SMU

SINGAPORE MANAGEMENT
UNIVERSITY

# Natural Language Processing on The Federal Reserve System

Stephen Kusrianto Jiang Jin Stanley Wong Hung Yee Lim Seng Hui

QF634 – Applied Quantitative Research Methods

December 2022

# Abstract

The influence that Central Banks have on the overall state of the economy with regards to their monetary policy and forward guidance is well known to many in the financial markets. But how important is the policymaker's choice of words? In this paper, we applied computational linguistics methods to assess the documents from Federal Open Market Committee (FOMC)'s meeting minutes, speeches of Federal Reserve Officials and Summary of Commentary on Current Economic Conditions (Beige Book). Applying sentiment analysis and topic modelling from natural language processing (NLP) models, we propose a trading strategy and an unique asset allocation strategy inspired by the All-Weather portfolio from legendary hedge fund manager, Ray Dalio.

**Keywords:** Central Bank, Natural language Processing, Topic modelling, Latent Dirichlet allocation, Sentiment analysis, Valence Aware Dictionary for Sentiment Reasoning, Cosine Similarity.

# 1 Introduction

There have been vast literatures on central bank policies (Blinder et al., 2008)[1], many of which suggested that communication is an essential toolkit to maintain monetary stability and achieve long-term macro-objectives set by the central bank.  Furthermore, central bank tone aids in the signaling of future policy decisions and the state of the economy (Hubert et al., 2021)[2].

Thus, central bank speeches and statements hold significant information. However, drawing meaningful insights has proven to be difficult given their generalized context in nature with less focus on financial indicators. With this in mind, we used computational linguistics methods to identify the tone of central bank speeches and generated topic models surrounding central bank messages.

To generate the topic models, we performed Latent Dirichlet allocation (LDA), an unsupervised machine learning method on central bank transcripts. We performed semi-supervised LDA by applying word2vec methods (Mikolov, T et al., 2013)[3] and guided our LDA models to targeted topic models attempting methodology as proposed by (Xue, M. 2019). Based on the topic distribution over central bank transcripts across time, we developed an asset allocation strategy inspired by the all-weather portfolio (Ray Dalio., 2011)[4] from Bridgewater Associates.

On the sentiment analysis side, 3 different methods were used, namely Textblob, Vader sentiments and Loughran-McDonald dictionary sentiments. Trading signals were created from the sentiment scores, and it was backtested on various asset classes.

# 2 Topic Modeling Approach

## 2.1 Data Methodologies

**Data Acquisition**

For data acquisition we utilized web scrapping method to extract our data of Federal Reserve Speech (2006 – 2022), Beige Book (1996-2021), Meeting Minutes (1995 – 2022), Statements (1994 – 2022). We used Selenium Webdriver[6] and Fedtools Python library scrapper to automate our scrapping process.

**Text Extraction and Cleanup**

We remove stop words, or words that appear too frequently in the text corpus, punctuation and ensure all letters are lowercased by using regular expression package library to ensure word format is suitable for computational encoding. Next, we lemmatized the words in our corpus using Natural Language Tool Kit (NLTK) to map all different forms of words to its base word – as known as, lemma.

**Tokenization of Text**

We tokenize the words by splitting individual words into a list of its single component in order to encode for computational analysis when we apply LDA models.

**Feature Engineering**

We apply Word2Vec and Document Term Matrix word vectorization methodologies to derive our input features to run our topic modelling.  Word2Vec represents each distinct word with a particular list of numbers called a vector. The vectors of words are assigned equal weightage from its neural network hidden layer such that the vectors capture semantic and syntactic qualities of words using cosine similarity scoring to measure the level of semantic similarity between the words represented by those vectors. For our project, we assigned top 50 similarity words based on target words ‘growth’ and ‘inflation’ to be used as labels for our input feature for LDA model.

While Document Term Matrix maps the tokenized words based on their occurrence frequencies into matrix form for LDA analysis.

**Topic Modelling**

We apply Latent Dirichlet Allocation (LDA), (Blei, et.al 2003)[10] and Ensemble Latent Dirichlet Allocation (ELDA) methodologies (Brigl, T. 2019)[12] for topic modelling. LDA draws semantic meaningful connection between the document corpus based on probability distribution of most relevant topics to map with the probability distribution of words.

As LDA has its limitation of model reproducibility issues on different rounds of model training, hence ELDA is introduced to addresses the issue by training multiple topic models and discard topics that do not occur across ensembles allowing for more consistent and reliable topic.

## 2.2 Findings, Insights and Model Selection

For the purpose of generating signals that are beneficial for our asset allocation strategy, we trained a multitude of LDA models with different parameters, with different types of data sources as their input features. We found that the topics generated from *Beige Books* do not provide any relevance to our strategy. As such, we excluded such models early in our model shortlisting process. In contrast, topic models from *Fed Minutes* in general provides reliable insights on the state of market, from the point of view economic regime. Below are our shortlisted models:

| Model Name | K Topics Set | Actual K Topics | Num Models | Num Passes | Perplexity Score | Umass Coherence | Runtime |
|---|---|---|---|---|---|---|---|
| **Full Fed** | 4 | 4 | 25 | 80 | -6.2335 | -0.3327 | 135 min 26 sec |
| **Full Fed W2V** | 4 | 4 | 10 | 50 | -6.2813 | -0.3591 | 33 min 45 sec |
| **Fed Minutes** | 4 | 2 | 25 | 100 | -5.9194 | -0.0177 | 48 min 48 sec |
| **Fed Statements W2V** | 4 | 3 | 25 | 50 | -5.0525 | -0.5673 | 8 min 22 sec |
| **Fed Speech W2V** | 4 | 4 | 25 | 50 | -6.9291 | -0.3639 | 163 min 34 sec |

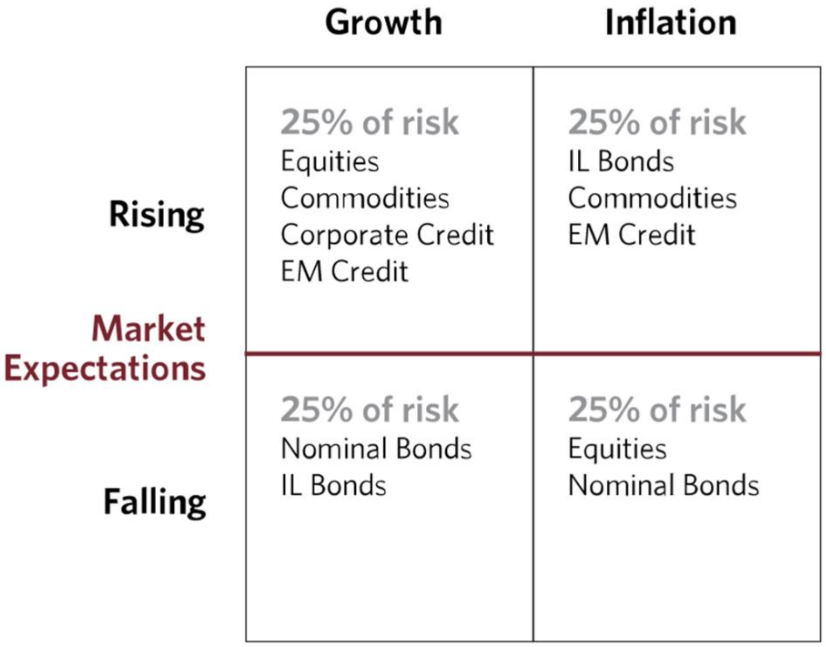
It is worth noting that all the above are Ensemble LDA Models. We realized that Ensemble LDA maintains a high degree of reliability, while minimizing variance and pruning non-reoccurring topics over a single LDA instance. However, it is indeed computationally more expensive to train.

Hyperparameter tuning we have done for LDA is as shown in the table above in *K* topics, and the number of passes applied is what we find most optimal. For ELDA we applied the number of passes in the table above as the most optimal so far. For Word2Vec vocabulary training on our corpus we applied the parameters of window = 100 words encompassing the target word, and 5 epoch as the most optimal setting for logical results based on similar words.

The shortlisting criteria are primarily based on the model’s ability to link topics toward inflation and growth as a theme in the set of documents that form our corpus. We also selected top 5 models with the lowest Perplexity Score (Blei et al., 2003)[1] and UMass Topic Coherence Score closest to zero (Mimno et al., 2011)[3]. A lower perplexity score implies that a trained model is able to linearize semantic relationships better, hence making it possible to predict topics from new datapoints with higher confidence. UMass Coherence score of zero implies perfect coherence of topics generated, meaning the model can accurately distinguish topics that are semantically interpretable, versus topics that are merely of statistical inference.

We have contesting models regarding the two metrics. In terms of lowest perplexity, our best model, the semi-supervised *'Fed Speech W2V'* wins, but in terms of UMass *Fed Minutes* is the closest to perfect coherence. Nevertheless, as quoted from Rob J Hyndman, *"A model which fits the data well does not necessarily forecast well."* (Hyndman, 2014)[4]. Thus, we move on to the next step, which is to backtest the signals and benchmark our model's strategy performance against the original All-Weather Portfolio.

## 2.3 Trading Strategy



| Market Expectations | Growth | Inflation |
|---|---|---|
| Rising | 25% of risk<br>Equities<br>Commodities<br>Corporate Credit<br>EM Credit | 25% of risk<br>IL Bonds<br>Commodities<br>EM Credit |
| Falling | 25% of risk<br>Nominal Bonds<br>IL Bonds | 25% of risk<br>Equities<br>Nominal Bonds |

Source: The All Weather Story. (2012, Jan). Bridgewater Associates.
https://www.bridgewater.com/research-and-insights/the-all-weather-story

Using the signals generated from the LDA models, we implement an asset allocation strategy where the assets we will be investing in depend on the market expectation and the topic model that is generated out.

To determine if the market expectation is rising or falling, we will be using the VIX index as a proxy. If the VIX for that day is above 20, the market expectation is falling, and vice versa.

For example: If the VIX is below 20 and the topic model from the latest federal reserve minutes is growth, it means that we are in the quadrant representing rising market expectations and growth. Therefore, we will be investing in equities, commodities, corporate credit and EM (Emerging Markets) credit.

To replicate each of the asset classes in the above quadrant, we will be selecting exchange-traded fund (ETF) representing each of the asset classes as shown in the table below. The ETF is chosen based on the size of asset under management (AUM) and whether it is a tracking ETF or not.

| Asset Class | ETF |
|---|---|
| Equity | SPY |
| Commodities | DBC |
| Corporate Credit | LQD |
| Emerging Market(EM) credit | EMB |
| Nominal bonds | BND |
| Inflation Linked (IL) bonds | TIP |

Here are the results of our backtest results for shortlisted Ensemble LDA models:

| Model Name | CAGR | Annualized Sharpe | Annualized Vol | Maximum Drawdown Period | Maximum Drawdown (%) |
|---|---|---|---|---|---|
| Full Fed | 4.06% | 0.5893 | 6.53% | 526 Days | 32.44% |
| Full Fed W2V | 4.21% | 0.6078 | 6.57% | 531 Days | 32.90% |
| Fed Minutes | 4.82% | 0.8086 | 5.76% | 350 Days | 20.54% |
| Fed Statements W2V | 4.56% | 0.7608 | 5.78% | 349 Days | 20.43% |
| Fed Speech W2V | 5.21% | 0.9574 | 5.03% | 279 Days | 20.28% |

When we compare the *Full Fed* model with the *Full Fed W2V* model, there is a slight increase in CAGR and annualized Sharpe Ratio. This may imply that using word2vec, in conjunction with Ensemble LDA to semi-supervise training contributes to improved model performance.

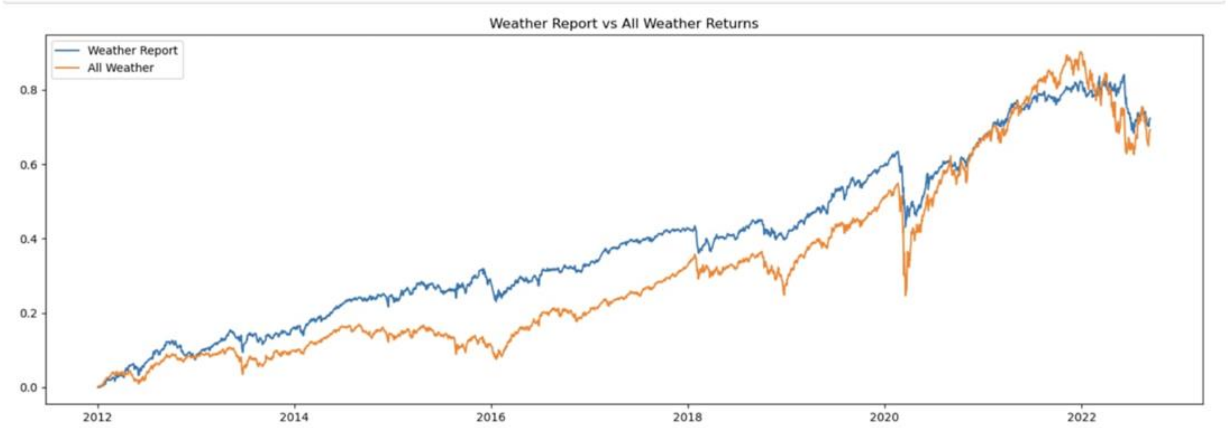


Fig: Fed Speech W2V model (best performing model)

The best performing model by far is the *Fed Speech W2V*. It has the lowest annualized volatility with the highest compounded annual growth rate (CAGR) and Sharpe ratio. It is interesting to note that using entire range of federal reserve documents does not generate the highest performance. In fact amongst the shortlisted models, Full Fed models perform the worst.

# 3 Sentiment Analysis Approach

## 3.1 Data Methodologies

Text data consist of information from the federal reserve system:

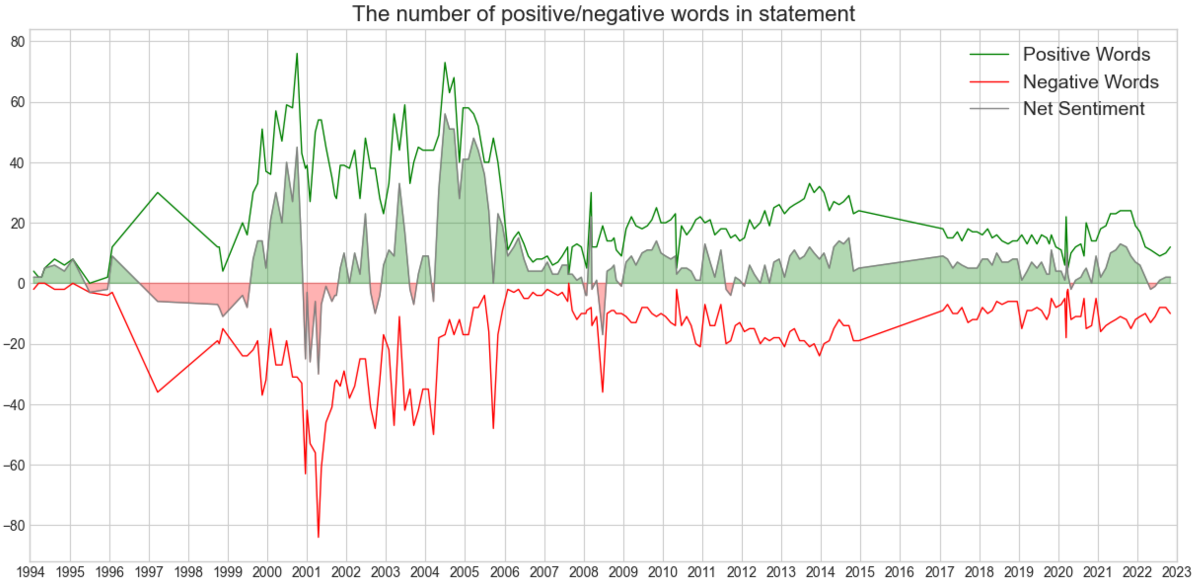
1. Beige book: economic conditions before meetings
2. Statements: policy decision immediately after meetings; 8 times a year
3. Minutes: detailed record of meeting 3 weeks after statements
4. Speeches: by fed officials in between meetings

Non-text data consist of ETFs from equity, bonds, and commodities & others:

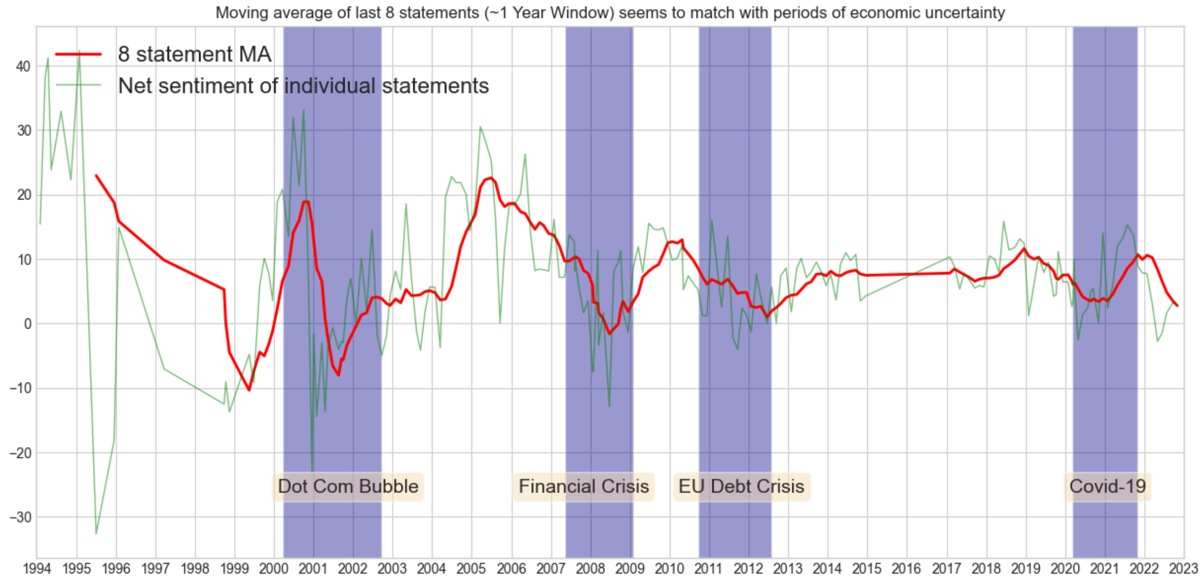
1. **SPY:** SPDR S&P 500 ETF Trust
2. **BND:** Vanguard Total Bond Market ETF
3. **TLT:** iShares 20+ Year Treasury Bond ETF
4. **SHY:** iShares 1-3 Year Treasury Bond ETF
5. **BIL:** SPDR Bloomberg 1-3 Month T-Bill ETF
6. **TIP:** iShares TIPS Bond ETF
7. **GLD:** SPDR Gold Shares
8. **DBC:** Invesco DB Commodity Index Tracking Fund
9. **VNQ:** Vanguard Real Estate ETF

A combined data frame is created to perform analysis.

## 3.2 Exploratory Data Analysis



We found that number of positive words tend to be inversely proportional to number of negative words. The average is on the positive side. Next we apply the moving average to the net sentiment to see the trend plot with recession period.



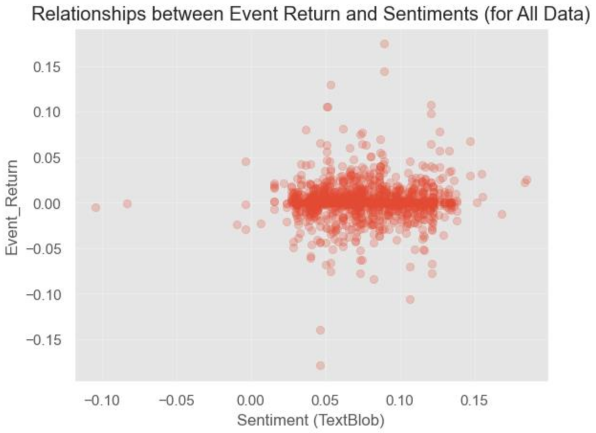
The 8-statement moving average goes down during economic uncertainty. This suggests the net sentiment correlates with macro-economic environment, but it can be a noisy for prediction. Overall, the data makes sense and we can use this for NLP.

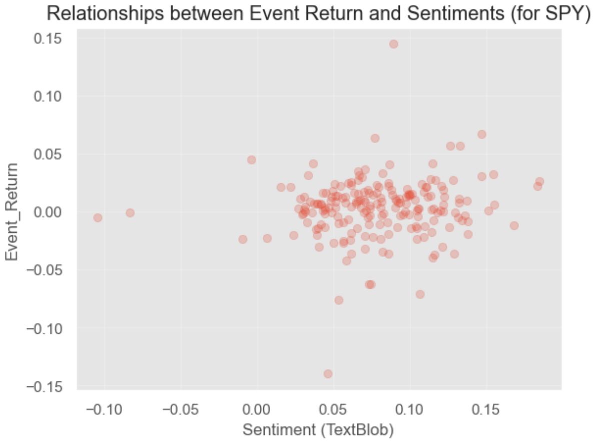
## 3.3 Interpretation of Results

### NLP Technique 1: TextBlob

This is a pre-trained model based on the Naïve-Bayes classification algorithm that maps:

1. Polarity score: convert sentences into a numerical value of sentiment between -1 to +1 (positive-negative).
2. Subjectivity score: convert sentences into a numerical value of subjectivity between 0 to 1 (objective-subjective).

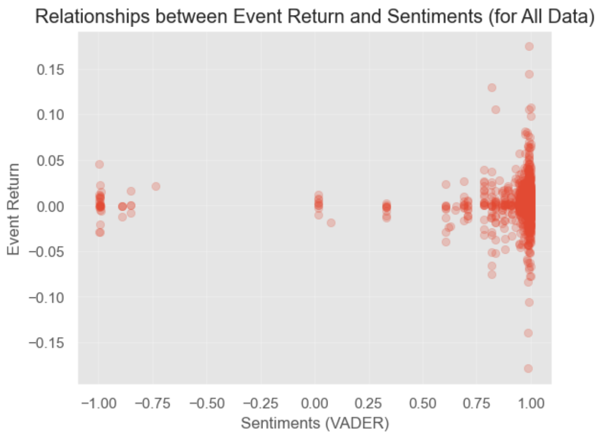


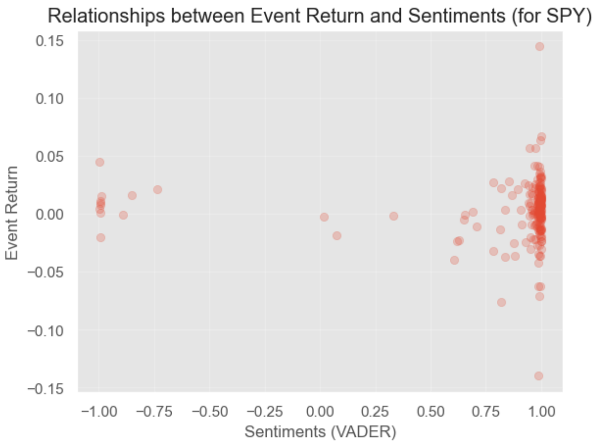


There isn’t strong correlation between event returns and sentiments. Most of the points on the scatterplot cluster around 0 which indicates a weak signal.

## NLP Technique 2: VADER

This is a pre-built sentiment analysis model included in the NLTK package. Lexicons are special dictionary or vocabularies created for analyzing sentiments. We train this VADER model based on a financial lexicon (~20k words) and apply it to analyze sentiments of fed statements.

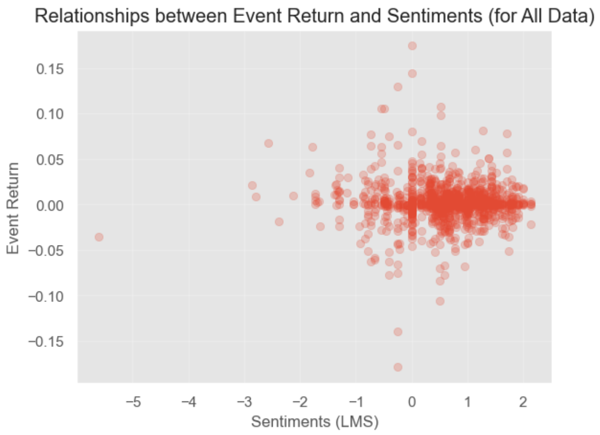


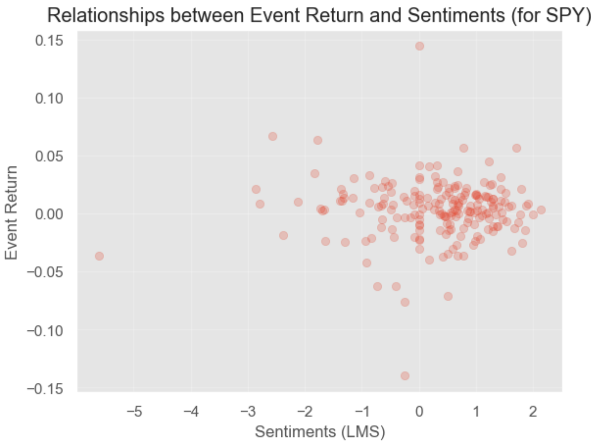


The correlation between event returns and sentiments shows an improvement compared to the TextBlob technique. However, most points still cluster between 0.75 to 1.00.

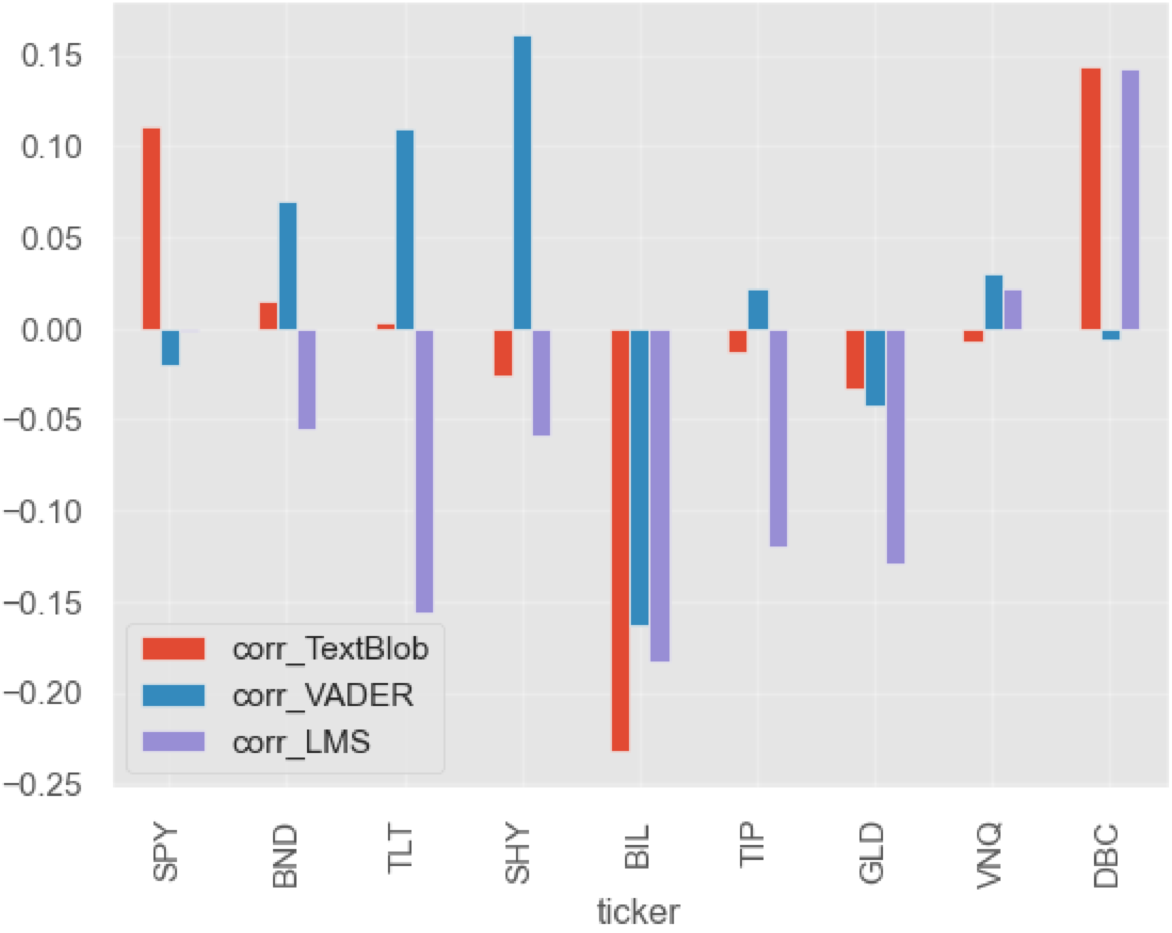
## NLP Technique 3: LMS

This is the gold standard that examines tokens from all 10-K type filings for the full EDGAR 10-K archive and earnings calls from CapIQ. Sentiment categories include negative, positive, uncertainty, litigious, strong modal, weak modal, and constraining. We train the model based on LMS lexicon (~4k words) and apply it to analyze sentiments of fed statements.





## Model Comparisons



When economic conditions are favorable, fed sentiments are positive and the policy decision is to raise interest rates. Conversely, when economic conditions are unfavorable, fed sentiments are negative and the policy decision is to decrease interest rates.

Insights that we discover:

1. Increasing interest rates environment is bad for all asset classes, except commodities (DBC) and real estate (VNQ);
2. Commodities and real estate tend to outperform the market during an inflationary growth economic condition;
3. Long-term treasury (TLT) is more sensitive to interest rates than short-term treasury (SHY); and
4. T-Bills (BIL) is the most sensitive to interest rates.

Hence, we decided to use NLP technique 3 LMS to backtest our trading strategy.

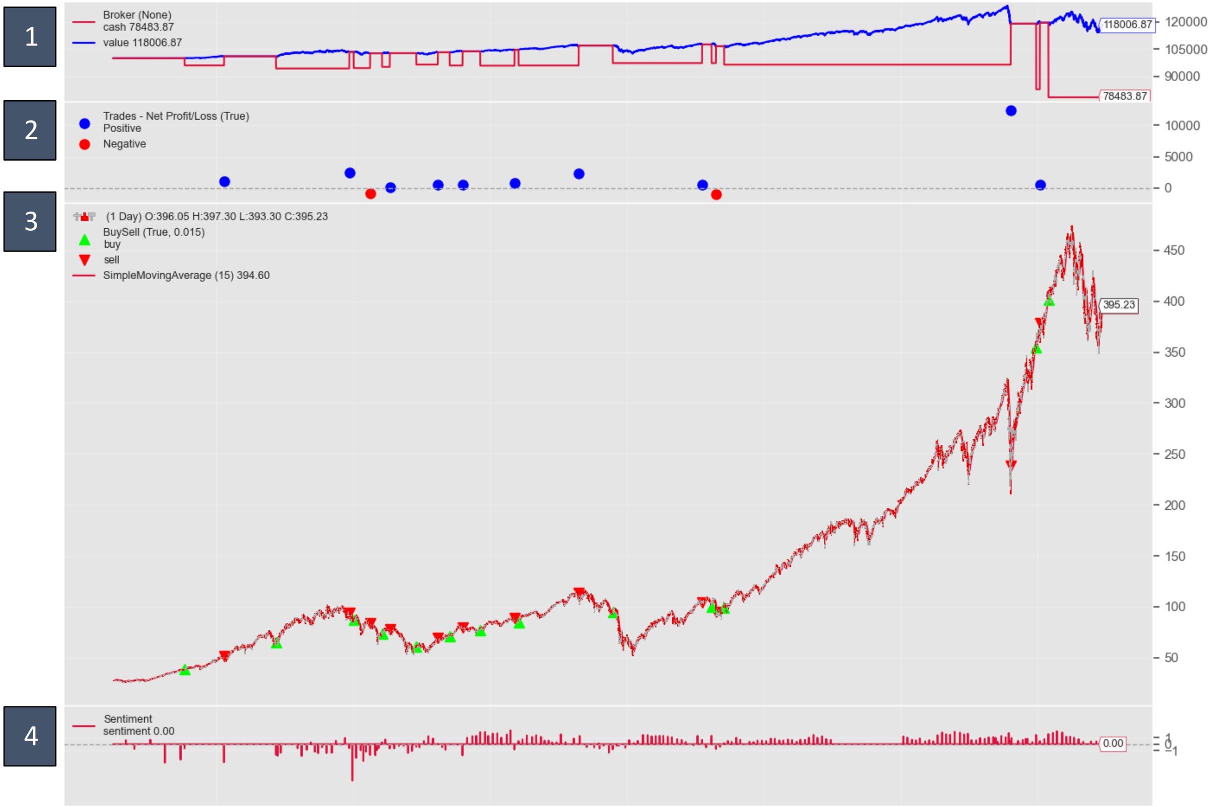
## Backtesting Strategy

We start with an initial capital of $100,000 to buy a stock when the change in sentiment score (current sentiment score - previous sentiment score) is greater than 0.5 and sell a stock when the change in sentiment score is less than -0.5. We check the 15-day moving average when buying or selling with position size of 100 units per trade.

### Performance of Trading Strategy (for All Data)

| | SPY | BND | TLT | SHY | BIL | TIP | GLD | VNQ | DBC |
|---|---|---|---|---|---|---|---|---|---|
| **Per Unit Start Price** | 27.54 | 46.73 | 41.34 | 58.12 | 82.90 | 55.84 | 44.38 | 22.74 | 22.05 |
| **Strategy Profit** | 18,006.86 | 199.33 | 2,075.47 | 2,125.70 | 465.73 | 6,372.85 | 14,335.00 | 1,918.22 | 389.39 |

### Plotting the Trading Strategy (for SPY)



**Panel 1:** The first panel is the Cash Value Observer, which, as the name implies, keeps track of the cash and total portfolio value during the life of the backtesting run. As we can see in this panel, we started with $100,000.00 and the final value at the end is $118,006.87.

**Panel 2:** The second panel is Trade Observer, which shows, at the end of a trade, the actual profit and loss. A trade is defined as opening a position and taking the position back to zero (directly or crossing over from long to short or short to long).

**Panel 3:** The third panel is Buy/Sell Observer, which plots (in addition to the prices) where buy and sell operations have taken place.

**Panel 4:** The fourth panel shows the sentiment score.

The profitable strategy is a proof of concept that sentiment data can be used in different ways for the trading strategy. Sentiment scores can be used as directional signal and ideally create a long-short portfolio, by buying the ETFs with a positive score and selling the ETFs with negative score.

The sentiments can also be used as additional features over and above other features (such as correlated stocks and technical indicators) in a supervised machine learning model to predict the price or come up with a trading strategy.

There can be many ways to create a trading strategy based on sentiments, by varying the threshold for change in sentiment and changing trading position size based on the initial cash available. Hence, we added a sensitivity analysis of +/- 10% to measure how it affects trading performance.

<u>Sensitivity Analysis (for All Data)</u>

| **SPY** | | **Position size** | | |
|---|---|---|---|---|
| | | **90** | **100** | **110** |
| **Change in sentiment** | **0.55** | +2.10% | +13.44% | +24.78% |
| | **0.5** | -10.00% | $18,007 | +10.00% |
| | **0.45** | -31.63% | -24.04% | -16.44% |
| **BND** | | **Position size** | | |
| | | **90** | **100** | **110** |
| **Change in sentiment** | **0.55** | -158.79% | -165.33% | -171.86% |
| | **0.5** | -10.05% | $199 | +10.05% |
| | **0.45** | -49.25% | -43.72% | -37.69% |
| **TLT** | | **Position size** | | |
| | | **90** | **100** | **110** |
| **Change in sentiment** | **0.55** | -27.08% | -18.99% | -10.89% |
| | **0.5** | -10.02% | $2,075 | +10.02% |
| | **0.45** | -9.98% | NM | +10.02% |
| **SHY** | | **Position size** | | |
| | | **90** | **100** | **110** |
| **Change in sentiment** | **0.55** | -26.25 | -18.06% | -7.61% |
| | **0.5** | -10.02% | $2,126 | +9.97% |
| | **0.45** | -24.84 | -16.51% | -8.14% |
| **BIL** | | **Position size** | | |
| | | **90** | **100** | **110** |
| **Change in sentiment** | **0.55** | -33.05 | -25.54% | -18.03% |
| | **0.5** | -10.09% | $466 | +9.87% |

| | | | | |
|---|---|---|---|---|
| | 0.45 | +5.36% | +16.95% | +28.76% |
| TIP | | Position size | | |
| | | 90 | 100 | 110 |
| Change in sentiment | 0.55 | -10.00% | NM | +10.00% |
| | 0.5 | -10.00% | $6,373 | +10.00% |
| | 0.45 | -23.44% | -14.92% | -6.42% |
| GLD | | Position size | | |
| | | 90 | 100 | 110 |
| Change in sentiment | 0.55 | -16.85 | -7.62% | +1.62% |
| | 0.5 | -10.00% | $14,335 | +9.37% |
| | 0.45 | -12.84% | -3.15% | +6.53% |
| VNQ | | Position size | | |
| | | 90 | 100 | 110 |
| Change in sentiment | 0.55 | +60.06% | +77.84% | +95.62% |
| | 0.5 | -10.01% | $1,918 | +10.01% |
| | 0.45 | -54.85% | -49.84 | -44.84% |
| DBC | | Position size | | |
| | | 90 | 100 | 110 |
| Change in sentiment | 0.55 | -37.19 | -30.15% | -23.37% |
| | 0.5 | -12.06% | $398 | +7.54% |
| | 0.45 | -165.08% | -283.42% | -301.76% |

# 4 Conclusion

## 4.1 Sentiment Analysis: Limitations and Future Work

1. Lack of training data (8 statements per year)
2. Data trade-offs (statements are timely but lack context compared to minutes; minutes gives more insights but lagged by 3 weeks giving little applicability)
3. Data quality (irrelevant and long paragraphs, ML good at learning ~500 words, splitting words by overlapping 200/50 but lose context, etc.)
4. Use models such a TFIDF, LSTM/RNN, and BERT.

## 4.2 Topic Modelling: Limitations and Future Work

(Naushan, H 2020) explained the limitations of topic modelling of LDA that affects its model performance in the presence of low data input or short document length, exchangeability issue that may affect the accuracy of the results. Word2vec has its limitation on the small dataset input size that may not be able to interpret words with multiple meanings, and inability to handle out-of-vocabulary (OOV) words as explained by (Horn. F. 2017).

(Angelov, D. 2020), proposed Top2Vec as a newer model that aims to resolve these limitations. With in-built preprocessing tools, and the introduction of concept of joint embedding of document and word vectors to draw more accurate semantical topic modelling that search for the centroid of a dense cluster of documents and words. It also applies high dimensionality reduction using UMAP for more accurate results prediction. Our group aims to attempt this methodology for our future project.

As (Sbalchiero, S., & Eder, M. 2020) alluded that LDA as a topic modelling technique as a classifier is a relatively weak solution in general because they are a generative model, whereas classification is a discriminative problem. Hence for our future project we consider using transformer like BERTopic for larger corpus or Non-negative matrix factorization for smaller ones and supervised machine learning with hyperparameter tuning with macroeconomic indicators as input feature to generate signals.

i) Strategy Limitations of weather forecast portfolio

Firstly, for the all-weather portfolio (benchmark used) that was plotted, there was no rebalancing at all while the actual all-weather portfolio rebalances on a quarterly or annual basis.

Secondly, using VIX as proxy for market expectations is a very simplistic inference.

i) Strategy Futureworks

Instead of using VIX, we can use social media data to map out market expectations and use it in the asset allocation strategy.

We can also use retail topic models to compare with central bank topic models and come up with insights that can be useful in generating trading signals.

Lastly, using data from different central banks to generate topic models that can be useful in deciphering regulator’s stance on the country’s economy.

# 5 References

1. Blinder, A. S., Ehrmann, M., Fratzscher, M., de Haan, J., & Jansen, D.-J. (2008). Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence. *Journal of Economic Literature*, *46*(4), 910–945. http://www.jstor.org/stable/27647085
2. Hubert, P., & Labondance, F. (2021). The signaling effects of Central Bank tone. *European Economic Review*, *133*, 103684. https://doi.org/10.1016/j.euroecorev.2021.103684
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, September 7). *Efficient estimation of word representations in vector space*. arXiv.org. Retrieved December 14, 2022, from https://arxiv.org/abs/1301.3781
4. Xue, M. (2019, January). A text retrieval algorithm based on the hybrid LDA and Word2Vec model. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)* (pp. 373-376). IEEE. https://doi.org/10.1109/ICITBS.2019.00098
5. Dalio, R. (2011, August). *Engineering targeted returns and risks* . Retrieved December 14, 2022, from https://bridgewater.brightspotcdn.com/fa/e3/d09e72bd401a8414c5c0bdaf88bb/bridgewater-associates-engineering-targeted-returns-and-risks-aug-2011.pdf
6. Selenium. (n.d.). *The Selenium Browser Automation Project.* https://www.selenium.dev/documentation/webdriver/
7. Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*. https://doi.org/10.48550/arXiv.1309.4168
8. Řehůřek, R., & Sojka, P. (2010) *Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. (pp. 45-50). ELRA. http://is.muni.cz/publication/884893/en
9. PYPI. (n.d.). *Fedtools 0.0.7.* Fedtools · PyPI
10. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent dirichlet allocation - Journal of Machine Learning Research*. Journal of Machine Learning Research. Retrieved December 14, 2022, from https://jmlr.org/papers/volume3/blei03a/blei03a.pdf
11. Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Chapter 1. NLP: A Primer. *Practical natural language processing: a comprehensive guide to building real-world NLP systems (1st Eds)*. (pp. 3-36). O'Reilly Media, Inc.
12. Brigl, T. (2019). *Extracting Reliable Topics using Ensemble Latent Dirichlet Allocation* [Bachelor Thesis]. Technische Hochschule Ingolstadt. Munich. https://www.sezanzeb.de/machine_learning/ensemble_LDA/
13. Naushan, H. (2020, Dec 3). Topic Modeling with Latent Dirichlet Allocation. In *Medium*. https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8
14. Horn, F. (2017). Context encoders as a simple but powerful extension of word2vec. *arXiv preprint arXiv:1706.02496*. https://doi.org/10.48550/arXiv.1706.02496
15. Mimno, D., Wallach, H. M., Talley, E., Leenders, M., &amp; McCallum, A. (2011). Optimizing semantic coherence in topic models - Cornell University. Cornell Bowers Information Science. Retrieved December 14, 2022, from https://mimno.infosci.cornell.edu/papers/mimno-semantic-emnlp.pdf

16. Hyndman, R. J. (2014). *Measuring forecast accuracy - Rob J. Hyndman*. Rob J Hyndman. Retrieved December 15, 2022, from https://robjhyndman.com/papers/forecast-accuracy.pdf
17. Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. Quality & Quantity, 54, 1095-1108. https://link.springer.com/article/10.1007/s11135-020-00976-w