Market Regime Modeling with Natural Language Processing

Dhruv Baid (dbaid) Prajakta Phadke (prajphad) Uday Sharma (udayshar)

Pittsburgh Campus

Abstract

This project used natural language processing (NLP) to study, detect, and predict changes in market regimes to improve existing trading strategies. We used text data from the Federal Open Market Committee (FOMC) meeting minutes and price data of the S&P 500. We used Non-Negative Matrix Factorization (NMF) to study the characteristics of regimes. We labeled regimes using a hidden Markov model and spherical K-means clustering. Using the labels as the response and price as covariates, we fit a naive Bayes model to predict the regime. We also try adding the NMF model topic weights and word embeddings from FinBERT as covariates. The performance was evaluated on the test set by comparing the metrics of a regime-dependent trading strategy for different combinations of labeling methods and covariates. Using NMF, we found that text data helped identify regimes that corresponded to different macroeconomic topics. However, neither the regimes identified using text data nor the text covariates were able to improve the performance of the trading strategy.

1 Introduction

Our project aims to augment a prior study on regime detection with NLP methods to study, detect, and predict regime changes in financial markets with the intent of improving existing trading strategies. We do this for the S&P 500 index.

A regime change can be defined as a significant, and often abrupt, change in the behavior of financial markets. This could be limited to an asset class, region, or currency, or it could be universal. Understanding and predicting regime changes are of great importance to all participants in the market as they can cause asset prices to behave erratically. In the event of a regime change, trading strategies based on assumptions on price dynamics need to be re-calibrated or even scrapped altogether (Tsang and Jun, 2018). This is of particular interest to quantitative investors, whose models might implicitly be calibrated to particular regimes they are trained on. Past literature has found that the market has historically operated under one of two regimes – a stable regime and a distressed regime (Chen and Tsang, 2021).

Regime changes usually coincide with major changes in the global economy, policy, or regulations (Ang and Timmermann, 2011). NLP techniques are relevant for analyzing speeches, announcements, or company filings which accompany these changes. The most significant data, arguably, are minutes from the eight annual meetings held by the FOMC (The Federal Reserve, n.d.-b), which describe the broad economic and financial conditions, the Fed's chosen monetary policy stance, and its risk assessments for its long-term goals of price stability and sustainable economic growth (The Federal Reserve, n.d.-a). We used the FOMC meeting minutes from 1985-02-13 to 2023-06-14, downloaded from a Kaggle dataset (Bickerton, 2023) that was scraped from the Fed's website. In addition to the text data, we also had daily price data for the corresponding dates. We converted the traditional price time series, which are sampled at fixed intervals, into directional change (DC) data (Chen and Tsang, 2021). Under this event-based framework, we constructed a subset of the price data that has an alternating sequence of uptrends and downtrends (Petrov et al., 2019).

To study the regimes, we performed an unsupervised learning task where we used NMF to construct a time series of the topic weights for these documents and assigned topics to each document based on the highest weight. To detect and predict regimes, we performed a supervised learning task. We first used spherical K-means to create regime labels on the training set using the text data. We also used a hidden Markov model as a benchmark model to get regime labels using the price data. Using these labels, we fitted three naive Bayes classifiers – one that only used the price as a covariate, and two that used different word embeddings. For the test set, we evaluated the model performance by comparing the results between regime-dependent and control trading strategies. We sought to answer the following research questions:

1. Do the regimes, as identified by unsupervised NLP methods, show any distinct characteristics?
2. Does using text data in identifying regimes improve the performance of our trading strategy?
3. Does using text data as covariates to predict regime help improve our trading strategy?

2 Methods

2.1 Data

2.1.1 Text Data

We used the text of the FOMC meeting minutes for the eight regularly scheduled annual meetings. The policy decisions and actions are detailed in the meeting minutes issued by the FOMC three weeks after each meeting (The Federal Reserve, n.d.-b). These minutes have been published under three categories: `minutes, historical minutes`, and `minutes of actions` (The Federal Reserve, n.d.-c). Each meeting generates a minutes document, which constitutes a single observational unit of text or document in our dataset.

The FOMC meeting minutes are available from 1936 (The Federal Reserve, n.d.-c). However, we focused our analysis on 1985 onwards. This is generally considered to be the modern financial era, with the markets, regulatory, and banking systems remaining relatively consistent (James, 2002). Meeting minutes are also available for the entirety of this period, from February 1985 to June 2023. There are a total of 309 documents and 6716 unique words in the dataset vocabulary.

Every meeting minute document published by the FOMC follows a particular structure (Jegadeesh and Wu, 2015): a roll call, followed by the meeting. This roll call consists of the names and designations of the attendees and does not contain any pertinent information about policy. We noticed that the phrases "unanimous vote" or "the manager" always indicated the end of the roll call and used that as a marker for relevant data by truncating the text that came before it. All the documents had at least one of the two phrases occur after the roll call.

We tokenized the data using `nltk`, converted all words to lowercase, and removed all non-alphabetic characters and stop words. We used a customized stop words list that included both the standard English stop words and stop words relevant to FOMC data, such as `minutes`, `members`, `board`, `meeting`, `reserve`, and so on. After stop word removal, the vocabulary size decreased to 5057 words. We do not perform stemming or lemmatization to preserve maximum tonality and information in the document.

For unsupervised learning, we use the TF-IDF representation since this representation emphasizes unique and discriminatory words, and hence provides more value than using raw counts. This is important since we wish to identify regime changes, which are often indicated by distinct changes in market conditions.

2.1.2 Price Data

We used Yahoo Finance (and its Python API) to obtain daily price data from January 1, 1985 to June 30, 2023 for the S&P 500. We used a directional change (DC) indicator instead of the raw price. Under the DC framework, we sample a subset of the time points where the price moves beyond a certain threshold. The advantage of this approach is that it allows us to focus on significant market movements and filters away a lot of the noise commonly present in financial data. The method is explained in detail in Appendix A.

2.2 Unsupervised Learning

We utilized two methods of unsupervised learning on the meeting minutes data. We performed topic modeling using NMF on the documents to understand the topic, and hence regimes, that could be inferred from the FOMC data and to see how they changed over time. We also clustered the data using spherical K-means clustering to label the training data for the supervised learning task.

2.2.1 Topic Modeling using Non-Negative Matrix Factorization

Topic modeling is useful to summarize and understand the themes underlying the text data. FOMC meeting minutes go into details regarding monetary policy decisions and market conditions so it is instructive to look at the broader themes. We also use topic modeling to conduct a temporal analysis of topics. Furthermore, topic modeling is an effective dimension-reduction technique when used in conjunction with a supervised learning approach.

We use NMF as our topic model since we want to study the distribution of topics (themes) over time. We cannot use LDA as that requires using word counts (instead of TF-IDF), and since we want to use weights (loadings) on topics to interpret the regimes. We do not use LSA as the results aren't as easily interpretable. To choose the optimal number of topics, we use K-fold cross-validation with 5 splits. We used coherence scores to assess the interpretability and effectiveness of topic generation. We compare and contrast the topics chosen by two coherence scores - C_V score and UMass score. The C_V score consists of both an indirect and direct measure, whereas the UMass score is a direct measure. After choosing the optimal number of topics, we assign each document the topic with the highest weight found using NMF.

In our analysis, we primarily focus on three aspects:

- Topic loadings of each document
- Unconditional distribution of each topic
- Behaviour of assigned topics over time/documents

We also perform a manual inspection of the topics generated using word clouds to analyze the interpretability of the topics.

2.2.2 Spherical K-Means Clustering

Clustering helps identify similarities among the documents. We can use clustering as a method to label data. We perform spherical K-means on the TF-IDF representation of the text data. For our use case, we want to generate regime labels for the training data. Hence, we want a hard clustering method, i.e., we want each observation to be assigned to only one cluster. That narrowed our choice to spherical K-means or hierarchical clustering.

An added benefit of K-means clustering is that we can explicitly choose the number of clusters K (versus choosing t for hierarchical clustering). This is useful for our project since this can reflect how we choose to define our regimes, e.g., we can have $K = 2$ if we think there are 2 regimes. We cluster the data into 2 clusters, where each represents a different regime. This labeling is then used to compare and contrast against the hidden Markov model-generated labels. There are different ways to generate regime labels, and this is one of them. There is no true regime labels to begin with, and therefore, no incorrect way of generating them.

We also generate word clouds to qualitatively analyze the words associated with each cluster to correlate frequently occurring words and the associated regimes.

2.3 Supervised Learning

2.3.1 Pipeline

Our goal is to develop a classifier that can predict the regime the market is in and use that to improve the performance of an existing trading strategy. We set up the pipeline as follows:

- We use two methods to generate labels for the training set from January 1, 1985 to December 31, 2019 – the spherical K-means clustering with $K = 2$ as described in the previous section, and a hidden Markov model on the DC indicator. When using the K-means clustering, as the text data are only available periodically, we backward-fill the label between meeting dates (except the final data point in the training set, where we must forward-fill). This ensures we have a label for every date we have the price for. The advantage of backward-filling is that the labeling is more accurate – when there is a change in the market, it takes some time for the Fed to detect it and discuss it in the next FOMC meeting, backfilling will ensure that we incorporate this information in the labels that come before it. We are backward-filling to label the data and not make predictions on it, so look-ahead bias is not a problem.

  The justification for using spherical K-means is given above. We chose to use hidden Markov models as the benchmark for price data as other studies found it to be an effective method of detecting the regime (Chen and Tsang, 2021).

- We use a naive Bayes classifier for the supervised learning task as we want to be able to predict the regime label for the test set. We chose to use it as it is relatively simple to use and computationally fast. It was also found to have the best performance for the regime modeling task when using hidden Markov models to label regimes (Baid et al., 2023).

- We compare the performance of the classifier on the test set from January 1, 2020, to June 30, 2023, for different combinations of the response and predictors. We use two sets of labels from before, along with three different predictors (separately) – the DC indicator for price data, the topic loadings from the NMF model described before, and word embeddings from FinBERT, a language model pre-trained on financial news articles.

The text data of interest are relevant for this task as we want to find out whether the FOMC meetings contain information that can help predict the regime. This has important implications for quantitative investors looking for additional data to improve their existing trading strategies. Although market participants commonly use the meetings to inform their decisions, we are conducting a more rigorous and quantitative analysis.

2.3.2 Text Representations

We use two different representations of text as covariates in the supervised learning task.

FinBERT Word Embeddings: BERT is a popular large-scale transformer model, and FinBERT is a BERT model pre-trained on financial news text. We can utilize word embeddings from the FinBERT model to use for our classification task. For each document, we get the 768-dimensional embedding by averaging over the FinBERT-generated embedding of each token (word) in that document. Using embeddings will allow us to capture the semantic meaning of the words by looking at the context they're used and might help us better identify the regime. For example, the embeddings might allow the model to better contrast between a "growth" and a "recession" regime.

Topic Loadings from Non-Negative Matrix Factorization: NMF, as described before, allowed us to generate the loadings for different topics for each document, i.e., discover some sort of latent structure in the document. This structure could be useful in predicting the regime as well – depending on what state the economy is in, the Fed might choose to focus on different topics in their meetings. For example, the Fed might focus more on inflationary pressures in a certain regime and on trade wars in another. The Python function that we used from `gensim` seems to have normalized the loadings to be probabilities, but that is not important for our project.

2.3.3 Evaluating the Learning Task

Our goal in predicting the regime is to improve the performance of an existing trading strategy. So, to evaluate the classifiers, we compare the performance of a regime-dependent trading strategy on the test set from January 1, 2020, to June 30, 2023, to two control strategies (mean-reverting and momentum-based) that don't look at the regimes. The strategies are described in detail in Appendix B. The regime-dependent trading strategy uses the regime predictions as an input, so a better performance would imply that our predictions are more useful. We chose not to use something like the misclassification rate as a metric since we would just be evaluating how well we predict some arbitrary labels that we came up with in the first place. Evaluating the trading strategy performance is a more direct way to look at the outcome we're interested in. The strategies are evaluated based on the Sharpe ratio, total profit, and max drawdown, which are explained in detail in Appendix B.

To avoid data leakage/look-ahead bias, we first label the training set (where the labels implicitly incorporate information from the future via back-filling) and train a classifier on it. The predictions are then made on the test set using covariates which only make use of information available upto the given date. For text data, we forward-fill the predictions and covariates between two data points so that it matches the frequency of the price data without introducing any look-ahead bias. A minor point – as the first FOMC meeting minutes data for the test set is available in February, we chose to back-fill it instead of forward-filling the previous data point to avoid leakage from the training set. The look-ahead bias it introduces is negligible as the FOMC meeting was actually held in January (although the data became available later), and the portion represents a small proportion of the test set.

3 Results

3.1 Unsupervised Learning

3.1.1 Topic Modeling

As described in the methods section, we used K-fold cross-validation with 5 splits on an NMF model to find the optimal number of topics on the training data. We used two coherence scores as the evaluating metric – C_V and UMass.

Table 1: Cross-validation using coherence score for optimal number of topics.

	2	5	7	10	optimal choice
C_V	0.356024	0.482005	0.4515078	**0.496230**	10
UMass	**-0.788143**	-1.131355	-1.390159	-1.237912	2

We used the `pyLDAvis` library to visualize the different topics and the top 10 words associated with each topic. These visualizations are interactive, therefore, we have attached a screenshot for each score in Figure 1. The inter-topic distance map is a 2-D visualization of the topics. The area is proportional to the number of words that belong to each topic. We looked at each topic separately and collectively as a whole.

For C_V score (Figure. 1a), there is significant overlap among topics 1, 2, 3, 5, and 6. This implies that these topics have more tokens in common than with, say, topic 7. Despite the shared area, the distribution of words within these topics is different and each identifies a distinct theme in the dataset. Topics 4, 7, 8, 9 and 10 are disjoint with no overlap. Based on the top 10 words in each topic, we were able to assign a theme to 8 of 10 topics. To do so, we used the macroeconomic significance of the top 10 words in each topic. Topics 4 and 5 were junk topics with no clear identifiable economic themes. Table 2 summarizes our topic interpretation.

Table 2: C_V score topic interpretation

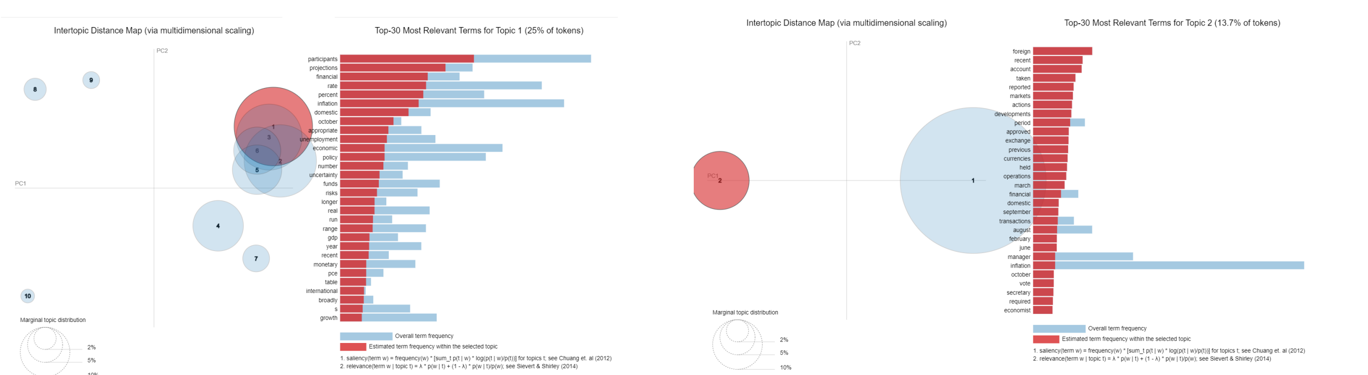
Topic #	1	2	3	6
Theme	Domestic Inflation Projection	Mid-year Labor Markets	Systems Open Market Account & Inflation	Mid-year FOMC Actions Review
Top 5 words (left to right)	participants, projections, financial, rate, percent	june, inflation, economic, labor, policy	soma, inflation, january, participants, rate	taken, actions, july, november, rate
Topic #	7	8	9	10
Theme	Elections & Discontinuances	Foreign Exchange and Accounts	Treasury & Securities Transactions	End-year FOMC Actions Review
Top 5 words (left to right)	connection, official, discontinuance, officers, cease	foreign, account, previous, exchange, currencies	november, september, securities, transactions, treasury	taken, actions, march, february, approved

For UMass score (Figure. 1b), the optimal number of topics chosen is 2. The topics are clearly demarcated and have no overlap. Furthermore, there is one clear topic that most tokens belong to (Topic 1). Table 3 summarizes our interpretation.

Table 3: UMass score topic interpretation

Topic #	1	2
Theme	Economic Policy & Growth	Foreign Exchange Markets
Top 5 words (left to right)	participants, inflation, rate, s, economic	foreign, recent, account, taken, reported

Figure 1: Intertopic distance map



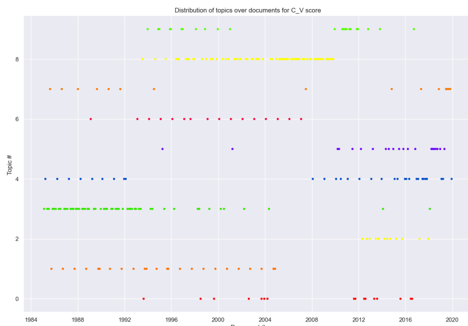
(a) for C_V score with tokens for topic 1

(b) for UMass score with tokens for topic 2

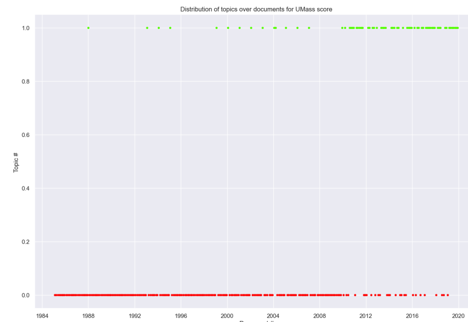
We also look at the unconditional distribution of each topic (code in Appendix C) and note that the disjoint topics have a higher frequency of being assigned probability 1 as compared to the topics that overlap i.e. the topic demarcation is more clear when the topics are disjoint.

Next, we looked at the topic assignment over time. As seen in Figure. 2a, for C_V score, there is no clear pattern for the topics assigned over time. We note that the topics are evenly and quite randomly distributed across the entire time series. The topics assigned by the UMass model (Figure. 2b), show a more significant pattern. From years 1985 to 2010, topic #1 tends to dominate. 2012 onwards, there is a fluctuating pattern, with topic #1 and #2 almost evenly distributed.

Figure 2: Topics assigned to FOMC meeting minutes over time



(a) for C_V score

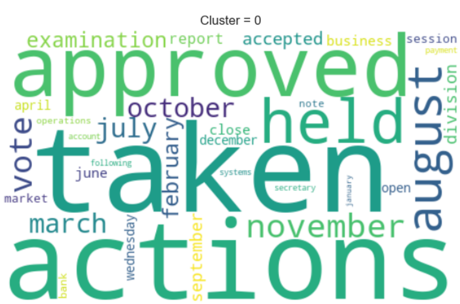


(b) for UMass score

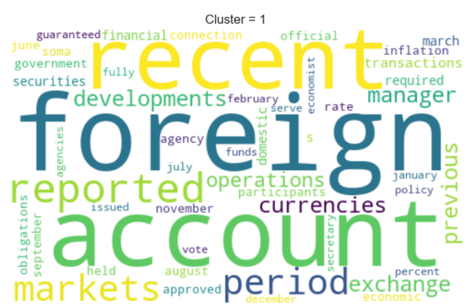
3.1.2 K-Means Clustering

We cluster the TF-IDF representation of the training data into 2 clusters to correspond to 2 regimes. The split of the 281 training set documents is 177 in cluster 0 and 104 in cluster 1. We also created word clouds to qualitatively understand the clusters (Figure. 3).

Figure 3: Top 50 words, word cloud



(a) for Cluster 0



(b) for Cluster 1

Interestingly, we observe that the word clouds for the clusters closely resemble the word distribution for the two topics chosen by the UMass topic model (Figure. 1b). However, the proportions of documents assigned to the clusters are not the same as the UMass score. The UMass score assigns significantly more documents to topic/cluster 0. Like the UMass score topics, cluster 0 corresponds to economic growth and cluster 1 corresponds to foreign exchange and markets.

3.2 Supervised Learning

We train the 6 classifiers and compare the performance of the regime-dependent trading strategy to the two control strategies on the test set. The results are given in Table 4.

From here, we can make a few observations:

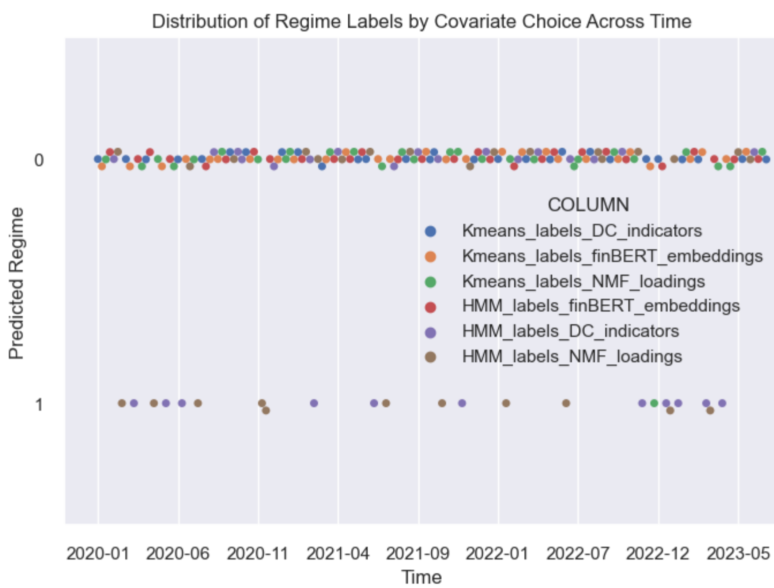
- For the hidden Markov model-produced labels, using text data as covariates instead of the price data (DC indicator) worsened the performance of the corresponding trading strategy in terms of Sharpe ratio and total profit. There was a slight improvement in the max drawdown, i.e., it was lower when using text data, however but it was not as significant (only 7% compared to a larger percentage drop in the other metrics). For the K-means-produced labels, the three strategies had almost equivalent performance. This shows that using text data as covariates for predicting regimes did not help in either case.

Table 4: Trading strategy metrics

Algorithm	Profit	Sharpe	MDD
Control 1 (mean-reverting)	0.531	0.541	0.154
Control 2 (momentum)	0.529	0.617	0.154
K-means labels, NMF loading covariates	0.531	0.540	0.154
K-means labels, FinBERT embeddings	0.531	0.541	0.154
K-means labels, DC covariates	0.531	0.541	0.154
HMM labels, NMF loading covariates	0.531	0.618	0.158
HMM labels, FinBERT covariates	0.531	0.541	0.154
HMM labels, DC covariates	0.902	0.710	0.223

- Between the K-means and HMM labels, for the same covariate, the models trained on the HMM labels either did equally well or better on all metrics except the max drawdown for DC covariates. This is strong evidence showing that using text data to identify regimes does a worse job than using HMMs in terms of the final trading strategy.
- Interestingly, we see that the the performance for the K-means–labelled models (and the HMM-labelled model with FinBERT covariates) is identical to the performance of the mean-reverting control strategy, which is agnostic to the regimes. We explore this further below.

Figure 4: Regime Labels by Covariate Choice Across Time



All the models that used K-means labels ended up only predicting the label 0 (with the exception of a singular data point). This explains why the performance on the mean-reverting control strategy (which is equivalent to the regime always being 0) is the same as the regime-dependent strategy.

4 Discussion

4.1 Unsupervised Learning

We make the following observations for unsupervised learning:

- In terms of topic interpretability and themes, C_V Score chooses topics that are easy to interpret and have distinct themes. Each topic covers an aspect of the macroeconomic reality at a given time. For instance, topic 1 covers the projection for inflation domestically – an important factor for FOMC interest rate decisions. Furthermore, the topic split also incorporates important information about the decisions themselves in a separate topic - topics 6 and 10.
- UMass score divides the dataset into two topics, neither of which conveys any distinctive information. The topics

are too generic, and each includes words that are contradictory in macroeconomic sentiment or do not convey any sentiment at all. For instance, the topic 1 in UMass has the suffix ‘s’ as the fourth-most frequently occurring word.

- Furthermore, we expect the regime changes to be frequent over time. The changing topics under C_V score validate that expectation. The UMass topics remain constant up to 2012. That is not supported by real-life market dynamics.

In conclusion, we believe that C_V score performs better. Therefore, we chose to use the topic probability distribution generated by the NMF model with 10 topics as covariates for the supervised learning problem.

K-means clustering splits the documents into 2 clusters. We use the cluster labels as an indication of the underlying regimes. Although we do not draw any independent conclusions from the clusters themselves, we evaluate them based on the performance of the supervised learning task in the next section.

4.2 Supervised Learning

We make the following conclusions:

- Using text data to identify regimes does not improve the existing trading strategy – neither from the regime-agnostic strategy nor from the regime-dependent strategy that uses price data to identify regimes. This is because we found that the classifier predicted the same label for all the points in the test set when trained on regimes identified using text data. Even when we used the same covariates as before, the text data-identified regimes worsened the trading strategy’s performance. This could be due to a couple of reasons. One possible reason is that we have very little data in a given year – there are only 8 meetings, and we are filling in the same label between two meeting dates. Two, it could be possible that the regimes, which we found to be related to economic growth and foreign exchange respectively (see Section 3.1.2), may not be that significant in the context of market behavior.
- For a given set of regime labels, using text data as covariates did not help improve the trading strategy from using price data as covariates. A major reason for this could be that by removing numeric data, the text data is not able to account for differences between, say, a 1% inflation and 6% inflation, which should correspond to very different regimes – our representations can only account for the word “inflation” being mentioned. Other possible reasons for this could be that the text data had a lower frequency than the price data and the values had to be forward-filled. For the NMF loadings, it could be the fact that the topics we found had significant overlap and a couple of topics were junk topics (see Section 3.1.1).
- Overall, our classifiers that used text data, either to generate the labels or as covariates, did a poor job at predicting the labels on the test set and ended up predicting the same label for the entire dataset.

4.3 Limitations and Future Research

- Our text data analysis focuses only on alphabetical data. However, in FOMC meeting minutes, we would expect to observe a significant amount of numerical data, and these would affect the interpretation of the text. For instance, we interpret “inflation increased by 6%” as more significant than “inflation increased by 2%”. A follow up study can utilize more sophisticated tokenization and modeling techniques, or they could extract the numerical data separately and treat them as additional covariates to the supervised learning problem.
- FOMC meetings occur only 8 times a year. Therefore, our NMF loadings and K-means-generated regime labels are based on input data with low frequency. Follow up studies can add text documents from regular financial news sources that can increase frequency of our data. Furthermore, we can augment the FOMC meeting minutes with numerical macroeconomic data from the FRED database e.g. unemployment rate, consumer index, and so on. This numeric data can provide a more holistic view of the drivers for regime changes.
- We use only the naive Bayes classifier to classify the regimes. However, NB assumes independence of covariates. This assumption may not be valid for NMF loadings (which are not necessarily orthogonal) or FinBERT embeddings. Follow up studies can expand the scope by using classifiers that account for covariate interactions such as random forest or deep learning classifiers.
- The regimes that we identified using K-means do not necessarily correspond to the same regimes that we detect using HMMs, so the comparisons that we make may not be apples to apples. A better idea might be to try different values of K until we find regimes that correspond to the characteristics shown by the HMM regimes, and then collapse the other clusters into a single “other” regime. Further, there are no “true” labels for regimes.

References

Ang, A., & Timmermann, A. (2011). *Regime changes and financial markets*. National Bureau of Economic Research.

Baid, D., Prasad, R., Vishnoi, S., & Sharma, U. (2023, April). Detecting market regime changes for trading. https://github.com/uday-31/detecting-market-regime-changes

Bickerton, E. (2023, August). Federal reserve fomc meetings text data. https://www.kaggle.com/datasets/edwardbickerton/fomc-text-data

Chen, J., & Tsang, E. P. (2021). *Detecting regime change in computational finance: Data science, machine learning and algorithmic trading* (First edition.). CRC Press, Taylor & Francis Group.

Fernando, J. (2022, February). Sharpe ratio definition. https://www.investopedia.com/terms/s/sharperatio.asp

Hayes, A. (2022, March). Maximum drawdown (mdd) definition. https://www.investopedia.com/terms/m/maximum-drawdown-mdd.asp

James, B. (2002). Globalization and the silent revolution of the 1980s. *Finance and Development*. https://www.imf.org/external/pubs/ft/fandd/2002/03/bought.htm

Jegadeesh, N., & Wu, D. (2015). *Deciphering fedspeak: The information content of FOMC meetings*. https://www.aeaweb.org/conference/2016/retrieve.php?pdfid=21466&tk=niAkBk3N

New York Stock Exhange. (n.d.). Nyse: Nyse market information. https://www.nyse.com/markets/nyse-arca/market-info

Petrov, V., Golub, A., & Olsen, R. (2019). Instantaneous volatility seasonality of high-frequency markets in directional-change intrinsic time. *Journal of Risk and Financial Management*, *12*(2). https://doi.org/10.3390/jrfm12020054

The Federal Reserve. (n.d.-a). The fed - federal open market committee. https://www.federalreserve.gov/monetarypolicy/fomc.htm

The Federal Reserve. (n.d.-b). Federal open market committee - calendar. https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm

The Federal Reserve. (n.d.-c). Fomc - historical data. https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm

Tsang, E., & Jun, C. (2018). Regime change detection using directional change indicators in the foreign exchange market to chart brexit. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*. https://doi.org/10.1109/TETCI.2017.2775235

Appendix A - Directional Change Indicators

Instead of only using the adjusted closing prices, we used both the opening and closing price and offset them by a factor of 12 hours on each date. Although the core trading session of the market is from 09:30 to 16:00 (New York Stock Exhange, n.d.), we chose to use a 12-hour offset so that the opening and closing prices were equispaced. This helped double the number of data points we had.

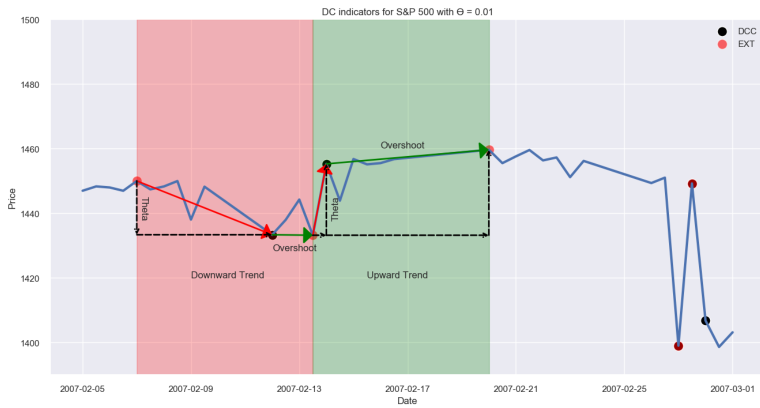


Figure 5: Market events under the DC framework.

The DC framework splits the market into 4 different events, as shown in Fig. 5. A DC event starts at an extreme point (P_{EXT}) and ends at a directional change confirmation point (P_{DCC}). The hyperparameter θ is the threshold percentage price change at which we can define a DCC event (see (1), from Chen and Tsang, 2021):

$$P_t = P_{\mathrm{DCC}} \iff \left| \frac{P_t - P_{\mathrm{EXT}}}{P_{\mathrm{EXT}}} \right| \geq \theta \tag{1}$$

We then transition to an overshoot event which continues until a reversal in trend happens. In the DC series, we sample the P_{EXT} and the P_{DCC} points from the entire time series.

We define the DC indicator Time-Adjusted Return (R) as per Chen and Tsang, 2021, which measures the change per unit time in the price between extreme events. This serves to quantify the directional changes observed in the market and is the input to our hidden Markov model:

$$R(n) = \frac{1}{t_{\mathrm{EXT}}(n) - t_{\mathrm{EXT}}(n-1)} \times \frac{P_{\mathrm{EXT}}(n) - P_{\mathrm{EXT}}(n-1)}{P_{\mathrm{EXT}}(n-1)} \tag{2}$$

For our project we used $\theta = 0.01$, which is close to half of the 1-day average daily volatility of the S&P 500. We wanted to choose a small enough threshold so that we have a lot more sample points.

Appendix B - Trading Strategies

We develop a novel way to develop our loss function for cross-validation. Extending the work by Chen and Tsang, 2021, we created three different trading strategies – two control strategies and a regime-dependent strategy. The first control strategy (CT1) is a basic mean-reverting strategy, while the second control strategy (CT2) is a basic momentum strategy. The regime-dependent strategy is mean-reverting during the low-volatility regimes and momentum-based during high-volatility regimes. We ensure that no look-ahead bias is present in any of the three strategies. These are described in detail below.

<table>
<tr><th colspan="2">Algorithms</th><th>Regime Dependent Strategy</th><th>Control Strategy – Mean Reversion (CT1)</th><th>Control Strategy – Momentum (CT2)</th></tr>
<tr><td rowspan="2">Normal regime (0) - low volatility</td><td>Open</td><td>Open contrarian position when |TMV| reaches 0.5</td><td>Open contrarian position when |TMV| reaches 0.5</td><td>Open trend following position when |TMV| reaches 0.5</td></tr>
<tr><td>Close</td><td>Close at next DCC or RCD</td><td>Close at next DCC</td><td>Close at next DCC</td></tr>
<tr><td rowspan="2">Abnormal regime (1) - high volatility</td><td>Open</td><td>Open trend following position when |TMV| reaches 0.5</td><td rowspan="2">Same as in normal regime</td><td rowspan="2">Same as in normal regime</td></tr>
<tr><td>Close</td><td>Close at next DCC or RCD</td></tr>
</table>

The three trading strategies which we implemented don't have a look-ahead bias which is one of the main problems which drive high profits. In the trading strategy, we define |**TMV**| at all points of the time series by calculating it from the last extreme point (P_{EXT}). We assume a frictionless market and no transaction costs during trading.

Control Strategies: We use two control strategies – one is momentum based, while the other is mean-reverting-based. In the momentum strategy, we take positions according to the market when |**TMV**| reaches a value of 0.5. We close this position at the next DCC event. In the mean reverting control strategy, we take a position opposite to the market according to the principle as defined above. None of the actions in the strategy depend upon the regime we are in or whether or not a regime change happens.

Regime-Dependent Strategy: In the regime-dependent strategy, we utilize the information about the market's current regime, which our two-class classifiers predict. We have different actions for the low-volatility and high-volatility regimes. We follow the same mean-reverting strategy as the control during the low vol times, but we close the position when a regime change is detected in addition to a DCC point. In the high-vol regime, we follow the market and trade on momentum. We open a trend-following position when the |**TMV**| reaches a value of 0.5. We close this position when a regime change is detected or when a DCC point is attained.

The following are the metrics used to evaluate the different strategies:

- PNL: This is the cumulative profit or loss generated by the trading strategy and is calculated by taking the cumulative sum of the daily returns. This assumes that on each day, a new position is entered and the previous position is closed, with a fixed notional of $1.
- Sharpe Ratio: This is used to measure the risk-adjusted return. Assuming a risk-free rate of zero, it is calculated by dividing the annualized return by the annualized volatility (Fernando, 2022).

$$\text{Sharpe Ratio} = \frac{r_p}{\sigma_p}$$

  where r_p and σ_p are the annualized return and standard deviation of the portfolio, respectively.

- Maximum Drawdown (MDD): Theoretically, MDD is the maximum observed loss from a peak to a trough of a portfolio before a new peak is attained (Hayes, 2022). It is calculated as shown in Equation 3. We use a proxy for this, which finds the minimum subarray of the returns using Kadane's algorithm. This is given by Equation 4.

$$\text{MDD} = \frac{\text{Trough Value} - \text{Peak Value}}{\text{Peak Value}} \tag{3}$$

$$\text{MDD} = \min_{0 \le i \le j \le N} \sum_{k=i}^{j} ret_k \tag{4}$$

Appendix C - Code

We collaborated using GitHub. The entire codebase can be found at this link. We have also attached a copy of the notebook below. For our project, we used code modules that we imported into the notebook. For the purposes of the submission, we have attached a copy of our notebook where we pasted all the module funtions into the beginning of the notebook for easy viewing.