1. HARNESSING INVESTOR SENTIMENT USING BIG DATA ANALYTICS
HARNESSING INVESTOR SENTIMENT USING BIG DATA ANALYTICS

This study examines the statistical and economic significance of investor sentiment, based on general business news, on stock market returns and volatility. Using big data analytics, our findings reveal that sentiment does not affect market returns. However, sentiment does influence volatility, with negative (positive) sentiment increasing (decreasing) volatility. Investor sentiment is also economically significant; we demonstrate that an ETF-based trading strategy can be used to capitalize on the predictive capability of investor sentiment. This paper summarizes the research findings made by Johnman, Vanstone and Gepp (2018) from a more practical perspective.

Academic research broadly classifies investors as either ‘noise traders’ (retail investors with random beliefs about future returns) or ‘rational arbitrageurs’ (sophisticated investors who hold more informed beliefs about future returns) (De Long et al., 1990). Investor sentiment research typically measures sentiment using textual data from financially focused news sources, such as The Wall Street Journal, which are more likely to be read by and influence sophisticated investors. The seminal papers of Tetlock (2007) and Tetlock, Saar-Tsenchansky and Macskassy (2008), further expanded by Ferguson et al. (2015), show that investor sentiment is capable of predicting asset prices, with positive (negative) sentiment predicting positive (negative) financial returns. Additionally, positive (negative) investor sentiment has been shown to decrease (increase) volatility in financial returns (Kumari and Mahakud, 2015). Negative sentiment’s effect on financial returns and volatility is usually found to be stronger than positive sentiment. Furthermore, news-based trading strategies can be developed to demonstrate the economic significance of the effects of investor sentiment on financial returns and volatility. This study extends prior work by examining the statistical and economic effect of sentiment, derived from business news published by the Guardian Media Group, on market returns and volatility. In contrast to previous literature, the data source used in this study is more likely to influence the investment choices of retail investors, who often do not have access to more financially focused news sources, such as Bloomberg. This allows us to gain insight into the effects of retail investor sentiment on market returns and volatility.
Use of Sentiment Analysis to Predict Financial Markets: An Overview

Using sentiment analysis in the context of a financial market usually comprises three key components: data sources, the sentiment analysis process, and a trading strategy. The sentiment analysis process involves extracting sentiment information from textual data and representing it in a numerical format, which we refer to as a sentiment analysis metric (SAM). Some of the literature stops at this point, merely examining the statistical significance of a SAM. However, this study also employs a trading strategy – rules to make trading decisions based on the SAM – to ascertain its economic significance.

Data Sources

The data sources comprise both the textual data for measurement of sentiment and the financial markets data for evaluation purposes. A range of textual data sources have been used, including news articles from prominent newspapers, message and discussion boards, Twitter, corporate announcements, macro-economic news announcements, and company annual reports. Financial markets data are typically stocks from major indexes (e.g. S&P 500), but have also included currency exchange rates and gold futures.

Sentiment Analysis Process

The sentiment analysis process involves three stages, namely feature extraction, feature representation and sentiment classification. In the feature extraction phase, features (variables) representative of investor sentiment are extracted from the text (typically as discrete words or phrases). These features are subsequently represented in a numerical format (e.g. the number of times a word appears). Finally, the sentiment classification phase involves processing the represented features to determine whether the text displays positive or negative sentiment. This processing often involves matching the features to a dictionary consisting of a collection of words that are pre-classified as being associated with positive or negative sentiment. Machine learning techniques, such as support vector machines or neural networks, are also often used during the sentiment classification phase.

Trading Strategy

The final component of the sentiment analysis process is the use of a trading strategy. A variety of trading strategies have been employed, including buying (or selling) an asset when a SAM is positive (or negative), or taking a long (or short) position in assets in the top (or bottom) section of investor sentiment rankings based on a SAM. The majority of trading strategies utilize short timeframes, typically daily or intraday (e.g. 20 minutes).

Utilizing Big Data Analytics to Conduct Sentiment Analysis

This study utilizes a dataset of 79,823 business news articles published by the Guardian Media Group between 02/01/2002 and 01/06/2016. The Guardian Media Group is a UK mass media group that publishes newspapers, including The Guardian, The Observer, and The Guardian Weekly. As the data source is a UK-based company with a large UK audience, the FTSE 100 is used as the source of financial markets data. Since the FTSE 100 cannot be directly traded, BlackRock’s iShares Core FTSE 100 UCITS ETF (ISF) is utilized for the trading strategy. The FTSE 100 and ETF data for the sample period are sourced from Bloomberg.

The unstructured textual dataset was collected utilizing the Guardian Media Group’s online application programming interface (API). Big data analytics is subsequently applied to the news articles to derive separate positive and negative SAMs. The sentiment analysis process is implemented by firstly defining each trading day’s news articles as those released online between 4.30pm of the previous day and 8am on the trading day (thus avoiding endogeneity concerns). The individual words (features) in each trading day’s news articles are matched against Loughran and McDonald’s (2011) finance-specific dictionary, with the count of positive (negative) words being used to create the positive (negative) SAM. To account for the varying number of news articles published per day, a ratio feature representation method is employed, with the positive (negative) SAM being divided by the total number of words in all articles for that day. The positive (negative) SAM is further standardized by subtracting the mean and dividing by the standard deviation of the previous calendar year’s fraction of positive (negative) words, as per Tetlock, Saar-Tsenchansky and Macskassy (2008).

Statistical and Economic Significance of Sentiment on Stock Market Returns and Volatility

Linear regression models are employed to measure the statistical significance of positive and negative sentiment (i.e. the SAMs) on daily excess returns and volatility in the FTSE 100 (Table 1). While sentiment has no discernable effect on returns, it does have a statistically significant effect on volatility, with negative (positive) sentiment increasing (decreasing) volatility. This suggests that the
behavior of retail investors based on sentiment does not influence market returns, but can add additional noise to the market, which increases volatility and may cause prices to temporarily deviate from their fundamental values.

It also highlights that sentiment measurements created from data sources targeted towards different types of investors can have different effects on financial markets. For example, Ferguson et al. (2015) and Tetlock, Saar-Tenschansky and Macskassy (2008) find that sentiment metrics created from data sources more likely to be read by sophisticated investors (e.g. Financial Times) have a statistically significant effect on market returns.

### TABLE 1: EFFECTS OF POSITIVE AND NEGATIVE SENTIMENT ON DAILY EXCESS RETURNS (PANEL A) AND DAILY VOLATILITY (PANEL B) IN THE FTSE 100

#### Panel A: Daily Excess Returns

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>T-Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.00008927</td>
<td>-0.437</td>
<td>0.662</td>
</tr>
<tr>
<td>Std_Positive</td>
<td>0.00006195</td>
<td>0.309</td>
<td>0.757</td>
</tr>
<tr>
<td>Std_Negative</td>
<td>0.00001652</td>
<td>0.088</td>
<td>0.930</td>
</tr>
<tr>
<td>F-Statistic</td>
<td></td>
<td>0.048</td>
<td>0.953</td>
</tr>
</tbody>
</table>

#### Panel B: Volatility

<table>
<thead>
<tr>
<th>Volatility Proxy</th>
<th>(Ri)^2 Coefficient</th>
<th>T-Statistic</th>
<th>P-Value</th>
<th>Hi – Li Coefficient</th>
<th>T-Stat</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.0001467</td>
<td>20.144</td>
<td>0.000 **</td>
<td>76.6704</td>
<td>101.375</td>
<td>0.000 **</td>
</tr>
<tr>
<td>Std_Positive</td>
<td>-0.0000175</td>
<td>-2.442</td>
<td>0.015 *</td>
<td>76.6704</td>
<td>101.375</td>
<td>0.000 **</td>
</tr>
<tr>
<td>Std_Negative</td>
<td>0.0000548</td>
<td>8.206</td>
<td>0.000 **</td>
<td>11.3370</td>
<td>16.338</td>
<td>0.000 **</td>
</tr>
<tr>
<td>F-Statistic</td>
<td>46.770</td>
<td>0.000 **</td>
<td>169.600</td>
<td>0.000 **</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Panel B presents the results of two volatility proxies used in prior research: squared returns (Ri)^2 and high-low range (Hi – Li). Statistical significance is denoted by * (5% level) and ** (1% level).
To determine the economic significance of these results, this study implements a short-term reversal trading strategy. The Active Strategy aims to exploit the additional volatility caused by retail investors by entering a trade when prices appear to have diverged from their fundamental values and exiting when they revert back. Specifically, the Active Strategy enters a trade at market close if the Std_Negative value for the day, calculated using news from the previous trading day’s close to today’s open, is greater than the 70th quantile of its values in the prior calendar year. The threshold ensures the strategy only takes positions when sentiment is substantially negative. Each trade uses all available equity and is exited at the market close of the day on which Std_Negative value falls below the 70% threshold.

As shown in Figure 1 and discussed in Johnman, Vanstone and Gepp (2018), the Active Strategy outperforms a simple Buy and Hold Strategy on both a risk-adjusted and absolute basis. The Active strategy has a larger Sharpe Ratio than the Buy and Hold Strategy, exhibiting higher returns and lower risk, as well as lower drawdown. Additionally, the Active Strategy’s average daily return for the days in which it is in the market is greater than that of the Buy and Hold Strategy. These results do not factor in practical market constraints such as transaction costs.
This study provides insight into how investors can harness retail investor sentiment using the power of big data analytics. Retail investors vastly outnumber sophisticated investors, and often only have access to general business news instead of specialized financial news. Although collectively they do not exert definitive effects on market returns in the FTSE 100, they can add noise to the market, thereby increasing volatility and potentially causing prices to temporarily deviate from their fundamental values. This is both a statistically and economically significant result, with this study demonstrating that an active trading strategy exploiting this short-term noise window outperforms a buy and hold strategy on a risk-adjusted basis. These findings reveal the potential and scope for sentiment analysis conducted using big data analytics, opening future avenues for trading strategy design. Future research could seek to replicate this analysis on multiple Australian indexes (e.g. ASX 200 and ASX 300), with there being evidence that the effects of sentiment on financial returns are stronger for stocks with lower market capitalization (Tetlock, Saar-Tsenchansky and Macskassy, 2008). Future research could also investigate the effect of practical constraints such as transactions costs. Additionally, the effect of sentiment on volatility could potentially be exploited with an options trading strategy. Such a strategy may also prove useful for hedging purposes.

References


