Power of the News: Tactical Asset Allocation Strategies Derived Using Sentiment Analysis

by

James McIndoe (CID: 00643676)

Department of Mathematics
Imperial College London
London SW7 2AZ
United Kingdom

Thesis submitted as part of the requirements for the award of the MSc in Mathematics and Finance, Imperial College London, 2017-2018

Declaration

The work contained in this thesis is my own work unless otherwise stated.

Signature and date:

Acknowledgements

I would like to thank both of my supervisors, Dr Mikko Pakkanen and Mr Jason Tannen, for all the guidance they have given me throughout this project. I would also like to thank both of my parents for the constant support they have shown me, not only throughout the Masters programme at Imperial, but through all of my endeavours.

Contents

1 Introduction

1.1 Motivation and Objectives

Eugene Fama's Efficient Market Hypothesis [14], stems from the assumption that all market participants behave in a rational manner, and that all information is quickly and fully reflected in the prices of financial assets. The resulting theoretical market renders technical and fundamental analysis useless, since the market, under these conditions, is a fair game. Certain market phenomena, such as the Momentum Effect [20] or the January Effect hypothesis, seem to contradict Fama's assumptions, giving rise to Behavioural Finance as an alternative school of thought.

The last decade has given rise to an exponential growth in the number of news sources with an increasing amount reported in real time. This has made the task of processing publicly available information an expensive and time-consuming challenge. Developments in natural language processing and machine learning methods has enabled the sophisticated investor to process information in a more timely manner than humanly possible. This dislocation has the potential to reveal inefficiencies in the markets.

This thesis examines the effect of investor sentiment in the commodity futures market. Using sentiment embedded in news reports, we develop several systematic strategies, in an effort to demonstrate the predictive power of the news on investor decision making.

1.2 Commodity futures market participants

The commodity futures market participants can be split into two broad categories: Commercial and Financial. Commercial participants, more commonly referred to as hedgers, use the futures market to hedge their underlying exposure to commodity price risk. While commercial participants comprise of both commodity producers and consumers, the desire to hedge adverse price movements for producers exceeds that of the consumers. For this reason, commercial participants are net sellers of commodity futures. As an example, a wheat farmer would enter the futures market to fix a price today, for future delivery of their crop, thus reducing the risk that they bear during harvest. Meanwhile, on the other side of this trade is the financial participants, who can also be split into two groups, speculators and index investors. Index investors, or passive investors, hold commodity futures as part of their asset allocation strategy, utilised both as a means to diversify their return drivers, and as a hedge against inflation. By definition, the passive investors are long-only commodity futures investors.

Speculators, or active investors, enter the futures market due to their belief that they can profit by successfully anticipating future movements in commodity prices by acting on new information on supply and demand fundamentals. As of May 2008, Goldman Sachs estimated that speculators made up 51.1% and 41.5% of the Chicago Board of Exchange (CBOT) Corn, and New York

Merchantile Exchange (NYMEX) crude oil futures and options positions respectively [16].

2 Literature Review

2.1 Investor sentiment and market activity

There have been many papers that have examined the relationship between investor sentiment and its effect on financial markets. Chaoenrook [7] showed that readings from the University of Michigan Consumer Sentiment Index was positively related to excess market return, while Da et al. [10] used the Financial and Economic Attitudes Revealed by Search (FEARS) index, which measures the volume of search engine queries related to household concerns, to predict market volatility and cashflows from equity mutual funds to bond funds.

The last decade has seen a rapid growth in the number of platforms used for dissemination of information, with the division between news and opinion becoming more blurred. Textual information processing has enabled researchers to analyse the impact of market sentiment from a number of sources. Antweiler and Frank [2] utilised the message boards of Yahoo! Finance and Raging Bull to conclude that an increase in the volume of messages predicts a subsequent increase in the trading volume and volatility of a stock, while Chen, De, Hu and Hwang [9] found that views expressed in articles and commentaries posted on Seeking Alpha, a popular forum for active investors, were able to predict future stock returns and earnings surprises.

In a time when company CEOs, and even the President of the United States, use social media platforms to disseminate opinions in real time, it is not surprising that these sources have also been evaluated on their ability to make predictions. Bollen, Mao and Zeng [5] used sentiment from Twitter to predict the direction of the next days' price movements of the Dow Jones Industrial Average (DJIA), while Vu et al. [36] used the same medium to measure the effect of sentiment on single name tech stocks. When measuring the effect of market sentiment, there are two core areas of focus. The first is its effect on trading volume and volatility. Alanyali, Moat and Preis [1] identified a relationship between the daily number of mentions of a company in the Financial Times and the daily transaction volume of the company's stock, while Dzielinski and Hasseltoft [13] used aggregate level and aggregate dispersion of news tone to predict realised volatility and changes in the VIX index.

The second is whether news sentiment can predict subsequent movements in price, with a number of behavioural finance biases cited as potential explanations for the empirical evidence. Barber et al. [3] argue that individual investors are more likely to buy attention grabbing stocks than sell them due to restrictions on short selling and scarcity of attention, and that this manifestation of availability bias applies only to individual investors and not institutional investors. Meanwhile De Bondt and Thaler [12] suggest that investors tend to overweight recent information, in violation

of Bayes rule. That is, that investors exhibit representativeness bias in forming their decisions. On the other hand, Barberis et al. [4] present empirical evidence demonstrating that markets underreact to positive information over a period of 1-12 months thus providing evidence of status quo bias in financial markets, i.e. that markets are slow to react to new information. This is in line with the findings of Sinha [32] who concludes that markets take almost 13 weeks to fully incorporate information from a news article, and that this finding is not restricted to small stocks or low analyst-coverage stocks. In addition, the "underreaction portfolio" Sinha created in this research demonstrates a high correlation to the momentum benchmark, suggesting that slow absorption of news might explain some portion of the momentum effect. Hillert, Jacobs and Muller [19] also focus on the relationship between news sentiment and coverage and suggest that media coverage can exacerbate investor biases, lending credibility to the underreaction-based explanation for the momentum effect. These findings coincide with those of Scott, Stumpp and Xu [31], who conclude that "the momentum-volume interaction is explainable as a delayed reaction to news about company fundamentals".

2.2 Text mining for market prediction

The process of training a natural language processor for market predictions can generally be broken down into three steps:

1. Data sourcing: collecting both textual data and market data
2. Feature pre-processing: the process of converting an unstructured textual message to a machine readable input for the classification algorithm
3. Application of a machine learning algorithm to classify the features to the target variable

We compare the various methods used in each of these stages separately when reviewing the existing research in this field.

Data sourcing

The majority of research in this area has focused on published news articles as their source of textual information. Sources such as the Financial Times [37], Wall Street Journal [35] and Forbes.com [28] have all been used in an attempt to forecast stock returns. This is due to the common belief that this source of market sentiment has less noise compared with alternative sources such as Twitter ([5] & [36]) or online message boards [11]. Meanwhile several pieces of research used corporate documents as their source of sentiment. Butler and Kešelj [6] focused on annual reports, while Li [22] used 10-K and 10-Q filings from SEC Edgar Website. While several researchers used foreign exchange rates as the target for their classification ([21] & [8]), the majority of attention in this

area has been forecasting returns of single name equities and equity indices with a horizon of up to 1 day.

Feature pre-processing

Before the unstructured textual message is used, it must be structured such that it can be processed by the classification algorithm. The main stages of pre-processing textual data are feature-selection and dimensionality reduction.

Feature selection

By far the most common method for feature selection is the "bag-of-words" model, also known as the "vector space model", which simply represents a textual message as a vector of words and the frequency of these words within the text, considering each word frequency as a feature. The drawback of this representation is that the order and content of these words are completely ignored. Jin et al. [21] and Mahajan et al. [23] use the latent Dirichlet allocation (LDA) method, categorizing words into unobserved groups, which become the features to be used as inputs to the classification algorithm.

Dimensionality reduction

When utilising unstructured textual information for classification in real time, it is important to have a limited number of features, both in terms of accuracy of prediction and computational efficiency. A number of different methods are observed among existing research. One approach is to use a predefined dictionary of keywords, constructed by market experts [27]. Alternatively, the Harvard Psychosociological dictionary used by Tetlock et al. [35] focused on words within the field of psychology and sociology. An alternative method, implemented by Soni et al. [33], is to create the dictionary dynamically based on the training text corpus.

Machine learning algorithms

A number of different machine learning algorithms have been applied to this field of research. Groth and Muntermann [17] utilised a Naïve Bayes approach to the classification problem. Although this algorithm works on the unrealistic assumption of independence between text features, it has still been successful in the field of text classification. Another popular method for classification is Support Vector Machine (SVM), which attempts to partition the vector space by a separating hyperplane with a maximum margin. Both of these methods were employed by Antweiler and Frank [2] with similar accuracy in classification. Rachlin et al. [28] and Vu et al. [36] both used the C4.5 algorithm [30], developed by Ross Quinlan, to generate a decision tree, with the output being a set of trend predicting rules. The last approach covered in this section is that of an ensemble

methodology, that combined a number of algorithms together. This has been demonstrated by Das and Chen [11], who used an ensemble of 5 machine learning algorithms, stacked together in a voting system.

3 Data

Sentiment Data

The dataset used for this research was provided by the a third-party data provider. By utilising the database of news articles from Dow Jones News Wire, the third party were able to source approximately 13 million articles for their commodity dataset from 13 different sources. In contrast to the existing research covered in Section 2, when training their sentiment classification engine, they used a sample of 125,000 news articles that had been labelled as positive, neutral or negative sentiment by industry experts. The engine therefore attempts to replicate the response of financial analysts to a news article, and not its contemporaneous impact on market characteristics, such as volatility or intra-day returns. As a result, the classification process does not rely on pre-defined dictionaries or a rule based approach, but rather, it learns the relationships that exist between words from the training data. To achieve this, the engine uses a set of machine learning algorithms, combined in an ensemble learning framework to assess the emotional tone embedded in news articles. The output is a labelled dataset with a sentiment score for each news article. Details of the sentiment dataset features can be found in Appendix A. While the specific makeup of the ensemble is proprietary, the three primary algorithms used are Support Vector Machine (SVM), Maximum Entropy and Conditional Random Fields. While the sentiment classification process is beyond the scope of this thesis, an intuitive explanation of the primary algorithms can be found in Section 4.4. We use a subset of the entire dataset; a list of 10 commodities, who have commodity tags that related to a specific future and not grouped into categories. For example, a commodity tag relating to 'grains' would not be included in our commodities subset. The 5,742,362 news articles span a period from 1 January 2000 to 31 May 2018. In all strategies, we aggregate the commodity Net Sentiment Score at midnight on Sunday.

Definition 3.1. The commodity **Net Sentiment Score** at time t, over a look-back period (L), is given by

$$NetSentimentScore_{t,L} = \log\left(\frac{1+\sum_{i\in(t-L,\ t]} WeiPosNews_i}{1+\sum_{j\in(t-L,\ t]} WeiNegNews_j}\right),$$

where

$WeiPosNews_i = Confidence_i \times Sentiment_i . \mathbb{1}_{\{Sentiment_i>0\}}$

$WeiNegNews_i = Confidence_i \times Sentiment_i . \mathbb{1}_{\{Sentiment_i<0\}}.$

Commodity futures data

For our commodity futures return data, we use a generic commodity futures series, which is a continuous futures contract, constructed by splicing together successive futures contracts on the underlying commodity. There are two elements to the construction method: the date on which to roll successive contracts, and the adjustment made to the raw contract price.

The price adjustments are required to eliminate "jumps" in the continuous contract history, caused by discontinuities in the prices of successive underlying futures contracts. For the price adjustment rule, we used the "Backwards ratio method", which multiplies contracts by a constant factor so as to eliminate jumps, working backwards from the current contract. Since this method works back from the current contract, this method necessitates full historical recalculation on every roll date.

The roll date rule utilised is the "end-to-end roll method", which allows you to use the front contract for as long as possible. While this method can create rollover risk for a live trading strategy, it is sufficient for our purposes of backtesting. The return on the commodity future at time t, over holding period h is given by

$$r_{t,h} = \frac{p_{t+h} - p_t}{p_t},$$

where

$r_{t,h}$ = the return of the commodity future at time t over holding period h

p_t = the price of the generic commodity future at time t.

4 Methodology

4.1 Portfolio Construction

4.1.1 Equally Weighted by Capital

When constructing the portfolio from our trading signals, we utilise two different approaches to capital allocation. In the first method for portfolio construction, the positions in the long (short) portfolio were equally weighted by capital, with the long and short portfolio exposures each set at 100%.

4.1.2 Risk adjusted Allocation

The second approach accounts for varying levels of volatility in each of the commodity futures return series. When using the risk adjusted allocation method, we consider the total variance in the long and short portfolios separately using a 3-year rolling window. The position weights for each of the holdings in the long (short) portfolios are given by

$$W_i = \frac{1/\sigma_i}{\sum_j 1/\sigma_j},$$

where

W_i = allocation weight of commodity i in the long (short) portfolio

σ_i = annualised volatility of commodity i in the long (short) portfolio.

By constructing the portfolio holdings using this method ensures an equal contribution to risk in the long (short) portfolios.

4.2 Signal generation

4.2.1 Ranked Sentiment Strategy

When constructing the Ranked Sentiment strategy, we consider the universe of commodities simultaneously. At time t, the aggregated Net Sentiment Score for each commodity is calculated and ranked. The buy (sell) signal is then calculated, such that portfolio went long (short) the top (bottom) n commodities by rank, with n to be determined in the model calibration process. The holding period for all positions is 1-week. A consequence of this construction is that the strategy is net neutral with respect to the commodity universe we are considering.

4.2.2 Extreme Sentiment Strategy

The Extreme Sentiment strategy considers each commodity independently when generating trading signals. In doing so, the best fit distributions for the Net Sentiment Score are first inferred, see Best fit distribtution below. The chosen distributions are then fit on a rolling window, refit for each trade decision. The buy (sell) thresholds for the strategy are then set to be selected p-values of the fitted distribution, with the threshold level to be determined in the model calibration process. We optimise the Extreme Sentiment strategy on a single commodity, crude oil, and then look to utilise one of the portfolio construction methods to build an entire portfolio of the 10 commodity universe we are considering. Since the strategy puts no constraints on the number of commodities in the long (short) portfolio, the portfolio has the ability to be long (short) during any holding period, with the exposure restricted to $\pm 100\%$.

Best fit distribution

To determine the underlying distribution of the Net Sentiment Score, we attempt to fit each of the distributions available in the `scipy.stats` package in python, to a number of sample periods in the training dataset. The goodness of fit is then measured according to the Sum of Squared Errors (SSE). The distributions are then ranked according to their goodness of fit. We use 1, 2 and 3 year discrete periods, as well as the entire training dataset, to look for stability in the choice

of distribution. The top 3 distributions by average rank were chosen to be tested in the model optimisation stage, against the Null hypothesis that the Net Sentiment Score is best modelled by the normal distribution.

4.2.3 Regression derived strategy

The third strategy employed uses a variety of statistical and machine learning methods, in an attempt to fit a model on a feature matrix, derived from the sentiment dataset. For both regression models, we set the commodity futures return series as the target variable. Details of the models used in the regression, as well as methods for model validation and dimensionality reduction of the feature matrix an be found in Section 4.3. Once the model is fit, the portfolio takes a long (short) position in the commodity future if the model predicts a positive (negative) return.

Construction of the feature matrix

When constructing the Ranked Sentiment and Extreme Sentiment strategies, we consider only the value of the Net Sentiment Score, and make no consideration for the change in sentiment over time. In an effort to capture this in our model, when constructing the feature matrix for our regression model, we include the absolute value, as well as the change from previous period (δ) and the change in delta (γ). We use three aggregation periods (windows), aggregating the net sentiment over 3 days, 5 days, 7 days. We also make use of the 'Events' tag in the news, calculating the above metrics for 'All News', 'Shipping News', 'Derivative Pricing News', 'Production News' and 'Inventory News'. Finally, when calculating the δ and γ metrics we calculate them both as the change from the previous window and from the previous week. As an example, the '3 day delta window' calculates the difference between Friday-to-Sunday aggregation and Tueday-to-Thursday aggregation, while the '3 day delta week' looks at the change between the current Friday-to-Sunday aggregation and the previous Friday-to-Sunday. As a result the 7 day window only has a 'week' feature as these different delta methods would result in the same feature. We therefore have 65 features included in the matrix. Due to the large number of features in our matrix, Principal Components Analysis (PCA) is used for dimensionality reduction. We use the Principal components matrix with 80% variation retained (PCA80), as well as PCA90 and PCA95 as the feature matrices for the regression.

4.3 Statistical Methods

Principal Components Analysis

When creating the feature matrix for our regression model, factors were selected according to our intuition of the drivers of predictive power for the dataset. While some of the factors may

indeed explain some variability of the futures returns, others may introduce unwanted noise into the matrix. In addition, since all the features are derived from the underlying sentiment data, we can assume there is a large amount of linear dependence between some of the factors. Thus before carrying out the regression, we attempt to reduce the dimensionality of the feature matrix using Principal Components Analysis (PCA). We formalise this process in the following paragraphs, drawn from the lecture notes of Angrew Ng [26].

We write our feature matrix $X \in \mathbb{R}^{m \times n}$ as $\{x^{(i)}; i = 1, ..., m\}$, such that each $x^{(i)} \in \mathbb{R}^n$ is an observation of our feature matrix. Considering each observation (row) of our matrix as a point in $\mathbb{R}^n$ the PCA algorithm extracts the direction (vector), along which the data has the most variability. Before running the PCA algorithm, it is important that each of our data points have a mean of zero and unit variance, which is achieved through normalisation of the sample data. The process for this is:

1. calculate $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$
2. replace $x^{(i)}$ with $x^{(i)} - \mu$
3. calculate $\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)})^2$
4. replace each $x_j^{(i)}$ with $x_j^{(i)} / \sigma_j$

Now that we have normalised the data, we can apply the PCA algorithm to find the direction of largest variance. That is, we choose a unit-length vector u so as to maximise:

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)^T} u \right)^2 &= \frac{1}{m} \sum_{i=1}^{m} u^T x^{(i)} x^{(i)^T} u \\ &= u^T \left(\frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)^T} \right) u \end{aligned}$$

Thus, maximising this, gives the principal eigenvector of $\Sigma = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)^T}$, the empirical covariance matrix of our data. We can iterate this procedure to project our m-dimensional data onto a k-dimensional subspace, where the $\{u_i; i = 1, ..., k\}$ are our top k eigenvectors of Σ, and form a new, orthogonal basis for the subspace. Thus we can represent $x^{(i)}$ in this basis by computing:

$$z^{(i)} = \begin{bmatrix} u_1^T x^{(i)} \\ u_2^T x^{(i)} \\ \vdots \\ u_k^T x^{(i)} \end{bmatrix} \in \mathbb{R}^k$$

Thus, we have reduced our observation $x^{(i)} \in \mathbb{R}^n$ into a lower, k-dimensional, approximation $z^{(i)} \in \mathbb{R}^k$ using the first k-principal components of the data.

4.3.1 Regression

Linear regression [18] is used in construction of the final trading strategy to predict the return of the commodity future, using the feature matrix as an input. The general form for the multivariate regression model, in matrix notation, is:

$$Y = X\theta + \epsilon,$$

where $X \in \mathbb{R}^{m\times n}$ is the feature matrix, derived from sentiment data, $Y \in \mathbb{R}^m$ is the return of the commodity future, $\theta \in \mathbb{R}^n$ are the coefficients of the regression model, and $\epsilon \in \mathbb{R}^m$ is the error term of the fit, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The accuracy of model fit is measured using the cost function

$$J(\theta) = \frac{1}{2}\|Y - X\theta\|_2^2.$$

Thus, we can calibrate θ by minimising the cost function, that is, solving

$$\min_\theta J(\theta) = \min_\theta \frac{1}{2}\|Y - X\theta\|_2^2.$$

One of the main limitations of ordinary least squares regression is that, when the dimension of the feature matrix is large, this often leads to overfitting and results in an unstable model, a model with high variability. For this reason, we turn to Ridge regression and LASSO regression.

Ridge

Ridge regression attempts to reduce the variance of the fitted model by introducing an L_2-penalty on the parameter θ, at the expense of an increase in bias. The cost function becomes

$$J(\theta, \lambda) = \frac{1}{2}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2.$$

The parameter λ controls the Variance-Bias tradeoff and needs to be optimised. This is achieved through k-fold cross validation. The L_2-penalty shrinks the coefficients compared to regular regression.

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression makes use of an L_1-penalty in the cost function. The resulting algorithm performs feature selection, which is particularly important when the feature matrix is very large with potentially highly correlated features. The cost function for LASSO regression is

$$J(\theta, \lambda) = \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_1.$$

The observation that our feature matrix has highly correlated features, guides are intuition that Lasso regression should perform better, due to its feature selection characteristic.

4.3.2 K-fold cross-validation

When training our regression model, we need to evaluate how well the fitted model estimates the target parameter, the commodity futures return. K-fold cross validation provides a data efficient way to train and validate our model, by using the entire in-sample data, S, for both stages of the process. Details of the k-fold algorithm are presented below, drawn from the notes of Andrew Ng [25].

Suppose we want to select among a finite set of models $\mathcal{M} = \{M_1, ..., M_d\}$ the model that most accurately predicts our target variable. The models $M_j \in \mathcal{M}$ can be the same regression model with a different choice of penalisation parameter λ, or different models, i.e. LASSO and Ridge. The process for choosing the optimal model and optimal value of the hyper-parameter is as follows:

1. Split the training data S into k disjoint subsets, each with m/k training examples. We label these subsets $S_1, ..., S_k$
2. For each model M_i, we evaluate as follows:
   - For $j = 1, ...k$
     Train model M_i on $S_1 \cup ...S_{j-1} \cup S_{j+1} \cup ...S_k$
     (i.e train model on all the data except S_j)
     to get some hypothesis h_{ij}
     (i.e. model i trained on all subsets except S_j)
     Test the hypothesis h_{ij} on S_j, to get $\hat{\epsilon}_{h_{ij}}$
   - Calculate the estimated generalisation error of model M_i by calculating the average over k error readings, i.e. $\hat{\epsilon}_{h_i} = \frac{1}{k} \sum_{j=1}^{k} \hat{\epsilon}_{h_{ij}}$
3. Choose the model M_i with the lowest estimated generalisation error, and retrain that model on the entire training set S. The resulting output is our final model.

4.4 Sentiment classification algorithms

Support Vector Machine

The Support Vector Machine is considered by many to be among the best off-the-shelf supervised learning algorithm. The process of classification using the Support Vector Machine is that of partitioning a vector space containing the vector representation of the input features such that it separates those training examples with a positive label, from those with a negative label. Details of this process are explained below, drawn from the lecture notes of Andrew Ng [24]. Formally, consider a training set $\{(x^{(i)}, y^{(i)}) : i = 1, ..., m\}$ where $x^{(i)} \in \mathbb{R}^n$ is an observation of our feature matrix and $y^{(i)} \in \{-1, 1\}$ is a binary label of the data. The classifier is then expressed as

$$h_{w,b}(x) = g(w^T x + b),$$

where $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ otherwise. Thus, given a training example $(x^{(i)}, y^{(i)})$, the functional margin of (w, b) with respect to the training example is defined to be

$$\hat{\gamma}^{(i)} := y^{(i)}(w^T x^{(i)} + b).$$

Thus, if $y^{(i)} = 1$, then, for the functional margin to be positive and large (i.e. for our classification to be correct and confident), we need $w^T x^{(i)} + b$ to be large and positive. Conversely, if $y^{(i)} = -1$, we need $w^T x^{(i)} + b$ to be large and negative. The dependence on magnitude of $w^T x^{(i)} + b$ requires a normalisation constraint to be put on the vectors w and b. Replacing (w, b) with $(w/\|w\|_2, b/\|w\|_2)$ restricts the freedom of scale, ensuring that the magnitude of our functional margin is meaningful. We therefore define the geometric margin of (w, b) with respect to a training example $(x^{(i)}, y^{(i)})$ to be

$$\hat{\gamma}^{(i)} := y^{(i)} \left(\left(\frac{w}{\|w\|} \right)^T x^{(i)} + \frac{b}{\|w\|} \right).$$

Then, given our training set $S = \{ (x^{(i)}, y^{(i)}) \; ; \; i = 1, ..., m\}$, we define the geometric margin with respect to S to be the smallest of the geometric margins, i.e.

$$\gamma := \min_{i=1,...,m} \gamma^{(i)}.$$

The hyperplane that maximally separates the positive and negative labels is given by the solution to the following optimisation problem:

$$\begin{aligned} \max_{\gamma, w, b} \quad & \gamma \\ \text{subject to} \quad & y^{(i)} \left(w^T x^{(i)} + b \right) \geq \gamma, \; i = 1, ..., m \\ & \|w\| = 1. \end{aligned}$$

Maximum Entropy

The Maximum Entropy classifier is a probabilistic classifier that learns the joint probability distribution from the labeled training dataset. In doing so, the model does not assume anything about the distribution, other than what is observed in the training data. The Entropy of a probability distribution is the measure of uncertainty, and is maximised when the distribution is as uniform as possible. The parameters of the joint distribution are then inferred using maximum likelihood parameter inference, using an iterative method such as a gradient-based method. Ratnaparkhi [29] demonstrates the application of the Maximum Entropy classifier for natural language processing.

Conditional Random Fields

When using the Conditional Random Fields classifier, the focus is modelling the conditional distribution $p(y|x)$ where y is the classification of the training example, and x is the input features

for prediction. Therefore, dependencies that involve only variables in x, play no role in the conditional model and thus, a conditional model can have a much simpler structure than a joint model, such as the Maximum entropy classifier. For a thorough description of the modelling process for Conditional Random Fields, see the works of Charles Sutton and Andrew McCallum [34].

5 Model calibration

Descriptive statistics

To be able to compare the cumulative returns over different periods, we calculate the annualised return ($r_{An.}$) and annualised volatility ($\sigma_{An.}$), given by

$$r_{An.} = \prod_{i=1}^{n}(1 + r_i)^{52/n},$$

where

r_i = return in holding week i, and

$$\sigma_{An.} = \sigma_{W.} \times \sqrt{52},$$

where

$\sigma_{W.}$ = standard deviation of our weekly return series.

When evaluating the return profile of our trading strategies, we make use of a number of metrics, commonly used in finance.

Definition 5.1. The **beta** of our portfolio with respect to the Buy & Hold benchmark ($\beta_{B\&H}$) is defined to be

$$\beta_{B\&H} := \frac{\text{Cov}[r_p, r_{B\&H}]}{\text{Var}[r_{B\&H}]},$$

where

r_p = the weekly return series of our portfolio

$r_{B\&H}$ = the weekly return series of our Buy and Hold benchmark.

Definition 5.2. The **correlation coefficient** of our portfolio with respect to the Momentum benchmark ($\rho_{Mom.}$) is defined to be

$$\rho_{Mom.} := \frac{\text{Cov}[r_p, r_{Mom.}]}{\sigma_{r_p}\sigma_{r_{Mom.}}},$$

where

r_p = the weekly return series of our portfolio

$r_{Mom.}$ = the weekly return series of our Momentum benchmark

σ_{r_p} = the volatility of the weekly return series of our portfolio

$\sigma_{r_{Mom.}}$ = the volatility of the weekly return series of our Momentum benchmark.

Definition 5.3. The **Sharpe Ratio** (SR) is defined to be the average return in excess of the risk free rate, per unit of volatility. That is,

$$SR := \frac{r_p - r_f}{\sigma_p},$$

where

r_p = the annualised return of the portfolio

r_f = the annualised risk-free rate of return

σ_p = the annualised volatility of the portfolio return series.

The main drawback of the Sharpe Ratio is that, in order for the annualised volatility to be an accurate measure of risk, the distribution of portfolio returns must be normally distributed. When the portfolio returns do not follow the Gaussian distribution, the Return over Max Drawdown ($RoMaD$) is often preferred as a measure of risk adjusted return.

Definition 5.4. The **Max Drawdown** (MDD) of a portfolio is defined to be the maximum loss from a peak to a trough of the cumulative return series, stated in percentage terms, that is

$$MDD := \frac{\text{Peak Value} - \text{Trough Value}}{\text{Peak Value}}.$$

Definition 5.5. The **Return over Max Drawdown** ($RoMaD$) is defined to be the annualised return of the portfolio divided by the Max Drawdown,

$$RoMaD := \frac{r_{An.}}{MDD}.$$

Definition 5.6. The **Hit Ratio** of a strategy is the total proportion of trading decisions that were correct, calculated at the portfolio level,

$$HitRatio := \frac{\sum_{i=1}^{n} \mathbb{1}_{\{r_i > 0\}}}{n},$$

where

r_i = return in holding week i.

When calibrating the Extreme Sentiment strategy, we also look at the proportion of decisions that our portfolio took long (%L), neutral (%N), and short (%S) positions.

Method for comparing return profiles

When calibrating the models, we consider a number of choices for the hyper-parameters of the strategy, and compare the in-sample return profile for various choices of parameter. To capture the risk-return profile of the strategy, we consider the Return over Max Drawdown ($RoMaD$). This measure has been chosen in preference of the Sharpe ratio to capture the fatter tails in the portfolio return distribution.

The in-sample period for the model calibration is 01 January 2005 to 31 December 2014. Once an optimised model for each strategy has been selected, its out-of-sample performance will be analysed on the period from 01 January 2015 to 31 May 2018.

Buy and Hold benchmark

The Buy and Hold benchmark is representative of a passive investment in a portfolio of the 10 commodities, with weekly rebalancing. In order to maintain a like-for-like comparison, when evaluating a strategy that utilises a specific portfolio construction method, we allocate to positions in the Buy and Hold portfolio using the same methodology.

Momentum benchmark

Intuitively, the momentum signal is that of buying the previous winners and selling the losers. Thus, the momentum 'buy' signal is the commodity whose future had the greatest return over the previous period. In order to maintain a like-for-like comparison, the momentum portfolio uses the hyper-parameters and capital allocation method as that of the sentiment strategy that we are considering when constructing the portfolio. For example, if the Ranked Sentiment strategy has 3 commodities in the long (short) portfolio, equally weighted by capital, the momentum benchmark will hold the same number of commodities and utilise the same portfolio construction method, with the previous weeks futures returns as its signal.

Trading frictions

In reality, there are a number of trading frictions that would dampen the performance of our strategy. We have incorporated an execution lag of 1-day to reflect the time taken to enter a position. In addition, trading costs such as spread and commission would also dampen returns. Due to the complexity of accurately modelling the nuances that affect observed transaction costs, which are beyond the scope of this thesis, I have included a 4 basis point(bp) transaction cost, weighted by notional, for entering or exiting any position. That is, for example, the cost of selling the entire long portfolio is 4bps. All backtests in this research include both execution lag and transaction costs, unless stated otherwise. Both of these frictions limit the ability of the portfolio to run short term strategies, i.e. intra-day strategies.

5.1 Ranked sentiment strategy

When optimising the Ranked Sentiment strategy, there are a number of questions that need to be considered:

1. What is the optimal number of commodities to go long (short) in each holding period?

2. What is the optimal aggregation window to calculate the Net Sentiment Score over?
3. Should the positions be allocated on an equal capital or risk adjusted basis?
4. Should we include sentiment from repeated articles, or consider only novel news pieces in our aggregation?

We consider each of these questions in series, carrying through the results from the previous optimisation stages. The strategy variations are compared to the Buy and Hold and Momentum benchmarks to better understand the drivers of returns.

Diversification vs conviction

The first parameter to be optimised is the number of commodities to be held during each holding period. While a portfolio with more assets will benefit from a higher level of diversification in its returns, the level of conviction in the decisions to go long (short) will be lower. Intuitively, there is a balance to be struck between the incremental benefit of adding holdings to the portfolio at lower levels of conviction.

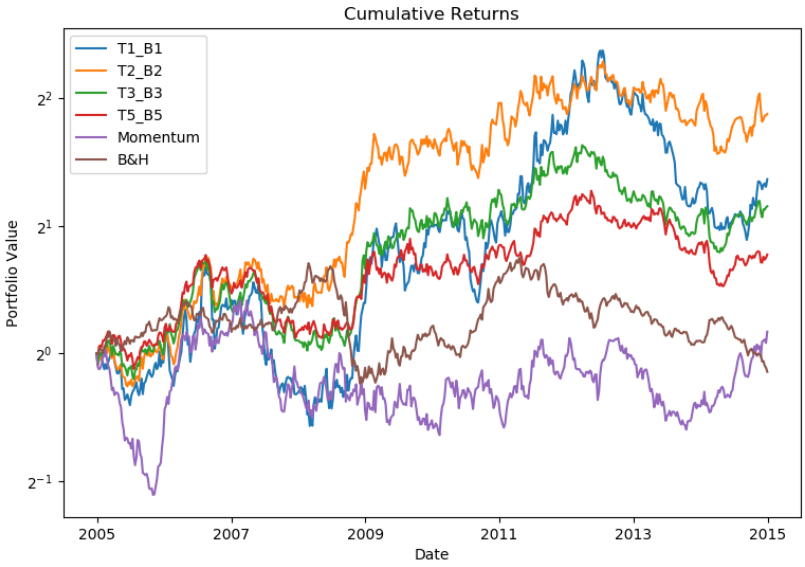

Figure 1: Diversification vs Conviction: number of commodities in long (short) portfolio

With reference to the $RoMaD$ metric in Table 1, the Top 2-Bottom 2 (T2-B2) portfolio has the best return profile. It is also worth noting that the beta coefficient to the Buy and Hold portfolio is close to zero, indicating that our Ranked Sentiment strategy has negligible market exposure with respect to the commodity universe that we are considering.

Metric	T1-B1	T2-B2	T3-B3	T5-B5	Mom.	B&H
$\beta_{B\&H}$	0.02	-0.03	-0.08	-0.05	-0.06	1
$\rho_{Mom.}$	0.18	0.17	0.17	0.15	1	-0.04
$r_{An.}(\%)$	9.94	13.92	8.33	5.53	1.19	-0.99
$\sigma_{An.}(\%)$	37.18	28.51	24.36	20.15	31.44	17.80
SR	0.27	0.49	0.34	0.27	0.04	-0.06
MDD (%)	64.27	39.36	44.18	40.37	54.52	47.96
$RoMaD$	0.15	0.35	0.19	0.14	0.02	-0.02
$HitRatio$ (%)	52.11	52.30	52.11	53.45	51.15	52.30

Table 1: Diversification vs Conviction: number of commodities in long (short) portfolio

Alpha decay of sentiment signal

The next consideration for us is the length of aggregation window to use in the Net Sentiment Score. To answer this question, we considered 4 aggregation windows, spanning from 1 week to 1 month, and consider the effect on $RoMaD$ gross and net of transaction costs.

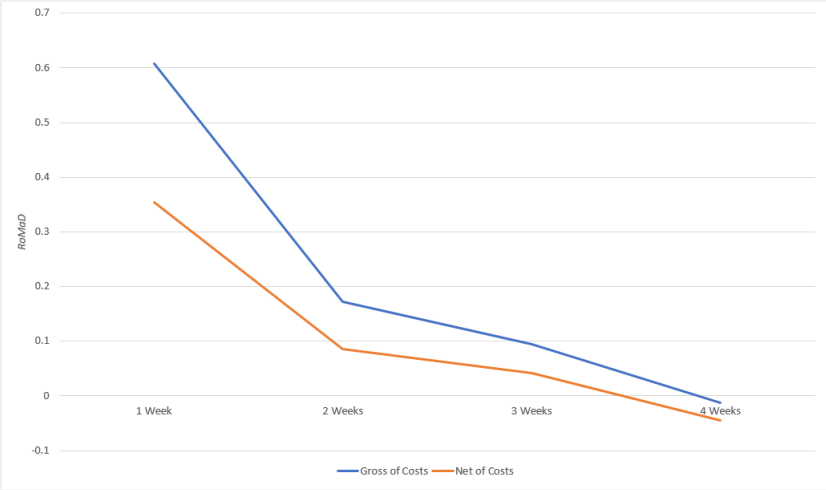

Figure 2: Alpha decay of sentiment signal - measured using the $RoMaD$ metric

Figure 2 clearly shows that the strength of the sentiment signal as a predictor of commodity future returns decays as the aggregation window is lengthened, with the sharpest rate of decay at the shortest end of the curve. This is a clear indication that, of the window lengths tested, 1-week is the optimal choice. In addition, we can see that the effect of transaction costs on our chosen measure of risk-adjusted return is lower, indicating that, as we increase the aggregation window, the portfolio turnover is lower.

Equally weighted by capital vs risk adjusted

When constructing the portfolio holdings, we need to decide whether to allocate to positions on a capital weighted or risk-adjusted basis. We consider the T2-B2 strategy with a 1 week aggregation period, as indicated by the previous sections.

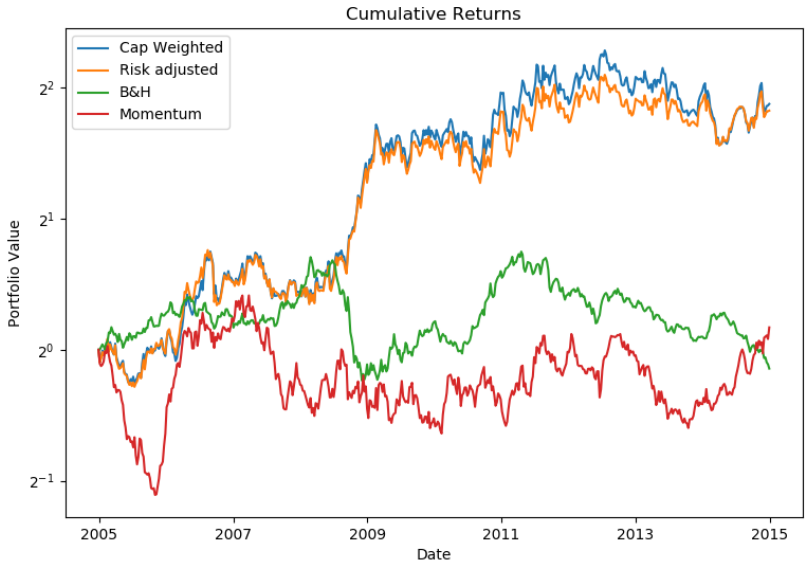

Figure 3: Equally weighted by capital vs risk adjusted allocation - cumulative return graph

Metric	**Cap. Weighted**	**Risk adjusted**	**Momentum**	**B&H**
$\beta_{B\&H}$	-0.03	-0.06	-0.06	1
$\rho_{Mom.}$	0.17	0.18	1	-0.04
$r_{An.}(\%)$	13.92	13.51	1.19	-0.99
$\sigma_{An.}(\%)$	28.51	27.56	31.44	17.80
SR	0.49	0.49	0.04	-0.06
MDD (%)	39.36	31.24	54.52	47.96
$RoMaD$	0.35	0.43	0.02	-0.02
$HitRatio$ (%)	52.30	51.34	51.15	52.30

Table 2: Equally weighted by capital vs risk adjusted positions

With reference to Table 2, we have that the Capital Weighted portfolio achieves a higher annualised return, outperforming the Risk adjusted portfolio by 41bps per annum. This comes at the cost of slightly higher volatility in returns, but more importantly for our optimisation process, with a higher realised Maximum drawdown. Comparing the $RoMaD$ metric of the two methods, we conclude that the risk adjusted portfolio has the preferred return profile.

Novelty vs Sentiment momentum

The final consideration to make is whether news articles that are part of a chain should be given the same consideration as the original news article. That is, should we exclude any sentiment contribution from an article whose content is a repetition of a previous article. This is part of a larger question of whether the sentiment effect is due to new news being priced into the market, or if the momentum of aggregated sentiment in the market can cause the price to drift. We test the T2-B2, risk adjusted portfolio, with a 1-week aggregation window, but excluding any news articles that has a novelty score bigger than 1; that is, if the story has already been received by the sentiment classification engine within the last 24 hours.

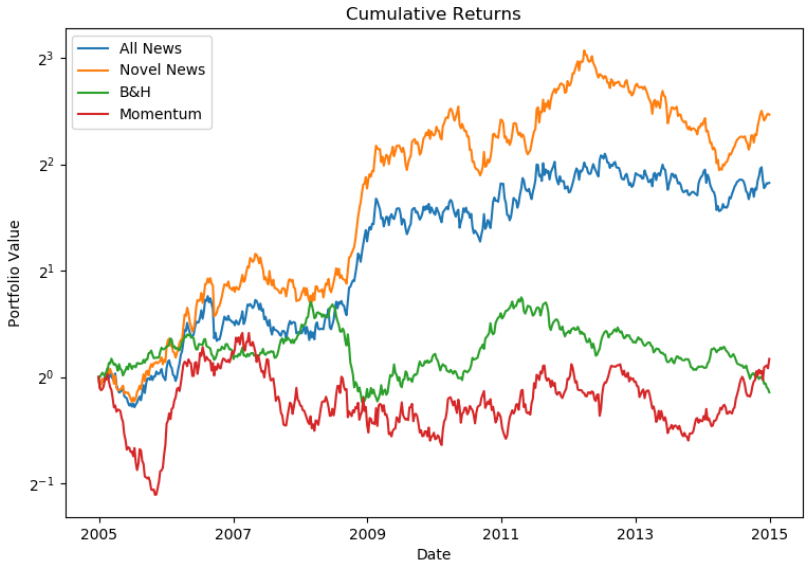

Figure 4: Comparison of returns with repeated news reports removed

Metric	**T2-B2**	**T2-B2 Novel**	**Momentum**	**B&H**
$\beta_{B\&H}$	-0.06	-0.13	-0.06	1
$\rho_{Mom.}$	0.18	0.22	1	-0.04
$r_{An.}(\%)$	13.51	18.68	1.19	-0.99
$\sigma_{An.}(\%)$	27.56	27.96	31.44	17.80
SR	0.49	0.67	0.04	-0.06
MDD (%)	31.24	54.18	54.52	47.96
$RoMaD$	0.43	0.34	0.02	-0.02
$HitRatio$ (%)	51.34	54.02	51.15	52.30

Table 3: statistics of including novel news only

Looking at Table 3 we can see that, by removing the repeated news reports, there has been an improvement in both the annualised return and the hit ratio of the strategy. This improvement, however, comes at the cost of a dramatic increase in the maximum drawdown. When comparing the strategy using our chosen metric, we conclude that, by including all news reports, not just novel reports, the risk-return profile of our strategy improves.

Optimal ranking strategy

To summarise the findings in this section, through a series of backtests, we have concluded that:

1. The optimal balance between diversification and conviction in positions is two positions in the long and the short portfolios
2. Using a 1-week aggregation window provides the strongest sentiment signal
3. Allocating to positions on a risk adjusted basis generates a superior return profile compared to positions that are equally weighted by capital
4. Retaining all news reports results in a strategy with a superior return profile, compared to using just novel news reports

The return statistics of the Ranked Sentiment strategy, both in-sample and out-of-sample, will be analysed in greater detail in Section 6.

5.2 Extreme sentiment strategy

When calibrating the Extreme Sentiment strategy, we considered the strategy on a single commodity, crude oil. Once the model has been optimised for the hyper-parameters in question, a total portfolio strategy is constructed using the model on the 10 commodities independently, with positions allocated on a risk-adjusted basis with respect to the long and short portfolios separately.

When choosing the hyper-parameters for the single asset strategy, there are a number of considerations:

1. What is the underlying distribution of the net sentiment score?
2. What is the optimal rolling window length to fit a distribution to?
3. What percentile thresholds shall we set for our buy (sell) signals?

We consider the first two questions simultaneously, before moving onto the question of signal thresholds.

Best fit distribution and Length of window

Having selected a list of 3 distributions that most closely fit the sample periods, we look to test their performance for providing signal thresholds against the Gaussian and Student's t-distribution. The

five distributions are tested using a rolling 1, 3 and 5 year window to calibrate the fitted distribution which is re-fit after each trading decision. For this test, we have selected signal thresholds of 80% (20%) for our buy (sell) signals. We include the results for the 1-year window below, with the sub-optimal results in the Appendix for completeness.

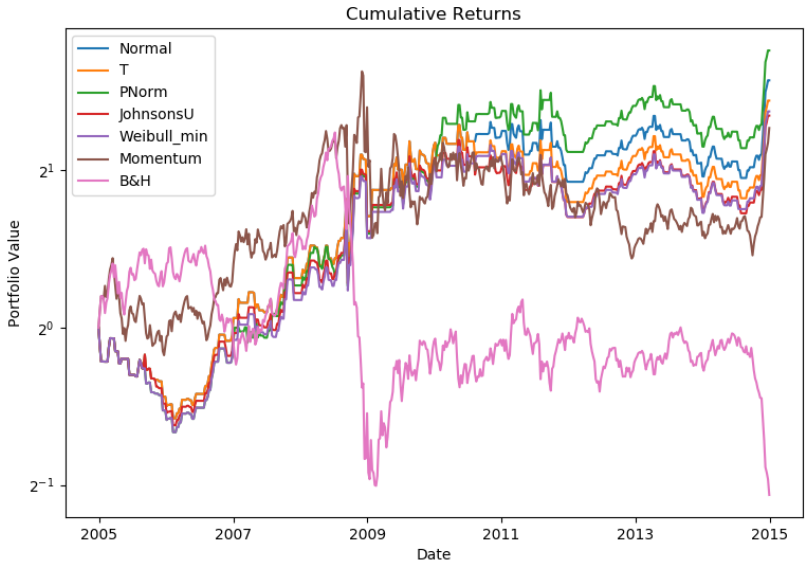

Figure 5: Fitted distribution - 1 year rolling window

Before considering our chosen metric for comparison, it is worth looking at the proportion of decisions that were long and short in our strategies. Comparing the figures in Table 4 with Tables 10 and 11, we can see that, as the window length shortens, our proportions tend towards our p-value thresholds, hinting that the 1-year fitted distribution is the optimal choice for our rolling window. In addition, the 1-year fitted distribution generates a superior risk-return profile, compared to the other two window lengths tested. We therefore conclude that the 1-year window is optimal for our distribution fitting. Turning our attention to the choice of distribution, we can see that, for both the 1 and 5 year windows, modelling the Net Sentiment Score with a Gaussian distribution yields the best in-sample return profile, with the 1-year fitted Gaussian distribution generating the best in-sample $RoMaD$ metric.

Metric	**Normal**	**T**	**Power Normal**	**Johnsons U**	**Weibull min**	**Mom.**	**B&H**
$\beta_{B\&H}$	-0.17	-0.16	-0.21	-0.16	-0.16	0.08	1
ρ_{Mom}	0.30	0.30	0.28	0.30	0.30	1	0.08
$r_{An.}$	11.52	10.53	12.99	9.80	10.00	9.20	-7.11
$\sigma_{An.}(\%)$	24.51	24.64	24.61	24.69	24.40	34.65	34.73
SR	0.47	0.43	0.53	0.40	0.41	0.27	-0.21
MDD (%)	29.72	29.72	33.86	31.82	33.86	56.13	79.72
$RoMaD$	0.39	0.35	0.38	0.31	0.3	0.16	-0.09
$HitRatio$ (%)	57.49	57.03	57.98	56.57	56.67	54.79	52.30
% L	25.67	26.05	23.37	26.25	24.52	52.49	100
% N	52.68	52.30	50.77	51.92	54.02	0	0
% S	21.65	21.65	25.86	21.84	21.46	47.51	0

Table 4: 1 year fitted distribution statistics

Level of Conviction

The final stage of the model calibration process is to optimise the signal thresholds for our fitted distribution. To achieve this goal, we test strategies employing buy and sell signals at three different percentiles, 70:30 , 80:20 and 85:15.

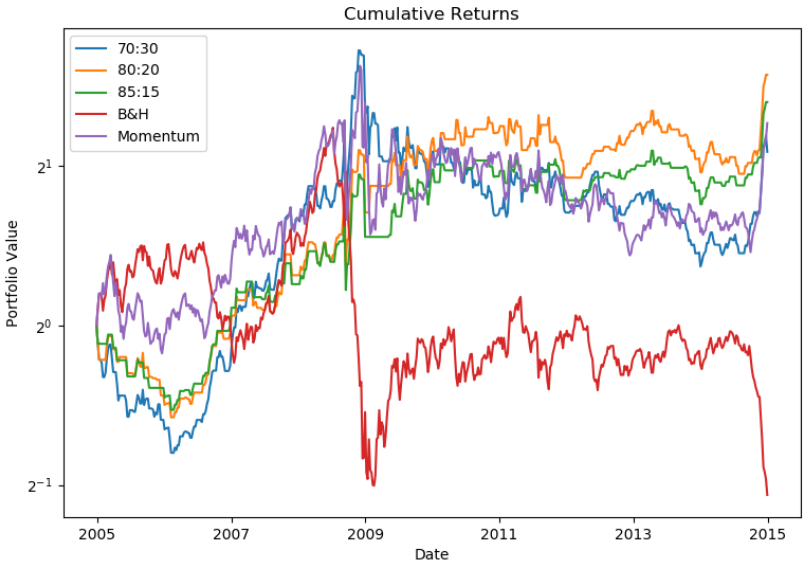

Figure 6: Conviction level for single asset strategy

Metric	**70:30**	**80:20**	**85:15**	**Momentum**	**B&H**
$\beta_{B\&H}$	-0.19	-0.17	-0.14	0.08	1
ρ_{Mom}	0.34	0.30	0.33	1	0.08
$r_{An.}(\%)$	7.84	11.52	10.20	9.20	-7.11
$\sigma_{An.}(\%)$	28.37	24.51	22.60	34.65	34.73
SR	0.28	0.47	0.45	0.27	-0.2
MDD (%)	60.89	29.72	27.89	56.13	79.72
$RoMaD$	0.13	0.39	0.37	0.16	-0.09
$HitRatio$ (%)	54.99	57.49	60.42	54.79	52.30
% L	36.59	25.67	18.77	52.49	100
% N	32.76	52.68	63.22	0	0
% S	30.65	21.65	18.01	47.51	0

Table 5: Single asset level of conviction

Considering the metrics in Table 5, we conclude that making use of the 80% (20%) percentiles of our fitted distribution for our buy (sell) signals produces the best risk-adjusted returns, as indicated by the $RoMaD$ metric.

Optimal extreme sentiment strategy

In summary, when optimising our Extreme Sentiment strategy on a single commodity, we have concluded that the optimal trading strategy is achieved by fitting a normal distribution to a 1-week rolling window, allocating positions to the long (short) portfolio using the risk-adjusted method.

5.3 Regression Derived strategy

When validating the regression models, we consider the coefficient of determination, R^2 coefficient, to select our optimal regression model. The R^2 is calculated as the average of the k R^2 coefficients output by our k-fold cross validation. When calibrating the two models, we considered a finite list of values for our regularisation parameter, and use $k = 10$ for our validation process. Results of the Ridge and LASSO regression fitting can be found in Tables 12 and 13 respectively.

With reference to our R^2 metrics, the best fit model for LASSO occurred when all regression parameters are set to 0, while the Ridge regression continued to improve as the regression coefficients shrink towards 0. Therefore, we conclude that the feature matrix input has no predictive power, when using a linear regression model, with respect to the commodity future return series. We abandon our regression approach to modelling the futures return, and do not present the performance of the model in Section 6.

6 Results

6.1 In-sample results

6.1.1 Ranked Sentiment strategy

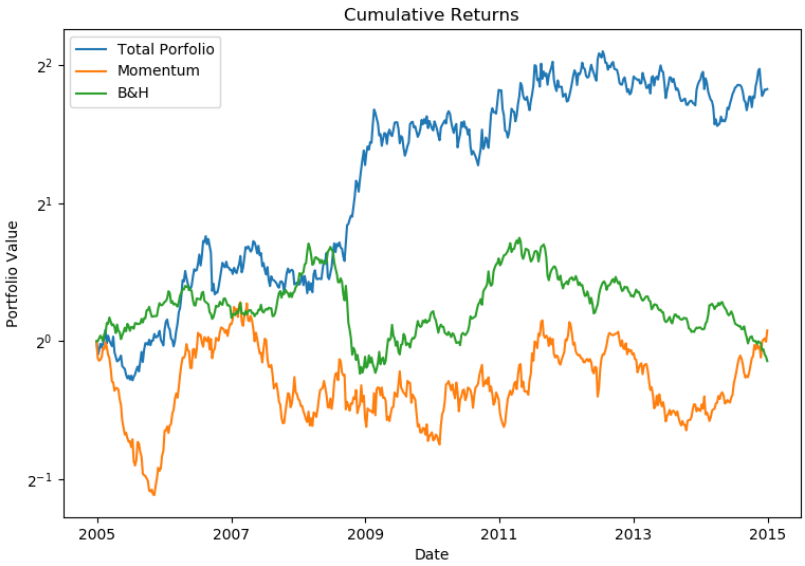

Figure 7: Ranked Sentiment strategy - Cumulative returns

Metric	Ranked Sentiment	Momentum	B&H
$\beta_{B\&H}$	-0.06	-0.06	1
$\rho_{Mom.}$	0.20	1	-0.04
$r_{An.}(\%)$	13.51	0.55	-0.99
$\sigma_{An.}(\%)$	27.56	30.44	17.80
SR	0.49	0.02	-0.06
MDD (%)	31.24	53.58	47.96
$RoMaD$	0.43	0.01	-0.02
$HitRatio$ (%)	51.34	50.00	52.30

Table 6: Ranked Sentiment strategy - return statistics

Comparing the return statistics in Table 6, the Ranked Sentiment strategy has outperformed both benchmarks on an absolute and risk-adjusted basis. Turning our attention to the cumulative return graph, Figure 7 reveals some interesting characteristics about the strategy. While both the sentiment strategy and momentum benchmark suffer losses in 2005, the losses observed in the sentiment strategy are muted compared to those of the momentum benchmark. In addition,

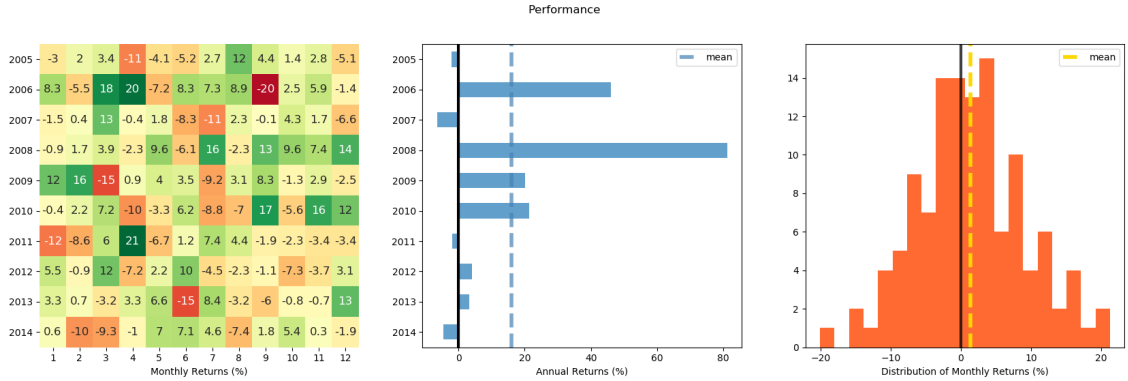

Figure 8: Ranked Sentiment strategy - Distribution of returns

when commodity prices slumped in 2008, the Ranked Sentiment strategy achieved its best period of performance, lending credibility to the findings of Garcia [15], that investor sentiment has a prominent effect during bad times. This is made more clear in the aggregated returns shown in Figure 8, where the strategy consistently generated excess returns in the last four months of 2008 and into the beginning of 2009.

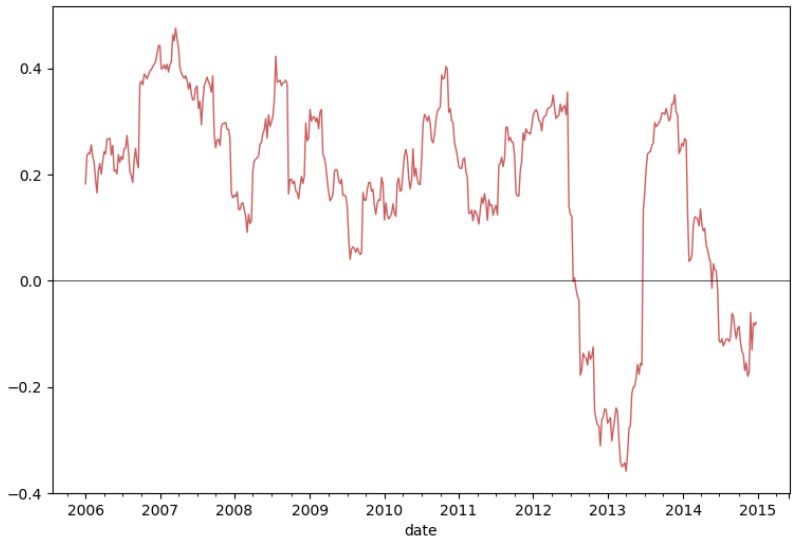

Figure 9: Ranked Sentiment strategy - 1-year rolling ρ_{Mom}.

Turning our attention to the rolling correlation of our strategy with the momentum benchmark, we can see that our strategy consistently has a rolling 1-year correlation with momentum in the corridor of 0.1-0.4, except in the period of 2012-2013 when it briefly falls into negative territory. This indicates that the driver of returns in our strategy may be intrinsically linked to that of momentum, lending credibility to the underreaction based explanation for the momentum effect

([12], [31] & [19]).

6.1.2 Extreme Sentiment strategy

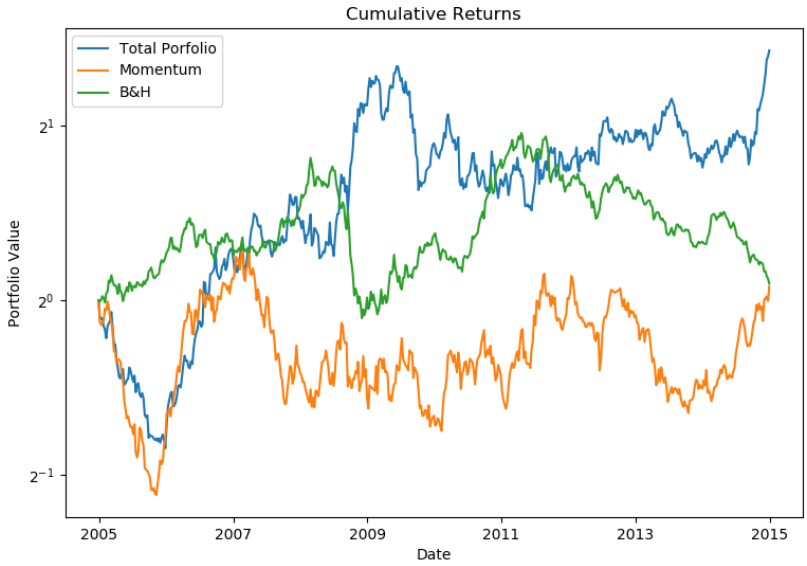

Figure 10: Extreme Sentiment Strategy - Cumulative returns

Metric	Extreme Sentiment	Momentum	B&H
$\beta_{B\&H}$	-0.17	-0.05	1
$\rho_{Mom.}$	0.22	1	-0.04
$r_{An.}(\%)$	10.43	0.55	-0.99
$\sigma_{An.}(\%)$	27.44	30.44	17.80
SR	0.38	0.02	-0.06
MDD (%)	43.52	53.58	47.96
$RoMaD$	0.24	0.01	-0.02
$HitRatio$ (%)	52.68	50.00	52.30

Table 7: Extreme Sentiment strategy - In-sample return statistics

From Table 7 we can see that the Extreme Sentiment strategy also outperforms both benchmarks from both an absolute and risk adjusted perspective, but not to the same extent as the Ranked Sentiment strategy. When analysing the best periods of performance for this strategy, we can see that, similar to that of the Ranked Sentiment strategy, the period of best performance was at the end of 2008. The Extreme Sentiment strategy also generated strong returns at the end of 2014, another bad period of performance for the Buy & Hold benchmark. This finding

indicates that Net Sentiment Score observations in the lower tail of the distribution may have the best predictive power with regards to asset returns, and that the Extreme Sentiment strategy is better positioned to exploit this relationship. However during the period of greatest drawdown for the momentum strategy, 2005, our Extreme Sentiment strategy participates to a far greater extent than the Ranked Sentiment strategy, with monthly aggregated returns displayed in Figure 11.

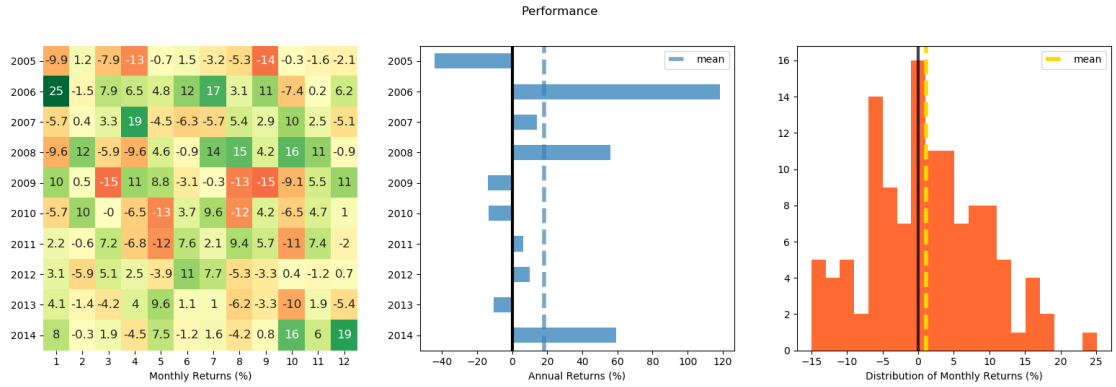

Figure 11: Extreme Sentiment strategy - Distribution of returns

When looking at the rolling correlation with graph in Figure 12, the strategy demonstrates a similar relationship to that of the Ranked Sentiment strategy, although demonstrating larger deviations from the average.

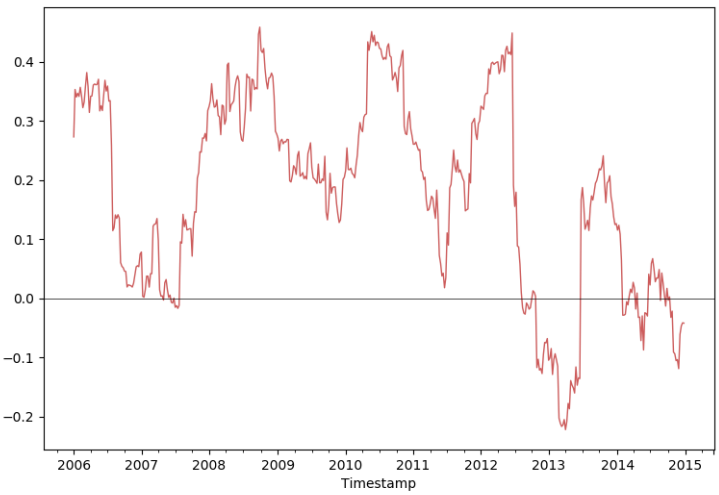

Figure 12: Extreme Sentiment Strategy - 1-year rolling $\rho_{Mom.}$

6.2 Out-of-sample results

6.2.1 Ranked Sentiment strategy

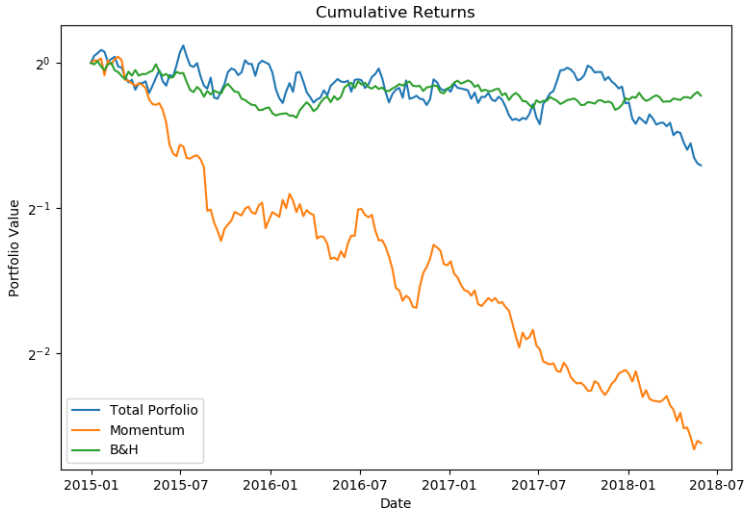

Figure 13: Ranked Sentiment strategy - Out-of-sample Cumulative returns

Metric	Ranked Sentiment	Momentum	B&H
$\beta_{B\&H}$	-0.14	0.01	1
$\rho_{Mom.}$	0.11	1	-0.04
$r_{An.}(\%)$	-13.44	-41.47	-2.61
$\sigma_{An.}(\%)$	26.25	30.38	11.56
SR	-0.51	-1.37	0.23
MDD (%)	43.62	84.68	20.52
$RoMaD$	-0.31	-0.49	-0.13
$HitRatio$ (%)	46.63	37.08	51.69

Table 8: Ranked Sentiment strategy - return statistics

In contrast to our in-sample period, the out-of-sample period is a period of low volatility for the Buy & Hold benchmark, where returns were largely flat over the period. While the Ranked Sentiment strategy outperformed the Momentum benchmark, which lost nearly 85% of its value over the period, it underperforms the Buy and Hold portfolio.

6.2.2 Extreme Sentiment strategy

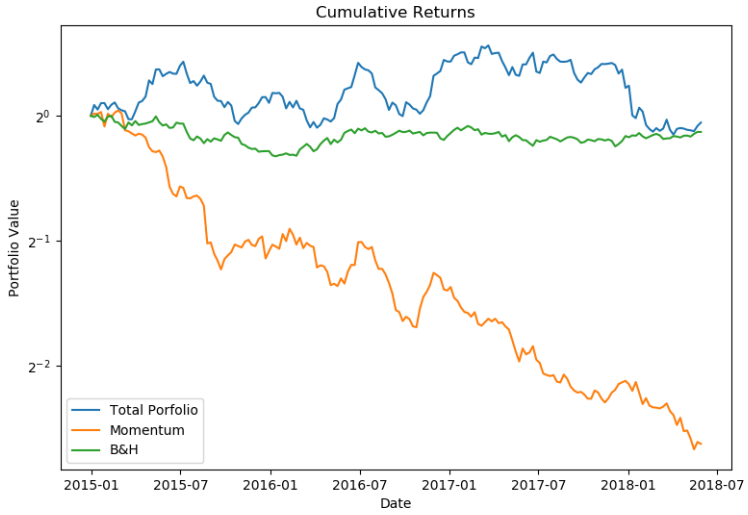

Figure 14: Extreme Sentiment Strategy - Out-of-sample Cumulative returns

Metric	**Extreme Sentiment**	**Momentum**	**B&H**
$\beta_{B\&H}$	-0.18	0.01	1
$\rho_{Mom.}$	0.19	1	0
$r_{An.}(\%)$	-1.09	-41.47	-4.52
$\sigma_{An.}(\%)$	27.70	30.40	12.70
SR	-0.04	-1.36	0.04
MDD (%)	38.99	84.68	23.74
$RoMaD$	-0.03	-0.49	-0.19
$HitRatio$ (%)	48.90	37.10	50.00

Table 9: Extreme Sentiment strategy - Out of sample return statistics

The Extreme Sentiment strategy outperforms both benchmarks, as well as the the Ranked Sentiment strategy in the out-of-sample period, restricting losses to 1.09% per annum, and reducing the maximum drawdown observed.

7 Discussion

We have implemented several systematic trading strategies, in an attempt to exploit market inefficiencies, born from investor reaction to market sentiment. In doing so, both strategies responded quickly in times of market distress, producing their best periods of returns during a market downturn, lending credibility to the findings of Garcia [15], that people react more to information when primed into negative mood states. In addition, the correlation metric of our strategies with the Momentum benchmark hint that the driver of returns may be intrinsically linked to that of momentum. This finding is most evident in the single asset Extreme Sentiment strategy, which exhibited a correlation of 0.34 with a momentum strategy based on crude oil. This result lends credibility to the under-reaction based explanation for the momentum effect ([12], [31] & [19]).

During the in-sample period, both strategies were able to capitalise on market distress, with the Extreme Sentiment portfolio being slightly more responsive, generating significant returns in both 2008 and 2015. Meanwhile the Ranked Sentiment strategy demonstrated better downside protection during the largest drawdown for the momentum strategy, losing 31% compared to 44% for the Extreme Sentiment strategy. While neither strategies were able to generate strong returns during the out-of-sample period, when volatility was low and the commodity futures market was flat, both demonstrated protection against the losses incurred by the momentum portfolio.

Although our results are inconclusive regarding which strategy is preferable, the results are still encouraging with regards to a sentiment derived trading strategy. To improve these strategies there are a number of avenues we would like to explore further. The first is to include a sentiment dispersion factor into our models, in an attempt to forecast periods of higher market volatility, and position the portfolio accordingly. In addition, further efforts are required to measure the sentiment-momentum effect over varying time horizons and market conditions, with a view to further strengthen our evidence that the underreaction to news articles is also apparent in the commodities futures market.

A Data

Sentiment Dataset Features

Feature	Details
Timestamp	The publication time stamp of the news item
StoryID	Unique ID assigned to every news item that the sentiment engine processes
Commodity	The commodity that the article concerns
Sentiment	Trinary score - the highest probable state of sentiment found in the news item
Confidence	The probability of the trinary sentiment score (above) matching a known sentiment state alone
Novelty	Identifies new stories: 1 = new story, 2 or greater means the item is linked to a previous story that was published within the last 24 hours
HeadlineOnly	Identifies if the item is headline only or if there is a body as well
AutoMessage	Identifies if the story is generated as part of a schedule delivery
Events	Identifies the subject or topic of the news item
Source	A tag that identifies the publication name or source type
Prob_POS	The raw probability that the article sentiment is positive
Prob_NEU	The raw probability that the article sentiment is neutral
Prob_NEG	The raw probability that the article sentiment is negative

Commodity return series

Commodity	Ticker
Coffee	KC1 A:00_0_R Comdty
Copper	HG1 A:00_0_R Comdty
Corn	C 1 A:00_0_R Comdty
Cotton	CT1 A:00_0_R Comdty
Gold	GC1 A:00_0_R Comdty
Natural Gas	NG1 A:00_0_R Comdty
Silver	SI1 A:00_0_R Comdty
Sugar	SB1 A:00_0_R Comdty
Wheat	W 1 A:00_0_R Comdty
Oil	CL1 A:00_0_R Comdty

B Model Calibration

B.1 Extreme Sentiment Strategy

Rolling window length

Metric	Normal	T	Power Normal	Johnsons U	Weibull min	Mom.	B&H
$\beta_{B\&H}$	-0.09	-0.09	-0.10	-0.09	-0.10	0.08	1
ρ_{Mom}	0.32	0.33	0.33	0.33	0.32	1	0.08
$r_{An.}(\%)$	7.18	6.35	5.78	6.35	6.01	9.20	-7.11
$\sigma_{An.}(\%)$	24.18	24.38	24.34	24.38	24.19	34.65	34.73
SR	0.3	0.26	0.24	0.26	0.25	0.27	-0.21
MDD (%)	52.04	52.04	52.04	52.04	53.22	56.13	79.72
$RoMaD$	0.14	0.12	0.11	0.12	0.11	0.16	-0.09
$HitRatio$ (%)	58.72	58.40	57.98	58.40	58.77	54.79	52.30
% L	31.61	31.80	31.61	31.80	30.27	52.49	100
% N	54.98	54.41	54.41	54.41	56.32	0	0
% S	13.41	13.79	13.98	13.79	13.41	47.51	0

Table 10: 5 year fitted distribution statistics

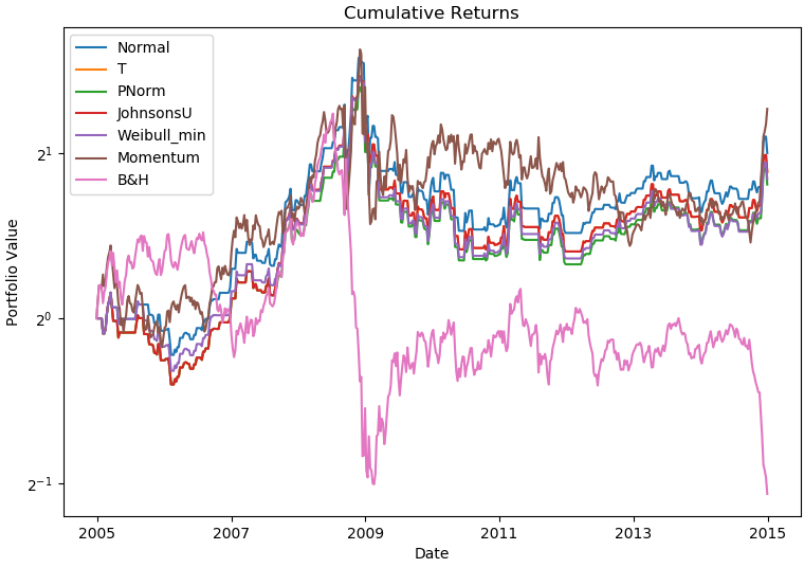

Figure 15: Fitted distribution - 5 year rolling window

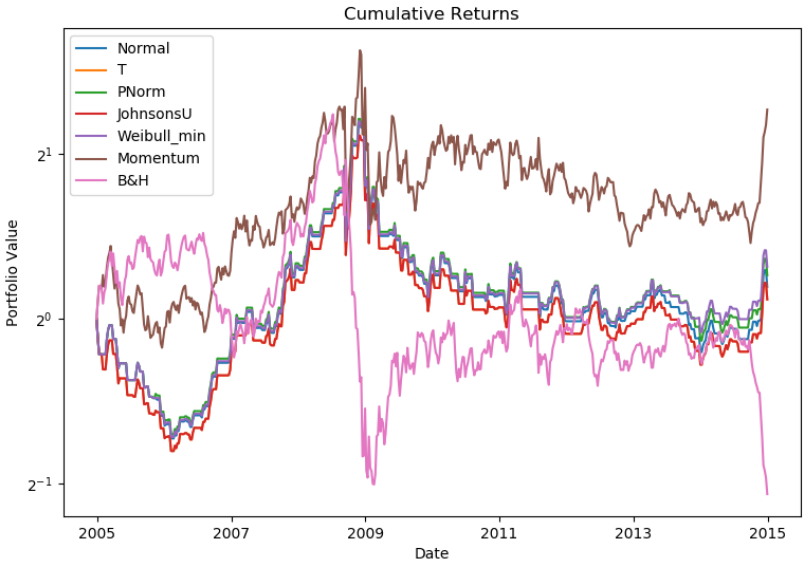

Figure 16: Fitted distribution - 3 year rolling window

Metric	**Normal**	**T**	**Power Normal**	**Johnsons U**	**Weibull min**	**Mom.**	**B&H**
$\beta_{B\&H}$	-0.16	-0.16	-0.16	-0.16	-0.16	0.08	1
ρ_{Mom}	0.30	0.30	0.30	0.30	0.30	1	0.08
$r_{An.}(\%)$	1.35	0.82	1.84	0.82	2.21	9.20	-7.11
$\sigma_{An.}(\%)$	24.23	24.32	24.22	24.32	24.18	34.65	34.73
SR	0.06	0.03	0.08	0.03	0.09	0.27	-0.21
MDD (%)	61.80	61.80	60.61	61.80	58.73	56.13	79.72
$RoMaD$	0.02	0.01	0.03	0.01	0.04	0.16	-0.09
$HitRatio$ (%)	56.22	55.93	56.65	55.93	57.21	54.79	52.30
% L	26.82	27.01	26.82	27.01	26.63	52.49	100
% N	55.36	54.79	55.36	54.79	56.13	0	0
% S	17.82	18.20	17.82	18.20	17.24	47.51	0

Table 11: 3 year fitted distribution statistics

B.2 Regression derived strategy

Feature Matrix	**1E-05**	**1E-03**	**2E-03**	**3E-03**	**7E-03**	**9E-03**	**1E-02**	**3E-02**	**7E-02**	**9E-02**	**1E-01**	**3E-01**	**1E+00**	**1E+01**	**3E+01**
Normalised	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.18	-0.17	-0.16	-0.15	-0.10
PCA80	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.06
PCA90	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09
PCA95	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.09	-0.08

Table 12: Ridge R^2 coefficients evaluated for various values of our hyper-parameter λ, using $k = 10$ for our k-fold cross validation

Feature Matrix	**1E-05**	**1E-03**	**2E-03**	**3E-03**	**7E-03**	**9E-03**	**1E-02**	**3E-02**	**7E-02**	**9E-02**	**1E-01**	**3E-01**	**1E+00**	**1E+01**	**3E+01**
Normalised	-0.17	-0.05	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
PCA80	-0.07	-0.04	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
PCA90	-0.10	-0.05	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
PCA95	-0.08	-0.05	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02

Table 13: LASSO R^2 coefficients evaluated for various values of our hyper-parameter λ, using $k = 10$ for our k-fold cross validation

References

[1] M. Alanyali, H. S. Moat, and T. Preis, 2013, Quantifying the relationship between financial news and the stock market, *Scientific Reports*, Vol. 3, pp. 3578

[2] W. Antweiler, and M. Z. Frank, 2002, Is All That Talk Just Noise? The Information Content Of Internet Stock Message Boards, *Journal of Finance*, Vol. 59, No. 3, pp. 1259-1294

[3] B. M. Barber, and T. Odean, 2008, All That Glitters: The Effect of Attention and News on the Buying Behaviour of Individual and Institutional Investors, *Review of Financial Studies*, Vol. 21, No. 2, pp. 785-818

[4] N. Barberis, A. Shleifer, and R. Vishny, 1998, A model of investor sentiment, *Journal of Financial Economics*, Vol. 49, No. 3, pp. 307-343

[5] J. Bollen, H. Mao, and X. Zeng, 2010, Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1-8

[6] M. Butler, and V. Kešelj, 2009, Financial forecasting using character n-gram analysis and readability scores of annual reports, *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, pp. 39-51

[7] A. Charoenrook, 2006, Does Sentiment Matter?, working paper, Vanderbilt University, https://www.researchgate.net/publication/228760482

[8] A. Chatrath, H. Miao, S. Ramchander, and S. Villupuram, 2014, Currency jumps, cojumps and the role of macro news, *Journal of International Money and Finance*, Vol. 40, pp. 42-62

[9] H. Chen, P. De, Y. Hu, and B. Hwang, 2014, Wisdom Of Crowds: The Value Of Stock Optinions Transmitted Through Social Media, *Review of Financial Studies*, Vol. 27, No. 5, pp. 1367-1403

[10] Z. Da, J. Engewlberg, and P. Gao, 2015, The Sum of All FEARS Investor Sentiment and Asset Prices, *Review of Financial Studies*, Vol. 28, No. 1, pp. 1-32

[11] S. M. Das, and M. Y. Chen, 2007, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, *Management Science*, Vol. 53, No. 9, pp. 1375-1388

[12] W. F. M. De Bondt, and R. Thaler, 1985, Does the Stock Market Overreact?, *Journal of Finance*, Vol. 40, No. 3, pp. 793-805

[13] M. Dzielinski, and H. Hasseltoft, 2012, Aggregate News Tone, Stock Returns, and Volatility, Working Paper, 10.2139/ssrn.2146706.

[14] E. F. Fama, 1970, Efficient Capital Markets: A Review of Theory and Empirical Work, *The Journal of Finance* , Vol. 25, No. 2, pp. 383-417

[15] D. Garcia, 2013, Sentiment During Recessions, *The Journal of Finance*, Vol. 68, No. 3, pp. 1267-1300

[16] D. Greely, and J. Currie, 2008, Speculators, Index Investors, and Commodity Prices, *Goldman Sachs Commodities Research*

[17] S. S. Groth, and J. Muntermann, 2011, An intraday market risk management approach based on textual analysis, *Decision Support Systems*, Vol. 50, No. 4, pp 680-691

[18] T. Hastie, R. Tibshirani, and J. Friedman, 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, *New York : Springer*, 2nd ed., pp. 61-73

[19] A. Hillert, H. Jacobs, and S. Müller, 2014, Media Makes Momentum, *Review of Financial Studies*, Vol. 27, No. 12, pp. 3467-3501

[20] N. Jegadeesh, and S. Titman, 1993, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *The Journal of Finance*, Vol. 48, No. 1, pp. 65-91

[21] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, and N. Ramakrishnan, 2013, Forex-Foreteller: Currency Trend Modeling using News Articles, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1470-1473

[22] F. Li, 2010, The Information Content of Forward-Looking Statements in Corporate Filings — A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research*, Vol. 48, No. 5, pp. 1049-1102

[23] A. Mahajan, L. Dey, and S. M. Haque, 2008, Mining Financial News for Major Events and Their Impacts on the Market, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 423-426

[24] A. Ng, 2017, CS229 Lecture notes, Part V, Support Vector Machines, pp. 1-5

[25] A. Ng, 2017, CS229 Lecture notes, Part VII, Regularization and model selection, pp. 3-4

[26] A. Ng, 2017, CS229 Lecture notes, Part XI, Principal components analysis, pp. 1-6

[27] D. Peramunetilleke, and R. K. Wong, 2002, Currency Exchange Rate Forecasting from News Headlines, *Proceedings of the 13th Australasian Database Conference*, Vol. 5, pp. 131-139

[28] G. Rachlin, M. Last, D. Alberg, and A. Kandel, 2007, ADMIRAL: A Data Mining Based Financial Trading System, *IEEE Symposium on Computational Intelligence and Data Mining*, pp. 720-725

[29] A. Ratnaparkhi, 1996, A Maximum Entropy Model for Part-Of-Speech Tagging, *Conference on Empirical Methods in Natural Language Processing*

[30] S. L. Salzberg, 1994, C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993, *Machine Learning*, Vol. 16, No.3, pp. 235-240

[31] J. Scott, M. Stumpp, and P. Xu, 2003, News, Not Trading Volume, Builds Momentum, *Financial Analysts Journal*, Vol. 59, No. 2, pp. 45-54

[32] N. R. Sinha, 2016, Underreaction to News in the US Stock Market, *Quarterly Journal of Finance*, Vol. 6, No. 2, pp. 1-46

[33] A. Soni, N. J. van Eck, and U. Kaymak, 2007, Prediction of Stock Price Movements Based on Concept Map Information, *IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, pp. 205-211

[34] C. Sutton, and A. McCallum, 2012, An Introduction to Conditional Random Fields, *Foundations and Trends in Machine Learning*, Vol. 4, Issue 4, pp 267-373

[35] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals *The Journal of Finance*, Vol. 63, No. 3, pp. 1437-1467

[36] T. T. Vu, S. Chang, Q. T. Ha, and N. Collier, 2012, An experiment in integrating sentiment features for tech stock prediction in twitter, *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data*, pp. 2338

[37] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang, 1998, Daily Stock Market Forecast from Textual Web Data, *IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3, pp. 2720-2725