immediate financial incentive for research into this area and an entire industry exists whose primary goal is the development of techniques to best predict market behaviour. Beyond the pursuit of profit, having a warning period before downturns could also potentially allow for protective or even preventative actions to be taken.

Our particular approach to the problem, and the contribution we believe separates us from previous similar work, is our combination of topic modeling and sentiment analysis into a single pipeline. Our goal is to automatically parse the document to identify different topics such as the energy or health care markets. Applying sentiment analysis to these topics individually will allow us to create predictions for each of the sectors tracked by the S&P 500. With this we can long or short the corresponding S&P 500 sector through publicly traded ETF Sector SPDRs.

Moreover, if we are able to generate a strategy which is able to outperform its index benchmark, the S&P 500, significantly, our argument that our unique pipeline of performing sentiment analysis after topic modeling is the differentiating factor that many other sentiment analysis quantitative trading strategies have been lacking. Additionally, as we are applying a more granular approach by correlating sentiment to individual S&P 500 sectors, we are developing a beta-neutral strategy. That is, our strategy will hopefully only outperform the S&P 500 or track it very closely. This will then result in a $\beta_{S\&P500}$ very close to 1. This serves as another important distinction which we believe will help us more accurately predict future index performance.

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

2. RELATED WORK

As mentioned in the previous section, the allure of financial gains has inspired a large amount of research into the topic of S&P predictions in general and sentiment analysis in particular.

One popular source of data for sentiment analysis is social media. A paper by Mittal and Goel[6] used sentiment analysis on twitter data in combination with values from the Dow Jones Industrial Average (DJIA) to predict stock prices. In comparison to our work this paper used a simpler approach to sentiment analysis and focused primarily on using a neural network to transform the output from the analysis into concrete stock predictions. The paper was based on an earlier 2010 work by Bollen et al. that claimed to predict daily DJIA fluctuations with 87% accuracy.[3]

Another common source for data is news articles. For example Azar in 2009 demonstrated that positive returns could be derived from text analysis of new sources although a qualification was made that these returns vanish when trading costs are accounted for.[1] In contrast to both of these existing works we have chosen an official government source for data rather than a source that reflects popular opinion. The benefits of this decision are that our information comes directly from the entity that decides policy. Downsides are that it limits the amount of data available to us and the argument could be made that stock prices are largely driven by public opinion.

Some previous work has been done on the subject of combining topic modeling and sentiment analysis by Lu et al.[5] However, they restrict their application to analyzing restaurant reviews. We hope that in the process of applying these techniques to our particular situation we will discover modifications that better suit the algorithms to this application.

3. ALGORITHMS AND EVALUATION

We propose to use a mixture of topic modeling and sentiment analysis to judge the opinions shared by the Fed on each sector. Each Fed Report can be expected to cover opinions about a subset of general sectors that affect US economy. A particular sector can be generalized to describe a mix of some topics. A topic modeling algorithm like LDA[2] will be used to the find the topics covered by each sentence or group of sentences. Based on our prior knowledge of the topical mix of each sector, we intend to feed certain seed words for each topic as a prior for the topic modeling algorithm, to improve accuracy. Once the topic model is extracted for each sentence, we can do the sentiment analysis.

We intend to start off with certain seed words for positive/negative sentiment and then grow them with their synonyms from WordNet[4]. This also helps us assign sentiments to yet unseen words. Depending on how frequently the synonyms of the unknown word appear in either positive or negative sentiment word lists, the strength of sentiment polarity can be assigned. To combine sentiments in each sentence, we could potentially use named entity tagging to identify the 'subject' and then define a word window around the subject within which the sentiments will be calculated.

Once the sentiment values have been tagged with each topic model, values for each topic will be considered to be a part of the feature space to predict the S&P Index. The performance of different regression based models (linear, tree-based, SVM etc..) will be evaluated based on their predictive powers on the respective S&P 500 sectors. With this, we then intend to long or short the corresponding S&P 500 sector through the publicly traded Sector SPDR ETFs.

Financial evaluation will be carried out by primarily attempting to identify the predictive power of our technique. As we are employing a near beta-neutral strategy – a strategy that seeks to only generate alpha over its benchmark, which for our case would be the S&P 500 – we seek to generate as high an information ratio as possible. The information ratio is a ratio of a portfolio's returns above the returns of a benchmark, the S&P 500, to the volatility of those returns. In simpler words, the information ratio is a risk-adjusted rate of return over our benchmark. This will be particularly helpful in determining whether our strategy is able to identify sentiment accurately enough to generate additional returns over what would have normally been achieved by the market over this time. On that note, another financial metric we will be paying particular attention to is CAPM (Capital Asset Pricing Model) α. α is the active return of a strategy against a market index used as a benchmark. As such, a high α can help verify whether our strategy is able to pinpoint the sectors that will either over- or underperform the benchmark as a whole. A high α would indicate that we our sentiment analysis on each of our topic models is correctly correlating sentiment to sector returns.

4. REFERENCES

[1] Pablo Daniel Azar. Sentiment analysis in financial news. B.a. thesis, Harvard College, Cambridge, Massachusetts, 2009.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), March 2011, Pages 1-8, 2010.

[4] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[5] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. Multi-aspect sentiment analysis with topic models. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 81–88, Washington, DC, USA, 2011. IEEE Computer Society.

[6] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis.