EJW
ECON JOURNAL WATCH
Scholarly Comments on
Academic Economics

Econ Journal Watch 14(3)
September 2017: 302–345

Shy of the Character Limit: “Twitter Mood Predicts the Stock Market” Revisited

Michael Lachanski and Steven Pav[1]

Link to Abstract

> *Derwent Capital Markets, a family-owned hedge fund, will offer investors the chance to use Twitter Inc. posts to gauge the mood of the stockmarket* [sic]*, said co-owner Paul Hawtin. … A paper by the University of Manchester and Indiana University published in October said the number of emotional words on Twitter could be used to predict daily moves in the Dow Jones Industrial Average. A change in emotions expressed online would be followed between two and six days later by a move in the index, the researchers said, and this information let them predict its movements with 87.6 percent accuracy.*
>
> —*Bloomberg News* (Jordan 2010)

> *The only dedicated “Twitter” hedge fund has shut down after deciding to offer its social media indicators to day traders instead.*
>
> —*Financial Times* (Mackintosh 2012)

The 2011 article “Twitter Mood Predicts the Stock Market” by Johan Bollen, Huina Mao, and Xiaojun Zeng, published in *Journal of Computational Science*, claims to use a proprietary algorithm’s estimates of collective ‘calm’-ness from Twitter

[1] The authors would like to thank for helpful comments Frank Fabozzi, Burton Malkiel, René Carmona, Mark TenPas, Paul Schepel, Ani Deshmukh, Jonathan Godbey, Yuliy Sannikov, Jim Liew, Andrew Sun, Jianqing Fan, Drew Griffen, Justine Yu, and participants in the 2014 and 2015 Financial Management Association conferences. Responsibility for error resides with the authors. The authors gratefully acknowledge financial support from Princeton University’s Economics Department, the Fred Fox ’39 Fund, the Eugene Wong ’55 Fund for Policy and Engineering, and the PIIRS Global Systemic Risk Initiative.

to achieve 86.7 percent accuracy in predicting the up-down sequence of the Dow Jones Industrial Average, a price-weighted stock market index of 30 large U.S. stocks. To exploit the finding, the authors teamed up with the hedge fund Derwent Capital Markets and raised tens of millions of dollars in capital. Less than one year after being founded, the hedge fund shut down and auctioned off its assets for approximately GBP 120,000, far below the reported break-even point of GBP 350,000 and the guidance price of GBP 5 million (Mackintosh 2012).[2] Still, the work of Bollen, Mao, and Zeng has led to a cottage industry of trying to use mood and other measures derived from text data for financial analysis.[3] The 2011 paper has garnered great attention from scholars, accumulating 2,514 Google Scholar citations and 678 Web of Science (Core Collection) citations as of April 24, 2017.

The piece that would become Bollen, Mao, and Zeng (2011) initially appeared in 2010 on arXiv, an open-source archive for papers in their prepublication form. It then received media coverage from *Huffington Post* (**link**), *Wired* (**link**), *MIT Technology Review* (**link**), *Time* (link **1**, **2**), *USA Today* (**link**), *The Atlantic* (**link**), *The Telegraph* (**link**), *The Daily Mail* (**link**), NPR (**link**), CNN (**link**), CNBC (**link**), and Fox Business Channel (**link**), among others. Perhaps the media hype resulted in a relaxation of normal quality-control practices.

The published article appeared not in a finance or economics journal but in Elsevier's *Journal of Computational Science*, which started in 2010. The article is not merely that journal's most highly cited paper: The article's Web of Science citation count exceeds that of the journal's next most-cited paper by more than 12-fold, and its Google citation count exceeds the same by more than 30-fold.[4] In many respects the article is an outlier.

[2] After shutting down its own trading operations, Derwent Capital Markets attempted to rebrand itself as a "trading platform" that sold Twitter mood time series to market participants. The fact that both ventures were called Derwent Capital Markets makes it difficult to know precisely when the hedge fund, typically referred to in news media as the "Derwent Capital Absolute Return Fund", failed. It appears that both ventures collectively lasted a total of less than one year from the start of operations. For more details, see Dew-Jones 2013; Milnes 2014.

[3] Logunov (2011), Logunov and Panchenko (2011), Chen and Lazer (2011), and Karabulut (2013) provide examples of more recent research that uses aggregated measures of public mood to predict market movements. Our literature review finds two similar studies for mood analysis that may predate Bollen, Mao, and Zeng (2011). Zhang et al. (2011) released their preliminary findings and published results at roughly the same time as Bollen, Mao, and Zeng (2011), and Gilbert and Karahalios (2010) released their results, based upon LiveJournal entries rather than Twitter, a few months prior. But neither of those has achieved as much fanfare as Bollen, Mao, and Zeng (2011), likely because neither makes such strong claims to be able to predict the Dow Jones's up-down sequence. Recent work in Twitter text mining for stock market prediction (e.g., Han 2012; Sprenger et al. 2014a; 2014b; Mao et al. 2015) has moved away from the approaches pioneered by Bollen, Mao, and Zeng.

[4] Using the Web of Science Core Collection on April 24, 2017, we tallied 678 citations to TMP. The next most-cited *Journal of Computational Science* article is Knyazkov et al. (2012) with 53 citations. As for Google citations, the comparison between those two papers is 2,514 versus 79.

The present paper is a critique of Bollen, Mao, and Zeng (2011). Our analysis sometimes becomes intricate, and we introduce abbreviations. Here we provide some primary abbreviations:

- BMZ stands for the set of authors Bollen, Mao, and Zeng.
- TMP, for "Twitter Mood Predicts…," stands for BMZ's paper published in 2011.
- DJIA, for Dow Jones Industrial Average.
- SR stands for Sharpe Ratio, a standard measure of portfolio performance.

No previous research, to our knowledge, has attempted to replicate the TMP findings in-sample. Here we attempt to replicate the results in TMP. Using a set of off-the-shelf content analysis algorithms on a subsample of tweets that includes 90 percent of BMZ's data, we fail to replicate TMP's "mood" effect. We suggest that BMZ's in-sample results stem from idiosyncratic features of the particular proprietary algorithm they used to fit noise in their dataset. BMZ throw out all Twitter data from 2007 and early 2008 without explanation. When we include the discarded data in our statistical analysis, we are unable to recover any statistically significant relationship between Twitter mood and the stock market, a result consistent with data snooping. We contend that the sophisticated mood analysis of the general micro-blogging public pioneered in TMP is not useful for predicting market returns at the daily frequency, and that the semblance of in-sample predictive power is driven not by their validity as risk-premia or even any behavioral relation but by a set of statistical biases.[5] We conclude that the appearance of predictive power is the result of a combination of three factors: (1) multiple-comparison bias driven by the high dimensionality of text data, (2) data snooping bias enabled both by sophisticated natural language processing algorithms and a long sample window,[6] and (3) publication bias.[7]

[5] While this essay focuses on TMP as an example of these biases, our criticisms apply with only slightly less force to most of the works profiled in Nassirtoussi et al. (2014); on the other hand, while we may be skeptical of the findings of Karabulut (2013), which echo those in TMP, few if any of our criticisms in *this* essay apply to his work, which uses a single measure of Gross National Happiness derived from text analysis of Facebook posts to anticipate next-day market changes.

[6] O'Connor (2014), especially in the third chapter of his dissertation, suggests that the heterogeneity of the results given by natural language processing techniques can give researchers an additional free parameter when conducting statistical analysis of text data, biasing their parameters toward statistically significant results.

[7] We cannot estimate the relative contributions of (2) and (3) to the final results presented in TMP because we do not have access to the underlying algorithm used by BMZ to construct their mood time series. For a quantitative evaluation of (1) see Lachanski (2016).

Spiritual followers of Eugene Fama (1970), we are skeptical of studies like TMP that claim to be able to use publicly available data to obtain Sharpe Ratios over 3 (see Lachanski 2016). We would not claim that mood and text analyses can never be profitably used.[8] But neither of the present authors has ever observed market participants successfully using Twitter text analysis to estimate collective mood states of the public for the purposes of predicting market indices, and we have seen no credible published studies since TMP suggesting it can be done.[9]

This essay includes the first attempt to replicate TMP in-sample. A number of failed out-of-sample attempts to replicate exist, but none use BMZ's 2008 data and so they leave open the possibility that the Twitter mood effect BMZ found was arbitraged away after TMP was uploaded to arXiv.[10] We avoid the problem by studying the same period of time, which should mean we have the same data as BMZ save for since-deleted tweets. We also extend BMZ's sample using tweets from before the sample window used in TMP, which allows us to evaluate the robustness of the Twitter mood effect. Conducting our statistical analysis on the extended window provides a natural check for possible selection bias. If selection bias on the sample window explains the statistically significant results found by BMZ in whole or in part, then we should find a much weaker Twitter mood effect in the extended sample of tweets than we do in BMZ's sub-sample.

Among the things we do in the present paper are the following:

1. We place BMZ's findings in the wider world of text-mining applications for empirical asset pricing.
2. We attempt to provide a complete documentation of the methods used in TMP.
3. Using a sample of Twitter data purchased from Twitter, we construct our own mood time series, including the time period covered in TMP: February 28, 2008 to November 3, 2008 (Eastern Standard Time) and using exactly the same search terms. Our sample should thus match that used in TMP, less the since-deleted and hence unavailable tweets. We provide enough detail to ensure reproducibility, and we perform the same hypothesis tests as in TMP. In Appendix II, we show that

[8] Tetlock (2007) provides evidence that the mood of journalists in contact with market participants can predict the DJIA.

[9] Derwent Capital Markets is neither the first nor the last failed hedge fund relying on collective mood data. Journalist Jonathan Yates (2011) quotes a MarketPsych manager saying that a combination of sentiment and collective mood analysis "was a viable strategy for us." Yates adds, "Revealed preference evinces that his firm abandoned it. No one does this because they are making too much money."

[10] At the time TMP was published, Twitter gave full access to all publicly available tweets for free. But it was difficult to collect, store, and then analyze a year's worth of tweets.

our time series replicates the visual characteristics of the time series presented in TMP.[11]

4. We present results from our own statistical analysis of BMZ's data. We use nonparametric, non-linear forecasting methods to attempt to capture any non-linear features of the effect of Twitter mood on the DJIA. We do not find statistically significant evidence of non-linear outperformance.

The reader might wonder how we got started on the present investigation. One of the present authors, Michael Lachanski, while doing a junior paper in finance at Princeton University, noticed that BMZ provided no motivation for using only tweets from February 28 to November 3, 2008, even though they had access to all 2006–2008 tweets. Lachanski (2014) conducted a Granger-causality analysis using data from November and December 2008 and found no significant effect of Twitter mood on changes in the DJIA. Independently, Steven Pav (2012a; b; 2013) had written publicly about methodological problems with BMZ (multiple comparisons, a putative predictive ability that would violate the efficient market hypothesis, etc.). Lachanski contacted Pav after completing his senior thesis on the same topic for the purpose of collaborating. That collaboration has yielded the present paper.

The danger zone

Although BMZ cite several empirical results in behavioral finance on sentiments (such as Edmans et al. 2007), in both TMP and in their public appearances, they do not attempt to relate their findings to equilibrium theories of investor sentiment (as is done, e.g., by DeLong et al. 1990). BMZ do not note that their results do not accord with other results in the behavioral finance literature; they compare their work to research that used Twitter buzz to predict box office receipts.[12]

[11] Appendix II explains the derivation of the framework in Table 2.1. In the literature review provided in Appendix II, we describe the methodology and findings of TMP in detail, taking care to document the ambiguities in TMP that prevent complete replication. Appendix II also discusses the theoretical framework, drawn from psychology, underlying TMP. For a picture of the text mining for finance and accounting literature from an empirical finance, machine learning, accounting, or interdisciplinary perspective, the interested reader is referred to four recent survey articles (Kearney and Liu 2014; Nassirtoussi et al. 2014; Loughran and McDonald 2016; Das 2014), and the review of investor sentiment theory by Baker and Wurgler (2007). Our appendices, code, and additional materials are available here.

[12] As it happens, the research on box office prediction using Twitter, namely Huberman and Sitar (2010), was subsequently discredited by Wong et al. (2012).

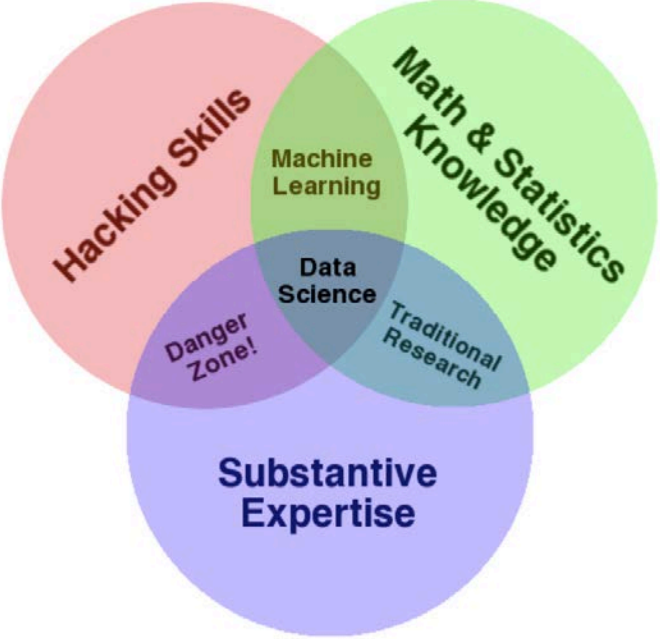


Figure 1.1. This figure corresponds to Figure 1 in Conway (2010).

TMP might be taken to be part of a trend of interdisciplinary research utilizing complex statistical tools. Such research, particularly when done in the private sector, has come to be called “data science” (Porter and Rafols 2009). We like the characterization given by Drew Conway (2010), who defines data science as any work in a particular field requiring “hacking skills,”[13] “math & statistics knowledge,” and “substantive expertise” of that field. In Conway’s typology, “traditional research” is the work in a particular field requiring math and statistics knowledge and domain expertise without hacking skills while “machine learning” tasks make use of math and statistics and hacking skills but do not require substantive expertise in the field being investigated. We believe that TMP makes a good example of work from what Conway calls the “Danger Zone” of data science (see Figure 1.1) because while BMZ have substantive expertise in the study of collective psychology, as well as considerable hacking skills, they make several statistical errors, as documented in Lachanski (2016). In fact, one academic re-

[13] Conway (2010) characterizes “hacking skills” in the following way: “For better or worse, data is a commodity traded electronically; therefore, in order to be in this market you need to speak hacker. This, however, does not require a background in computer science—in fact—many of the most impressive hackers I have met never took a single CS course. Being able to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically; these are the hacking skills that make for a successful data hacker.”

searcher has incorporated a close reading of TMP, as an example of the dangers of multiple-comparison bias when conducting inference on high-dimensional data, into his students' coursework.[14]

Animal spirits or statistical ghosts?

> Even apart from the instability due to speculation, there is the instability due to the characteristic of human nature that a large proportion of our positive activities depend on spontaneous optimism rather than on a mathematical expectation, whether moral or hedonistic or economic. Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits—a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities. (Keynes 1936, 161)

TMP claims to provide evidence in favor of what we call the 'empirical animal spirits hypothesis.'[15] The empirical animal spirits hypothesis is that current collective mood contains information about future asset prices not currently embedded in asset prices. The test of this hypothesis typically involves the scoring of text (usually news, as in Mitra and Mitra 2011, rather than tweets) with a single unipolar (e.g., is the text very subjective/emotional?) or bipolar measure (e.g., does the author feel positively or negatively about something?).

TMP scores tweet text according to seven bipolar dimensions of mood and finds that the "CALM" scores of tweets have more predictive power than traditional positive-negative measures. If true, it would suggest that current work in textual analysis for stock market prediction could be improved by incorporating measures that account for calmness or lack thereof rather than just positivity or negativity.

TMP found that only a single mood, CALM, was found to outperform traditional positive-negative measures of the emotional content of text when it comes to predicting the stock market. Four of the six mood measures tested exhibited no statistical significance and the sixth mood, "HAPPY," had a single model exhibiting statistical significance out of seven models estimated. Even without accounting for the multiple-comparison bias, given the correlation be-

[14] The researcher is Jacob Eisenstein of Georgia Tech (link).

[15] Not to be confused with the 'animal spirits hypothesis' in macroeconomics (e.g., Farmer and Guo 1994).

tween HAPPY and traditional positive-negative sentiment measure presented in TMP, this result is likely to be spurious.

BMZ frame their work as an attack on the efficient market hypothesis, which they imply is equivalent to the random walk hypothesis. As discussed in Lachanski (2016), such equivalence is incorrect; no adjustment for risk is conducted in TMP and so no explicit evidence is presented against the efficient market hypothesis.[16] TMP claims to provide evidence against the random walk hypothesis as a valid approximation to return series at the daily time horizon. Because well-cited evidence against the random walk hypothesis typically takes the form of high-frequency statistical tests (as in Niederhoffer and Osborne 1966, which uses tick data), or weekly (or longer) frequencies (as in Lo and MacKinlay 1988), we believe that it is valuable to assess BMZ's claim that data from online social media predicts changes in the stock market.

The methods of "Twitter Mood Predicts the Stock Market"

Brendan O'Connor (2014) defines text data as a "linear sequence of high-dimensional discrete variables (words)." Simple mean-variance investor models yield only two theoretical justifications for believing text data can forecast the stock market. The first justification is that information contained in soft variables like the tone of analyst reports or financial news enters asset prices only slowly. In this case, fundamental analysis of soft variables embedded in text data can allow us to forecast asset price changes. The second justification comes from noise-trader models (such as DeLong et al. 1990). Even media containing no new fundamental information can still be worth analyzing because of the effect that such media has on noise traders. The information and investor-sentiment theories make different predictions about the joint distribution of text-derived time series and asset prices. In Table 2.1, we compare BMZ's findings with the theoretical predictions made by the information and investor-sentiment theories, and we find that they do not appear to be compatible with either theory.

BMZ do not utilize a microfounded model to motivate their research. Instead, they characterize their work as an attempt to evaluate how psychological factors like mood impact equity prices. To do this, they conduct mood analysis on a sample of tweets. BMZ write: "We obtained a collection of public tweets

[16] However, Lachanski (2015) finds that the Sharpe Ratio obtainable by such a strategy would almost certainly violate no-arbitrage bounds on SR, given that the marginal investor has bounded risk aversion.

that was recorded from February 28 to December 19th, 2008 (9,853,498 tweets posted by approximately 2.7M users).” To do what they propose, they must devise a method for classifying and quantifying moods. To do so, BMZ create their own mood-state measures, which they say “are derived from an existing and well-vetted psychometric instrument, namely the Profile of Mood States (POMS-bi)” (BMZ 2011, 2).

The “POMS-bi” refers to the psychometric instrument developed by Maurice Lorr and Douglas McNair (1984). Lorr and McNair called it “bipolar” because it identifies six mood spectra, each comprised of two bipolar mood states, as visualized in Figure 2.1. We will call the mood states on the left hand side of Figure 2.1 negative mood states and the mood states on the right hand side of Figure 2.1 positive mood states. Unlike what a straightforward reading of TMP might lead one to think, the spectra labels CALM, ALERT, SURE, VITAL, KIND and HAPPY do not come from the POMS-Bi; those labels are the invention of BMZ.

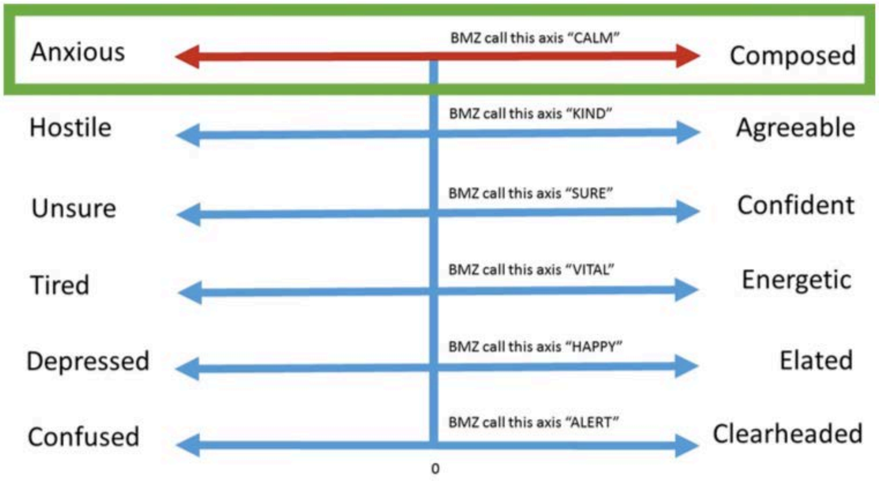

Anxious
BMZ call this axis “CALM”
Composed
Hostile
BMZ call this axis “KIND”
Agreeable
Unsure
BMZ call this axis “SURE”
Confident
Tired
BMZ call this axis “VITAL”
Energetic
Depressed
BMZ call this axis “HAPPY”
Elated
Confused
BMZ call this axis “ALERT”
Clearheaded
0

Figure 2.1. BMZ test for the presence of six mood effects. Only “CALM”-ness (framed in green and highlighted in red above) was found to Granger-cause DJIA increases.

The POMS-Bi associates each positive and negative mood state with six terms. For example, the POMS-Bi positive mood state “Composed” is associated with the word list: *composed*, *untroubled*, *peaceful*, *serene*, *calm*, and *relaxed*. The POMS-Bi negative mood “Anxious” is associated with the word list: *tense*, *nervous*, *jittery*, *shaky*, *anxious*, and *uneasy*. The Composed-Anxious spectrum consists of the twelve-word union of the Composed and Anxious word lists. Since BMZ find that only CALM, their construct based on the Composed-Anxious spectrum, has reasonable predictive power for changes in the DJIA, we only concern ourselves with tests of the effects of this particular mood. We will refer to that mood in general as Twitter mood, 'calm'-ness, Twitter calmness, collective mood, or the Composed-Anxious spectrum—i.e., all those terms are used synonymously in this paper from this point forward. When we are referring specifically to BMZ's mood time series as presented in TMP, we will call this CALM or the CALM time series.

Because of the small number of words in the POMS-Bi, BMZ might reasonably be concerned about covariate misclassification and data sparsity. If the language used to express mood on the internet differs from people's mental conceptions of their mood, then using the POMS-Bi terms will lead to systematic misclassification.

Consider the tweet: "I'm going to need to see Fleury's therapist after this. He's more composed than I am and I'm sitting at home eating spaghetti." A unigram bag-of-words system, using only the POMS-Bi terms, would have classified the tweet as more Composed than Anxious because it uses the word *composed*. However, in context the author appears to express anxiety relative to Fleury. Furthermore, because the POMS-Bi assigns only six words to the Composed word list, many tweets with terms expressing calmness (e.g., *undisturbed*, *unflappable*, or *placid*) will be missed by a system relying on word counts of POMS-Bi terms in tweets. Incorrectly assigning tweets to the calm category or neutral category are both forms of the misclassification problem. One way to decrease misclassification might be to add more unigrams to the model. For instance, the presence of the word *therapist* in the tweet might be taken as a sign of anxiety on the part of the Twitter user and mapped to the anxious half-scale. If all terms are term-weighted (i.e., we simply count the number of matches between our lexicon and the tweet), this would assign the whole tweet a Composed-Anxious score of 0, closer to the ground truth than our original classification. Another way to reduce misclassification might be to increase the complexity of the model by using a higher-order n-gram bag-of-words model. For instance, a mapping of the 4-gram "more composed than I" to the Anxious mood state would correctly classify the tweet above.

A subtler problem is that the use of a small lexicon on small documents like tweets we expect to obtain only a small number of term-matches per document. For our Twitter data, the modal number of a tweet's term matches, using any of the lexicons developed in the present paper, is zero. In fact, this sparsity of matches is common in many applications. Tim Loughran and Bill McDonald (2016) point out that with large lexicons much of the variance in word-count frequencies often comes from a small subset of words in the dictionary.[17] If such high-variance terms are included in a small lexicon, their variability will dominate the other words

[17] This empirical finding results from Zipf's law, discussed extensively in Manning and Schutze (2000).

in our lexicon; if they are excluded, we will have little variance in our sentiment series and consequently the estimators of the effect of sentiment on a variable of interest will be biased towards zero. These considerations suggest that the bias-variance tradeoff is best managed by including many high-variance terms. Since we cannot know which terms will have high-variance word counts before observing the documents, Loughran and McDonald (2016) suggest that one should err towards larger rather than smaller lexicons. For comparison, the Anxious lexicon of the POMS-Bi contains six words whereas the dictionary of Elaine Henry (2008), currently the smallest dictionary used in the text analysis for finance literature (Kearney and Liu 2014; Loughran and McDonald 2016), contains 85 negative terms.

Concerns about tweet misclassification and covariate sparsity motivate BMZ to use a lexicon with both more terms and higher-order n-grams than the POMS-Bi. BMZ write that:

> To make it [the POMS-Bi] applicable to Twitter mood analysis we expanded the original 72 terms of the POMS questionnaire to a lexicon of 964 associated terms by analyzing word co-occurrences in a collection of 4- and 5-grams computed by Google in 2006 from approximately 1 trillion word tokens observed in publicly accessible Webpages. … We match the terms used in each tweet against this lexicon. Each tweet term that matches an *n*-gram term is mapped back to its original POMS terms (in accordance with its co-occurrence weight) and via the POMS scoring table to its respective POMS dimension. (BMZ 2011, 2–3)

By expanding the POMS-Bi, BMZ can “capture a much wider variety of naturally occurring mood terms in Tweets and map them to their respective POMS mood dimensions” while also minimizing the misclassification and covariate concerns above. They multiply each tweet’s set of matches by a weight derived from term co-occurrences[18] before summing across mood states.

Since mood time series have no natural units, BMZ normalize them to enable comparison of the effects of changes in two different moods on stock prices in terms of changes in standard deviation in the moods. They write:

> [W]e normalize them to *z*-scores on the basis of a local mean and standard deviation within a sliding window of *k* days before and after the particular date. For example, the *z*-score of the time series X_t denoted $\mathbb{Z}_{X_t}$, is defined as:

[18] A co-occurrence typically refers to the above-chance frequent occurrence of two terms in a text corpus in a certain order.

$$\mathbb{Z}_{X_t} = \frac{X_t - \overline{x}\left(X_{t \pm k}\right)}{\sigma\left(X_{t \pm k}\right)} \qquad (1)$$

> where $\overline{x}(X_{t\pm k})$ and $\sigma(X_{t\pm k})$ represent the mean and standard deviation of the time series within the period $[t-k, t+k]$. This normalization causes all time series to fluctuate around a zero mean and be expressed on a scale of 1 standard deviation.
>
> The mentioned *z*-score normalization is intended to provide a common scale for comparisons of the [OpinionFinder] and GPOMS time series. However, to avoid so-called "in-sample" bias, we do *not* apply *z*-score normalization to the mood and DJIA time series that are used to test the prediction accuracy of our Self-Organizing Fuzzy Neural Network in Section 2.5. This analysis and our prediction results rest on the raw values for both time series and the DJIA. (BMZ 2011, 3)

> GPOMS and OpinionFinder time series were produced for 342,255 tweets in that period, and the daily Dow Jones Industrial Average (DJIA) was retrieved from Yahoo! Finance for each day. (ibid., 4)

We will generically refer to the technique expressed in expression (1) as 'local normalization.'

Ambiguities in "Twitter Mood Predicts the Stock Market"

As part of efforts to resolve ambiguities found in TMP, we have beginning in 2012 repeatedly reached out to BMZ for assistance but have not received any meaningful response.

One ambiguity is: On what time series was the linear time series analysis performed on? Was it performed on the normalized mood time series? BMZ specifically claim that they do not normalize their time series for the analysis presented in Section 2.5 of their paper, the non-linear models section. BMZ do not say whether or not they conducted their Granger causality analysis on the normalized mood time series.

Another ambiguity concerns their "scoring table." In their description of how they calculated the daily mood scores, BMZ refer to a "POMS scoring table." Our copy of the POMS-Bi exam does not contain anything labeled a "scoring table." Mood analysis, as described in Appendix II, provides many ways to adapt psychometric surveys to text data. One solution to this ambiguity might be provided by Bollen, Mao, and Alberto Pepe (2011, 451), who refer to a POMS-scoring function $\wp(t)$ and define it as follows:

> The POMS-scoring function $\wp(t)$ maps each tweet to a six-dimensional mood vector $m \in \mathbb{R}^6$. The entries of m represent the following six dimensions of mood: *Tension*, *Depression*, *Anger*, *Vigour*, *Fatigue*, and *Confusion*.
>
> The POMS-scoring function denoted $\wp(t)$ simply matches the terms extracted from each tweet to the set of POMS mood adjectives for each of POMS' 6 mood dimensions. Each tweet t is represented as the set w of n terms. The particular set of k POMS mood adjectives for dimension i is denoted as the set p_i. The POMS-scoring function, denoted $\wp$, can be defined as follows:
>
> $$\wp(t) \;\rightarrow\; m \in \mathbb{R}^6 = \left[\, \|w \cap p_1\|,\ \|w \cap p_2\|,\ \cdots,\ \|w \cap p_6\| \right]$$

If the POMS-scoring function defined above is the "POMS scoring table" (a simple count of term matches), then BMZ follow the procedure we use in our empirical replication.[19]

Another ambiguity concerns the defining of the sample. At the start of TMP, BMZ say they collected a sample of 9,853,498 tweets, but the sample BMZ analyze in Section 2.4 of their paper consists of 342,255 tweets.[20] BMZ do not describe how this subset of tweets was selected. One possibility is that the tweets selected for the analysis are those containing mood terms in the GPOMS lexicon. But that is unlikely to be the case, as we found 787,715 tweets classified as containing significant mood content using the *sentimentr* package (**link**) and our mood dictionaries over the timeframe analyzed in Section 2.4. One possibility is that they have a much smaller number of 3- and 4-gram matches. While we are unable to test this possibility, this would strongly suggest that BMZ discard the vast majority of mood variation in the data.

TMP's findings contra the literature

TMP purports to measure stock prices' sensitivity to mood changes with a finer granularity than do previous studies.[21] BMZ write:

[19] The POMS-scoring function as shown in Bollen, Mao, and Pepe (2011) is: $\wp(t) \;\rightarrow\; m \in \mathbb{R}^6 = \left[\, \|w \cap p_1\|,\ \|w \cap p_2\|,\ \ldots,\ w \cap p_6\| \right]$, which we believe contains a typo; we have corrected it in our quote.

[20] BMZ do not describe any subsampling procedure in Section 2.5 either; it seems reasonable to assume that they used their entire sample.

[21] This section substantially recapitulates Lachanski 2014.

> Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood. … if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for practical stock market prediction. Large surveys of public mood over representative samples of the population are generally expensive and time-consuming to conduct…. Some have therefore proposed indirect assessment of public mood or sentiment from the results of soccer games and from weather conditions. (BMZ 2011, 1–2)

Because BMZ do not provide the coefficients of their estimated autoregressive distributed lag (ARDL) models, we can only conduct a qualitative robustness check of their results. In all of their public appearances[22] and interviews BMZ suggest that Twitter 'calm'-ness has a positive effect on stock prices and stock price changes, as seemingly implied by graphics they supply (e.g., see Figure 3 in TMP, however see pages 331–334 of the present paper, showing that these graphics may be misleading).

To answer the question of whether Twitter mood predicts the stock market, BMZ follow a procedure presented by Mark Watson and James Stock (2010). BMZ obtain DJIA data from Yahoo! Finance and fit ARDL models to changes in stock prices from February 28, 2008 to November 3, 2008, removing weekends from their data. BMZ write:

> [W]e are concerned with the question whether other variations of the public's mood state correlate with changes in the stock market, in particular DJIA closing values. To answer this question, we apply the econometric technique of Granger causality analysis to the daily time series produced by GPOMS and OpinionFinder vs. the DJIA. Granger causality analysis rests on the assumption that if a variable X causes Y then changes in X will systematically occur before changes in Y. We will thus find that the lagged values of X will exhibit a statistically significant correlation with Y. … we are not testing actual causation but whether one time series has predictive information about the other or not.
>
> Our DJIA time series, denoted D_t, is defined to reflect daily changes in stock market value, i.e. its values are the delta between day t and day t−1: $D_t = DJIA_t − DJIA_{t−1}$. To test whether our mood time series predicts changes in stock market values we compare the variance explained by two linear models…
>
> We perform the Granger causality analysis according to model [*sic*] L_1 and L_2 … for the period of time between February 28 to November 3, 2008 to

[22] See, for example, Bollen's presentation on December 17, 2012 (**link**).

> exclude the exceptional public mood response to the Presidential Election and Thanksgiving from the comparison. (BMZ 2011, 4)

BMZ use two models L_1 and L_2, with the following structures, to evaluate the effect of mood at time t, denoted here by X_t, on equity price changes:

$$L_1 : D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-1} + \epsilon_t \tag{2}$$

$$L_2 : D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-1} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \epsilon_t \tag{3}$$

In these models, α, $\beta_{i,\, i \in \{1,\, \ldots,\, n\}}$, and $\gamma_{i,\, i \in \{1,\, \ldots,\, n\}}$ are parameters estimated from the data, D_t is as defined above by BMZ, and ϵ_t only requires the property that:

$$\begin{aligned} &E_t\left[\epsilon_t \;\middle|\; D_{t-1}, D_{t-2}, \ldots\right] = 0 \text{ for } L_1 \text{ and} \\ &E_t\left[\epsilon_t \;\middle|\; D_{t-1}, D_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots\right] = 0 \text{ for } L_2 \end{aligned} \tag{4}$$

The equations expressed in (3) are 'nested' ARDL models. We say that a linear model M_1 is nested in another linear model M_2 if all the terms in M_1 are also found in M_2. Since all the terms of the model L_1 are found in model L_2, we say that model L_1 is nested in model L_2. The nested structure allows us to conduct likelihood ratio tests. Furthermore, since our likelihood ratio tests are evaluating the statistical significance of past mood vector (X_{t-1}, X_{t-2}, …) on current DJIA changes D_t, we say that these likelihood ratio tests are testing for Granger causality. Watson and Stock (2010) define a Granger causality statistic as "the F-statistic testing the hypothesis that the coefficients on all the values of one of the variables" $X_{t-1}, X_{t-2}, \ldots, X_{t-q}$ are zero, and they state, "This null hypothesis implies that these regressors have no predictive content for Y_t beyond that contained in the other regressors, and the test of this null hypothesis is called the Granger causality test." BMZ evaluate the predictive power of Twitter mood using Granger causality for lags $n \in \{1, \ldots, 7\}$. For each n, the null and alternative hypotheses for the Granger causality tests are given by:

$$H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_n = 0 \tag{5}$$

$$H_A : \gamma_1 \neq 0 \;\vee\; \gamma_2 \neq 0 \;\vee\; \ldots \;\vee\; \gamma_n \neq 0 \tag{6}$$

BMZ's null hypothesis is that Twitter moods do not Granger-cause (i.e., do not predict) changes in DJIA values and the alternative hypothesis is that one or more of the lagged Twitter mood variables does Granger-cause (i.e., does predict) the stock market. BMZ deem F-statistics results to be statistically significant when p-values are less than 10 percent.

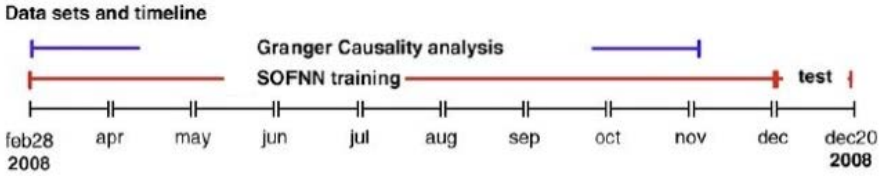


Figure 2.2. This figure corresponds with Figure 1 in TMP. Standard cross-validation of the models BMZ fit for their Granger-causality analysis would have tested their models, fitted using data from February 28 to November 3, with data from November and December 2008. BMZ conduct no such out-of-sample validation of their ARDL models in TMP. They do cross-validate their non-linear SOFNN model.

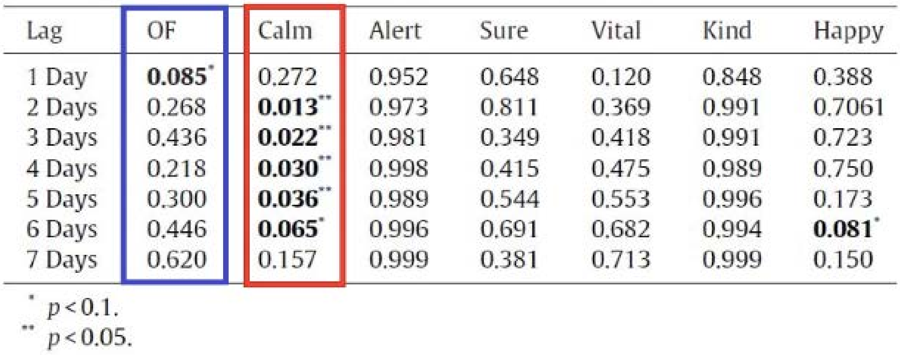
BMZ also want to check for non-linear relationships between mood and equity market behavior. First, they normalize all of their mood and equity market data to a [0,1] interval. We will refer to this type of normalization as a 'global normalization.' Next, they train a Self-Organizing Fuzzy Neural Net (SOFNN hereafter) from February 28, 2008 to November 28, 2008. They evaluate the SOFNN by testing its ability to predict the up-down sequence of DJIA prices from December 1 to December 19, 2008 (see Figure 2.2 above).

BMZ's results for their Granger causality test for nested models L_1 and L_2 are shown in Figure 2.3. BMZ find statistical significance at the 10th percentile for lags 2 through 6 of their ARDL model using a time series of Twitter 'calm'-ness. They also find that the first lag of the time series derived from the OpinionFinder tool, a standard text analysis tool, is statistically significant at the 10th percentile. Their SOFNN tool predicted correctly 13 out of 15 days tested or 86.7 percent accuracy; the probability of this occurring under the random walk hypothesis is 0.37 percent. BMZ interpret this finding as evidence that moods have non-linear effects on stock prices.

As a first pass at evaluating TMP's findings we can adapt the identification technique used by Paul Tetlock (2007) and schematized in Table 2.1.[23] In the well-cited literatures on news analytics and investor sentiment, the effects associated with information or high levels of linguistic sentiments in Twitter mood should appear immediately. The work on collective mood shocks cited in our Appendix II all suggest that changes in investor sentiments impact the stock market instantaneously. If Twitter mood predicts investor sentiment or information then high returns (positive changes in the DJIA) should occur the day after a high Twitter mood reading. TMP finds no next-day effect. What the p-values in Figure

[23] Like Tetlock (2007), who rules out the possibility that his pessimism time series forecasts positive sentiment, in Table 2.1 we do not present the possibility that TMP's CALM series forecasts negative sentiment.

2.3 suggest that it takes two days for changes in Twitter 'calm'-ness to have a statistically significant impact on equity prices. That suggests that, if the findings are not spurious, that the Twitter 'calm'-ness effect is not analogous to that found in previous financial economics research on collective mood changes.

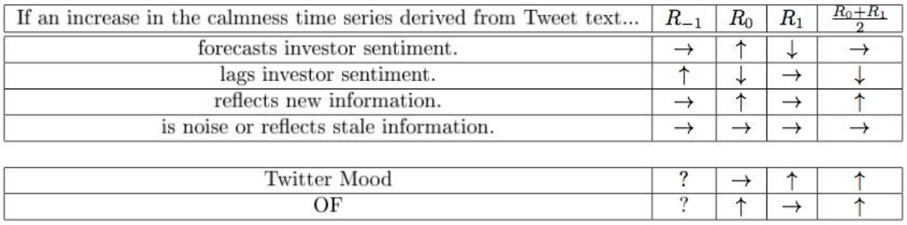


Lag	OF	Calm	Alert	Sure	Vital	Kind	Happy
1 Day	**0.085***	0.272	0.952	0.648	0.120	0.848	0.388
2 Days	0.268	**0.013****	0.973	0.811	0.369	0.991	0.7061
3 Days	0.436	**0.022****	0.981	0.349	0.418	0.991	0.723
4 Days	0.218	**0.030****	0.998	0.415	0.475	0.989	0.750
5 Days	0.300	**0.036****	0.989	0.544	0.553	0.996	0.173
6 Days	0.446	**0.065***	0.996	0.691	0.682	0.994	**0.081***
7 Days	0.620	0.157	0.999	0.381	0.713	0.999	0.150

* $p < 0.1$.
** $p < 0.05$.

Figure 2.3. This figure is a reproduction of Table 2 in TMP: "Statistical significance (p-values) of bivariate Granger-causality correlation between moods and DJIA in period February 28, 2008 to November 3, 2008." We have added the blue and red boxes around the results of interest. OF is an abbreviation BMZ use for the "OpinionFinder" tool.

TABLE 2.1.



If an increase in the calmness time series derived from Tweet text...	R_{-1}	R_0	R_1	$\frac{R_0+R_1}{2}$
forecasts investor sentiment.	→	↑	↓	→
lags investor sentiment.	↑	↓	→	↓
reflects new information.	→	↑	→	↑
is noise or reflects stale information.	→	→	→	→
Twitter Mood	?	→	↑	↑
OF	?	↑	→	↑

Notes: R_{-1} corresponds with returns from the period before a high calmness reading is recorded. R_0 corresponds with returns immediately after a high calmness reading is recorded and R_1 corresponds with a subsequent period. The →, ↑, and ↓ symbols signify average, high, and low returns respectively. Because the time series evaluated in TMP is so short, the statistical difference between differenced DJIA values D_t and returns R_t is not likely to explain the disagreements between TMP and the asset pricing literature. Notice that, in contrast with Twitter calmness, the OF-derived time series is consistent with the information interpretation of linguistic sentiment effects on asset prices given in Tetlock (2007).

The second major difference between TMP and the financial economics literature is that the statistically significant effects of collective mood shifts identified in TMP persist for several days. All research in the event study framework (reviewed in Appendix II) suggests that the effects of changes in collective mood effects on the aggregate market should vanish quickly, usually within 24 hours. In the news analytics literature, Tetlock et al. (2008), whose findings are representative of the literature, shows that the effect of new, firm-specific information on asset prices rapidly decays to zero and virtually all of the information in news reports is incorporated into asset prices within two days. Although Tetlock (2007) cannot reject the hypothesis that the effect of investor sentiment, at the weekly frequency, is zero, BMZ reject this hypothesis for their Twitter time series with p-value 0.036. In fact, the results BMZ obtain using their OpinionFinder tool (in the blue box in Figure 2.3) to measure Twitter sentiment, in both time-to-appear and persistence, are qualitatively more similar to those found in the financial economics literature for linguistic sentiment time series containing fundamental information than to the time series obtained with the GPOMS tool.

Overall, we find that in time-to-appear and persistence characteristics, Twitter mood's effects on asset prices as given by TMP are inconsistent with both the information interpretation of Twitter mood and the investor-sentiment interpretation of Twitter mood.[24]

BMZ's findings contra machine learning

Up-down sequence of the stock market indices is a benchmark rarely used to evaluate statistical models in financial econometrics, yet it is fairly common in the literature on textual analysis and machine learning as tools for stock market prediction.[25] Lachanski (2015) uses the survey of text mining for market prediction by Arman Nassirtoussi et al. (2014) to compile a list of the maximal predictive ability reported in several dozen studies. Lachanski (2015) finds that TMP's result implies higher price predictability than any published result in the news analytics or investor sentiment literatures (Loughran and McDonald 2016; Kearney and Liu 2014; Das 2014). Surprisingly, TMP also reports the highest level of stock market predictability of any study found in the literature on predicting the stock market using text analytics-based machine learning. There is no compelling reason to believe that the DJIA is especially predictable relative to the other financial instruments in these studies. The wide variety of financial instruments and methods used in the studies surveyed by Nassirtoussi et al. (2014) prevent us from conducting a formal meta-analysis, but the rank ordering of market predictability by study highlights TMP as the candidate most likely to be a statistical outlier.

[24] Curiously, Mao et al.'s follow-up work in 2015 recovers the co-movement pattern found in the standard investor sentiment-text analysis literature.

[25] Not without justification, much of the machine-learning work in text analytics for stock-market prediction is meant to be deployed in high-frequency trading strategies in which the primary criterion of success is the ability to predict the up-down tick sequence over a short time frame.

Persuasive critiques from the blogosphere and unpublished notes

BMZ's findings have generated controversy in the blogosphere. Zhichao Han (2012) points out that the mood normalization procedure in equation (1) takes in data from the future. BMZ appear to be aware that the forward-looking normalization procedure in equation (1) will introduce what they call "in-sample bias,"[26] writing that: "To avoid so-called 'in-sample' bias we do *not* apply *z*-score normalization to the mood and DJIA time series that are used to test the prediction accuracy of our Self-Organizing Fuzzy Neural Network in Section 2.5" (BMZ 2011, 3). Because of the ambiguity in the choice of *k*, we are unable to estimate the size of the bias that formula (1) induces. The normalization can be fatal for correct inference (as shown in Lachanski 2016 and below). There exist a number of available alternative formulas that would enable the interpretability of the raw mood time series without potentially inducing in-sample bias. Tetlock et al. (2008), wanting to ensure the stationarity of their sentiment time series, suggested the following normalization:

$$neg = \frac{X_t - X_{previous\ year}}{\sigma\left(X_{previous\ year}\right)}$$

where X_t is the output of a text-analysis algorithm applied to the news content for articles about a particular firm. This normalization would enable mood time series to be compared in terms of standard deviations above the previous year's mean for each mood. Instead of the problematic formula actually used in their paper, BMZ could simply have removed all future terms from $\mathbb{Z}_{X_t}$. Doing so would give the

[26] BMZ's meaning here of "in-sample" bias is unclear. We take their remark to refer to the fact that the Granger causality p-values are biased downward because BMZ's normalization procedure takes in data from the future, but this is not the standard use of the term. In-sample bias, or data snooping, typically occurs when the same sample is used for testing different models: Even if the null hypothesis is true, the probability of finding a significant p-value increases with the number of models tested. An alternative interpretation of BMZ, using the standard definition of in-sample bias, is that they worry their neural net tests and Granger causality tests will suffer from in-sample bias because they test a number of moods and the neural net's training period overlaps with the Granger-causality test period. But if this is what BMZ are referring to, then simply using the raw mood scores for the neural net tests will not eliminate the in-sample bias. A more appropriate solution to both problems would be to validate both the neural net and the ARDL models out-of-sample using the raw mood time series or, if those time series contain unit roots, the normalization proposed in Tetlock (2008).

following formula:

$$\mathbb{Z}^{*}_{X_t} = \frac{X_t - \overline{x}\left(X_{t,\, t-k}\right)}{\sigma\left(X_{t,\, t-k}\right)} \tag{7}$$

where $X_t - \overline{x}(X_{t,\, t-k})$ and $\sigma(X_{t,\, t-k})$ represent the mean and standard deviation of the time series within the period $[t{-}k, t]$ and k is arbitrary but fixed. This normalizes the mood time series to be expressed on a scale of one standard deviation (away from the current window's k-days mean) without being contaminated by information from after t. Volodymyr Kuleshov (2011) points out that, for Section 2.5 of TMP, normalizing both the training and test data to [0,1] before training the SOFNN introduces information from the future, but the size of the bias that this normalization induces is unclear.

Another criticism suggests that the implied SR is inconsistent with performance of the hedge fund Derwent Capital Markets. Pav (2012a) uses a Monte Carlo simulation to show that if the 86.7 percent accuracy in predicting the DJIA held for all time periods, then the strategy of timing the market using the techniques in TMP would have an SR greater than eight, far outside the range of published market timing strategies (Shen 2002). Lachanski (2015) shows that for realistic risk-free rates and transaction costs, the minimal SR obtained by such a strategy is greater than 5. Pav (2012b) points out that, given that Derwent Capital Markets—which had hired Bollen and Mao as consultants—shut down less than a year into operation, the market-predictability results in TMP are likely to have been misstated.

Another criticism suggests that BMZ draw incorrect conclusions from their non-linear statistics. In addition to the up-down accuracy, BMZ evaluate their SOFNN according to mean absolute percentage error (MAPE) defined as:

$$MAPE = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \left|\frac{A_t - F_t}{A_t}\right| \tag{8}$$

where T is the length of the time series, F_t is the forecast and A_t is the realized value. BMZ point out that using two time series, CALM and HAPPY, achieves the minimum MAPE (1.79 percent versus 1.83 percent for the CALM time series). Using the F-test, BMZ reject the hypothesis that 3 lags of the HAPPY mood time series contain additional information over just the 3 lags of the CALM time series. BMZ argue that these two facts together show that moods have a non-linear relationship with the DJIA.[27] But such reasoning is incorrect. First, the right-

[27] If TMP's mood time series has a non-linear relationship with the DJIA, then it is not analogous to the

hand side variables for both models are different (DJIA levels for the SOFNN tests, D_t for the regression tests). Second, BMZ never present the MAPE or up-down predictions for the linear models. Thus, the comparison is not a fair one because the models being compared are trained in-sample and penalized out-of-sample according to different error criteria. Third, BMZ do not show statistically significant improvements in MAPE for the HAPPY and CALM time series together over just the CALM time series.[28] BMZ cite the decrease in MAPE as evidence of HAPPY adding net information, but the up-down forecast accuracy using HAPPY and CALM decreases over just using CALM. Both small decreases in MAPE and the decrease in up-down accuracy are consistent with the HAPPY time series containing no information about stock prices. If the HAPPY time series is noise, all that can be concluded from BMZ's analysis is that the non-linear method obscures that fact while the nested ARDL model's F-statistic makes it obvious.

Did Twitter 'calm'-ness really predict the stock market?

Constructing Twitter mood time series

Following BMZ, we are concerned with tweets likely to contain explicit statements about the author's mood. We purchased the universe of tweets not containing URLs from Twitter matching the expressions 'I feel,' 'I am feeling,' 'I don't feel,' 'I'm,' 'Im,' 'I am,' and 'makes me' from January 1, 2007 to December 31, 2008. Server outages (and possibly user accounts and tweets deleted before we acquired our sample) have resulted in several days for which no tweets having these characteristics exist. For simplicity, we wish to work with a time series for which no data is missing at the daily level of aggregation. We start our sample from July 19, 2007, because it is the earliest that we can begin our time series analysis such that every day in our sample has at least one tweet that would have been

investor sentiment time series presented in Tetlock (2007), which Tetlock, using a semiparametric lowess method, finds to be approximately linear. This observation, by itself, does not rule out the information interpretation of the Twitter mood.

[28] BMZ put the MAPE of 1.79 percent in boldface and attach an asterisk, suggesting that this value is statistically significant at the 10 percent level. They present no statistical test indicating why this MAPE value is statistically significant (for comparison, BMZ do not bold the MAPE of 1.83 percent, suggesting this value is not significant at the 10 percent level). We suspect they do this simply to highlight the fact that this is the minimal MAPE achieved.

in BMZ's sample according to the filtering criterion above. Before this period, there are several days in our dataset for which there are no term-matches, likely because of both the small number of users at the time and Twitter server outages. That choice was made before conducting any time series analysis on the extended sample. It leaves us with 10,670,826 tweets in our dataset.

Our main objective in this section is to produce a plausible collective mood time series for bivariate analysis. We start with a Composed-Anxious lexicon ℓ that consists of a list of Composed and Anxious words. Our first Twitter mood time series is constructed by scoring each tweet, preprocessed using procedures described in Appendixes I and II, with the *sentimentr* R package. Our daily score is the mean score taken across all tweets with our sentiment scoring package.[29] Our second approach is to simply count the number of Composed and Anxious word matches from the dictionary and score each day's tweet with the following 'tone' function:

$$Twitter\_Tone_i = \frac{1}{T_i}\sum_{t=1}^{T_i}\left(\frac{C_t - A_t}{C_t + A_t}\right) \tag{9}$$

where day *i* receives score *Twitter_Tone*$_i$, and T_i is the number of tweets on day *i*. C_t and A_t are the number of matches between our Composed and Anxious lexicons, respectively, and tweet $t \in \{1,2, \ldots, T_i\}$. The functional form was selected because, in addition to its plausibility and usage in other research in financial applications using textual analysis, it produces a time series that appears to replicate several stylized features of the visualizations of the CALM series presented in TMP. We will call the mood time series generated via word count methods and the tone kernel X_t^{tone} and the time series generated by the *sentimentr* algorithm X_t^s.

BMZ's contention that their selection technique, and thus mood time series, will contain little to no information about fundamentals appears correct. Hand inspection of 3,000 tweets found no DJIA tickers, S&P 500 tickers, or anything that might conceivably be related to the fundamental value of an equity or equity mispricing. Our mood dictionary construction procedures and the characteristics of the Twitter mood time series we construct are described in Appendix II. We present both mood time series in Figure 3.1 and the summary statistics of these time series in Table 3.1.

[29] Our median and modal user each day uses zero mood terms from our lexicons. Thus, if we were to use either of these as a measure of central tendency we would have no variation in our mood time series and trivially reject the hypothesis that collective mood Granger-causes stock-market increases.

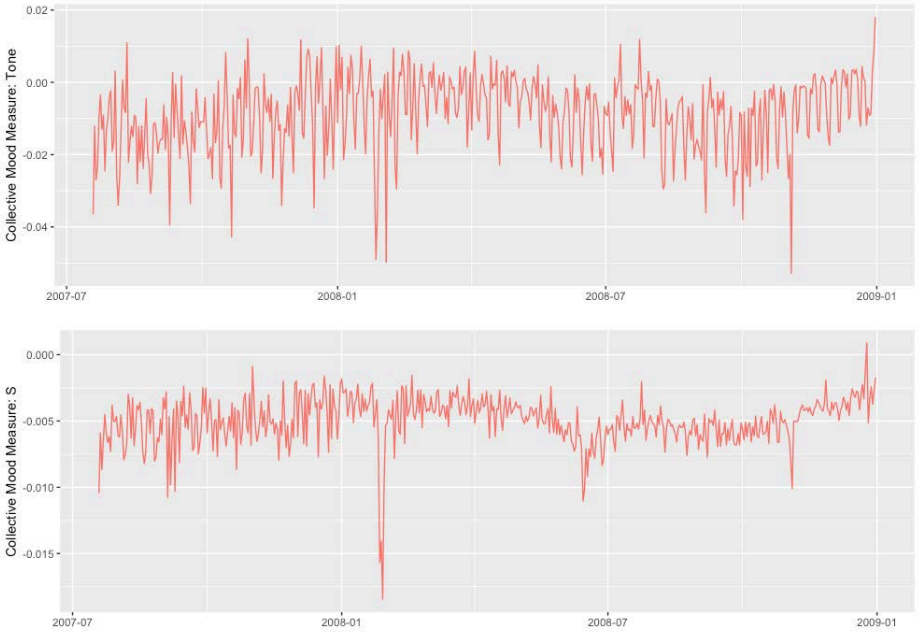


Figure 3.1. This figure contains our proxy variables for BMZ's CALM consisting of the tone (top panel) and *sentimentr* (bottom panel) generated time series from July 19, 2007, to December 31, 2008. There is no obvious unit root in either time series and they have a correlation of 0.61.

TABLE 3.1. Summary statistics for our mood time series reported to three decimal places

	X_t^s	X_t^{tone}	$\mathbb{Z}_{X_t}^{k=1}$	$\mathbb{Z}_{X_t}^{k=10}$
Minimum	−0.018	−0.053	−1.155	−3.275
1st Quartile	−0.006	−0.015	−0.697	−0.657
Median	−0.005	−0.007	0.045	0.202
3rd Quartile	−0.004	−0.001	0.674	0.725
Maximum	0.001	0.018	1.154	2.331
Mean	−0.005	−0.008	0.009	−0.001
Standard Deviation	0.002	0.011	0.748	0.971
Augmented Dickey Fuller Test p-value	< 0.01	< 0.01	< 0.01	< 0.01

Notes: The local normalization is applied to the X_t^{tone} because this series looks most similar to that presented in TMP. $\mathbb{Z}_{X_t}^{k=n}$ refers to the series X_t normalized accordingly with parameter k set to n. k=1 is chosen to illustrate the effects of local normalization and k=10 is chosen because this value yields a normalized DJIA difference series visually similar to that presented in TMP. We have 532 observations in our sample. We find no evidence of a unit root in our mood time series.

Like BMZ, we are interested in testing the ability of our collective mood time series to forecast the DJIA. We take the DJIA series that corresponds to our Twitter mood time series from July 19, 2007 to December 31, 2008 as our primitive, visualized in Figure 3.2.[30] Our main statistical exercises will treat July 19, 2007 to November 3, 2008 as the training set with the remainder of the data as our test set. We will see that, as usual, because of unit roots in our DJIA series, we cannot obtain unbiased estimators from our ARDL models without first differencing.[31] In Table 3.2 we provide unit root tests, summary statistics, and other univariate characterizations of the DJIA and DJIA difference distributions over this time frame and the time frame used by BMZ.

TABLE 3.2. Summary statistics for DJIA levels and DJIA differences

	Full sample		TMP subsample		Test set	
Series	$DJIA_t$	$\Delta DJIA_t$	$DJIA_t$	$\Delta DJIA_t$	$DJIA_t$	$\Delta DJIA_t$
Minimum	7552.29	−777.68	8175.77	−777.68	7552.29	−679.95
1st Quartile	11348.04	−129.07	11231.04	−127.28	8438.72	−216.43
Median	12391.57	−3.16	11724.75	−3.17	8585.40	5.10
3rd Quartile	13216.75	94.81	12531.26	91.27	8765.16	200.13
Maximum	14164.53	936.42	14164.53	936.42	9625.28	552.58
Mean	11941.73	−13.97	12350.91	−14.02	8586.47	−13.59
Standard Deviation	1657.88	211.24	1157.61	200.24	336.96	289.36
Augmented Dickey Fuller Test p-value	0.64	< 0.01	0.59	< 0.01	0.23	0.07
# Observations	368	368	178	178	40	40

Note: Statistics are for our full sample (July 19, 2007 to December 31, 2008), BMZ's linear test subsample (February 28, 2008 to November 3, 2008), and the test set we will use for comparing the predictive power of linear and non-linear models (November 4, 2008 to December 31, 2008).

In Table 3.2, we present the relevant summary statistics for DJIA levels and differences over the time period investigated. The Augmented Dickey-Fuller test (ADF hereafter) p-values indicate that we have a unit root in DJIA levels. There might be difficulties extrapolating from this time series to other time periods. For one, the three greatest and two least $\Delta DJIA_t$ values over the whole history of the DJIA occur over the time period studied in TMP. Four of the top ten largest up-or-down movements in DJIA history occur in TMP's linear causality tests. More importantly, one of the stylized facts of DJIA levels, positive trend, is rejected by our statistical analysis over this time period.

[30] We obtain the DJIA series from the *quantmod* package in R, which in turn, obtains the series from Yahoo! Finance.

[31] Our own preference for the functional form of the regression would be to take the first difference of logged DJIA levels so that we could interpret our regressions' left-hand sides as log-returns. This is not the approach taken by BMZ, and so for simplicity we present Granger-causality results only for level differences.

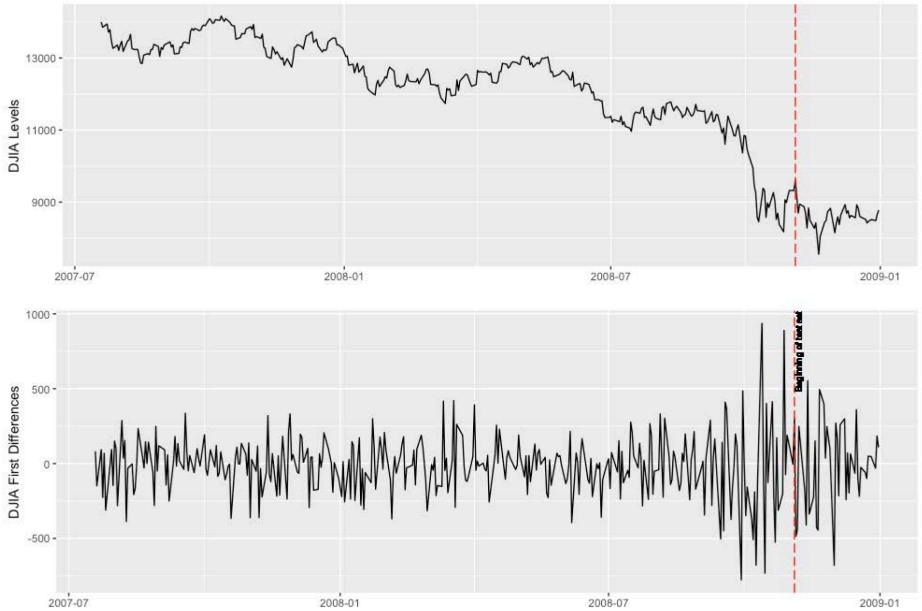


Figure 3.2. This figure contains our DJIA and differenced DJIA time series, July 19, 2007, to December 31, 2008. Our level and difference series contains 368 days (observations). The DJIA level time series contains an obvious unit root. Our full sample training (in-sample) tests are conducted on all the data points to the left of the red dotted line while our out-of-sample set consists of all the points to the right of the red dotted line. For reference, our training set includes TMP's training set (February 28, 2008 to November 3, 2008) as a subsample while our test set includes TMP's test set (December 1, 2008 to December 19, 2008) as a subsample. Thus, if our collective mood time series effectively approximates the CALM time series provided by the GPOMS tool, then our evaluation is a more powerful test of the Twitter mood effect than the in-sample and out-of-sample tests provided in TMP.

DJIA differences are heavy-tailed but predictable

BMZ write that they chose to use December 1, 2008 to December 19, 2008 as their test set for the SOFNN because "it was characterized by the stabilization of DJIA values after considerable volatility in previous months and the absence of any unusual or significant socio-cultural events" (p. 6). BMZ's characterization of the stock market over this 15-day period is accurate, but it is difficult to see how their findings are supposed to generalize to other periods. We display the Q-Q plot of the DJIA difference series analyzed in BMZ's Granger causality analysis, the DJIA difference series from July 31, 2000 to December 31, 2014, BMZ's test set, and our test set in Figure 3.3. Our Q-Q plot shows that BMZ's Granger causality analysis set, like the long-run DJIA difference series, covers a period with heavy-tailed DJIA differences, but that neither our test set nor BMZ's test set has heavy tails. The absence of heavy tails is important because it shows that:

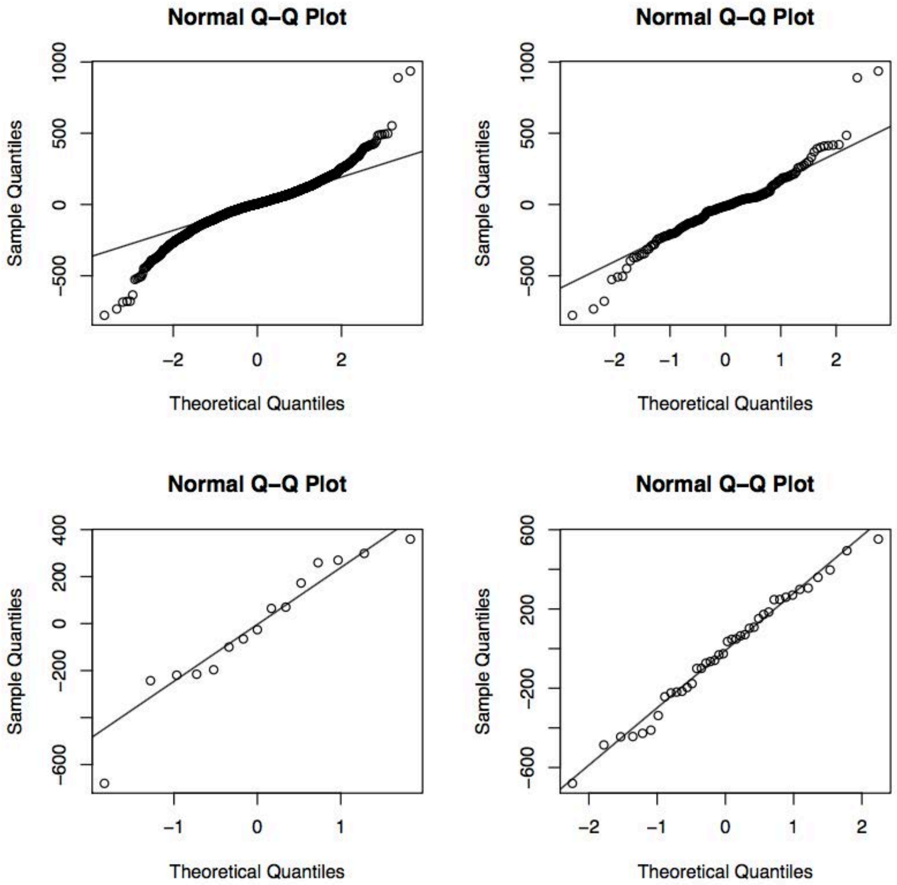


Figure 3.3. This Q-Q plot, clockwise from the top left panel, shows the 'long-run' DJIA difference series from July 31, 2000 to December 31, 2014; the difference series on which BMZ conducted their Granger causality analysis from February 28, 2008 to November 3, 2008; BMZ's test set; and our test set. Notice that the long-run DJIA series has heavy right and left tails (since the rightmost points on the curve are above the line characterizing the normal quantiles and the leftmost points on the curve are below the line characterizing the normal quantiles) but BMZ's test set does not. This can also be seen from Table 3.2 in which it is reported the DJIA difference series has negative excess kurtosis over a timeframe including the test set used in TMP.

1. Heavy-tailed returns, one of the stylized facts of the equity market return series, carry over to differenced prices as well and because:
2. BMZ's test set findings, with regards to their reported MAPE, are unlikely to generalize: As the daily movements of the stock market grow large, the average incorrect guess of the direction of the market adds more to the absolute prediction error. By selecting a period in which the tails of the equity market are lighter than average, BMZ minimize their reported MAPE.

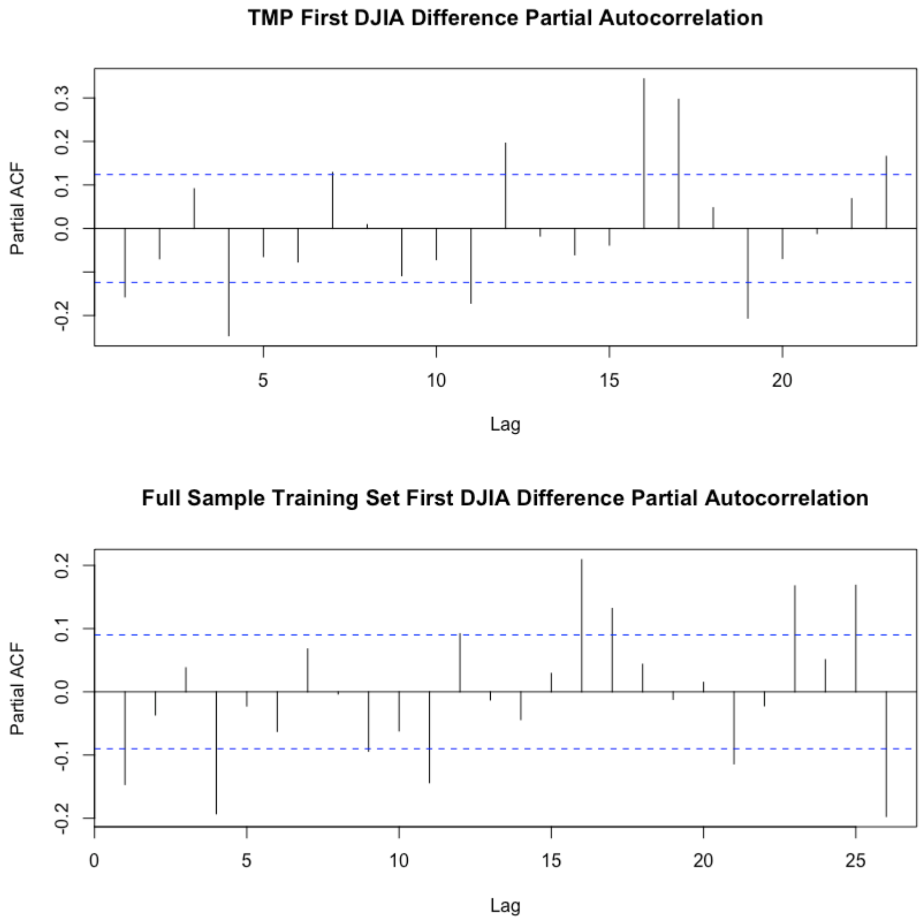


Figure 3.4. This figure shows the PACF plots for the first difference of the DJIA for the time period studied in Section 2.4 of TMP and, as a sanity check, in our own training set for the full sample. The blue dotted lines show 5 percent statistical significance bounds. Since the first and fourth lags exceed these bounds, we have statistically significant evidence in both datasets that the differenced DJIA time series is not i.i.d.

Several researchers have noted[32] that if the DJIA differences are independent and identically distributed (i.i.d.), then a single lower-order collective mood lag's significance can result in higher-order lags in TMP also appearing statistically significant, but with p-values declining in the order of the lag. In fact, as shown in Figure 3.4, the DJIA differences are not i.i.d. In Figure 3.4 we plot the partial autocorrelation function (PACF) over TMP's Granger causality analysis and our

[32] The point made in this sentence was related to us during a presentation by Lachanski at the Financial Management Association Annual Meeting in 2014.

training sets. If the correct statistical model of $\Delta DJIA_t$ is given by an AR model, then the statistically significant values returned by the PACF will correspond with statistically significant lags in that AR model. Even if the correct model is not in the AR family, significant lags confirm that the DJIA differences are not independent. Our plots show that the first two lags of the differenced DJIA series are statistically significant at the 5 percent level and so we reject such explanations for the p-value pattern found in TMP.

Parametric tests for the effect of collective mood on the DJIA

We use two statistical models and the mood time series X_t^{tone} and X_t^s to answer the questions originally posed by BMZ about the joint distribution of collective mood and the DJIA. First, we use ARDL models to assess whether or not collective mood changes linearly predict the stock market. We find no Twitter-mood effect in our extended sample. Five was chosen as the maximum lag length for our tests because TMP does not report statistically significant results for lag lengths greater than five. In Table 3.3 we display the minimal p-values for each model across different choices with respect to how moods were treated over weekends.

The lowest p-value occurs for an ARDL(5,5) model, i.e., the model with 5 $\Delta DJIA_t$ lags and 5 X_t^{tone} lags (taking the minimum value across weekends, holidays, and the last day of trading and assigning that to the last day of trading). This lowest-p-value finding increases our suspicion that the start date was chosen to minimize reported p-values. If our time series approximates their time series, then extending their time series actually lowers their reported p-values.[33] For the remainder of this section we are concerned only with the predictive power of this raw mood time series X_t^{tone}. None of the models we estimated reproduce the same p-value pattern shown in TMP. Overall, our in-sample evidence suggests that our measure of Twitter mood, constructed over the same time period as TMP's measure of Twitter mood and selected for its visual resemblance to the time series shown in TMP, has little ability to predict the stock market outside of the sample chosen by BMZ.

[33] Of the 1,155 models estimated in this section, the Akaike Information Criterion, but not the Bayesian Information Criterion, also selected the same lag length, the same data-cleaning procedure and the same start date as we did using the minimal p-value.

TABLE 3.3. P-values from t-tests and Granger-causality F-tests

Full sample: July 19, 2007 to November 3, 2008; Response variable: $\Delta DJIA_t$					
Test statistic	t-test	F-test	F-test	F-test	F-test
Lag Length	$\Delta DJIA_t$	X_t^s	X_t^{tone}	$Z_{X_t}^{k=1}$	$Z_{X_t}^{k=10}$
1	**0.005**	0.52	0.196	0.122	0.367
2	**0.024**	0.542	0.161	0.171	0.071
3	0.392	0.606	0.204	0.294	0.132
4	0.321	0.727	0.294	0.379	0.161
5	0.328	0.277	0.126	0.362	0.097
TMP subsample: February 28, 2008 to November 3, 2008; Response variable: $\Delta DJIA_t$					
Test statistic	t-test	F-test	F-test	F-test	F-test
Lag Length	$\Delta DJIA_t$	X_t^s	X_t^{tone}	$Z_{X_t}^{k=1}$	$Z_{X_t}^{k=10}$
1	**0.027**	0.327	0.078	0.49	0.5
2	**0.013**	0.381	0.084	0.694	0.658
3	0.389	0.069	0.054	0.549	0.343
4	0.561	0.107	0.075	0.694	0.399
5	0.205	0.113	**0.029**	0.699	0.156

Notes: Our first column of results consists of the p-values from t-tests performed on the coefficients of a single ARDL(5,0) which contains no mood covariates and five $\Delta DJIA_t$ lags. The remaining columns present the p-values from Granger causality F-tests of ARDL(i,0) models nested in ARDL(i,i) models where i is the lag length. This F-test corresponds to a test of condition (5) versus (6) in the nested models (3) and (4). The p-values shown have not been adjusted for multiple comparison bias. Each F-test's cell reports the minimum p-values across different ways of dealing utilizing mood time series data generated during non-trading days like weekends and holidays (see Appendix II for details) so as to bias these results in favor of lower p-values. Recall that we can (i) simply drop this additional mood data, (ii) assign the maximum daily mood value across contiguous non-trading days and the last trading day to the last trading day, or (iii) assign the minimum value across contiguous non-trading days and the last trading day to the last trading day. In general, (iii) generates lower p-values than (ii), which generates lower p-values than (i). Note that the missing data procedure for our mood time series giving the minimum p-value overall, taking the minimum value for mood across weekends, holidays, and the last day of trading before a weekend or holiday, results in X_t^{tone} having insignificant p-values for the first two lags. Simply dropping weekend mood values results in the first lagged p-value of X_t^{tone} being significant while the third lag is not. Thus, no single explanatory variable can reproduce the p-value pattern presented in TMP.

We produce the coefficients of X_t^{tone} for our model giving the lowest single p-value in Table 3.4. The coefficients accord with BMZ's description of the 'calm'-ness effect, but the equilibrium interpretation of this model is problematic for reasons given in our literature review. Furthermore, the sign of the coefficients for the models we estimate as significant in-sample contradicts the signs discovered in models estimated out-of-sample (as in Lachanski 2014). Finally, the signs of the effect of Twitter mood switch with the order of the lag, making it difficult to give any simple economic interpretation of the model.

TABLE 3.4. Coefficients of our ARDL model with five lags estimated on BMZ's subsample

Lag	Coefficient value	P-value (for individual coefficient)
1	1521.34	0.422
2	−693.83	0.719
3	4659.88	**0.015**
4	−877.50	0.652
5	4062.55	**0.037**

Notes: The p-value shows that collective significance is driven entirely by the third and fifth lags, but not the second as in TMP. An increase in 'calm'-ness five days ago, according to this model, will have a positive effect on the DJIA. According to this model, decrease of two standard deviations in calmness five days ago, for time periods matched on all other characteristics, forecasts a 73-point decrease in the DJIA today.

BMZ's normalization revisited

BMZ's detrending method pulls information from the future into present CALM values; that means that detrended Twitter mood's ability to forecast DJIA differences on days *t*+1 and *t*+2 likely arises in part from the day *t* or *t*−1's DJIA's ability to forecast Twitter mood on day *t*. Because the local normalization takes in data from the future, using a linear time-series analysis on a locally normalized mood time series may yield statistically significant results even when the underlying mood time series does not forecast the stock market.

First, we consider the case where the DJIA differences $\Delta DJIA_t$ are i.i.d. standard normal and the daily CALM score X_t is linear in the DJIA difference with some noise.

$$X_t = \beta \Delta DJIA_t + \epsilon_t, \tag{10}$$

where the ϵ_t are white noise. As in TMP, we then normalize each series with a centered window of size k, defining:

$$\mathbb{Z}_{X_t} = \frac{X_t - \frac{1}{2k+1}\Sigma_{|s-t| \leq k} X_s}{\sqrt{\frac{1}{2k}\left(\Sigma_{|u-t| \leq k}\left(X_u - \frac{1}{2k+1}\Sigma_{|s-t| \leq k} X_s\right)^2\right)}}$$

Let the numerator be the 'locally centered' mood:

$$\mathbb{C}_{X_t} = X_t - \frac{1}{2k+1}\sum_{|s-t| \leq k} X_s$$

It turns out that analyzing the numerator will be sufficient to capture the reverse causality we would like to describe. Consider the case where $\mathbb{C}_{X_t}$ appears to 'predict' $\Delta DJIA_u$ for $t < u \leq t+k$ even though the causality is reversed between X_t and $\Delta DJIA_t$. The correlation between these is:

$$\rho\left(\mathbb{C}_{X_t}, \Delta DJIA_u\right) = \frac{-\frac{\beta}{2k+1}}{\sqrt{\left(\beta^2 + \sigma^2\right)\left(\frac{2k}{2k+1}\right)}} \approx -\frac{\beta}{\sqrt{\beta^2 + \sigma^2}}\frac{1}{2k+1} \tag{11}$$

where σ^2 is the variance of ϵ_t.

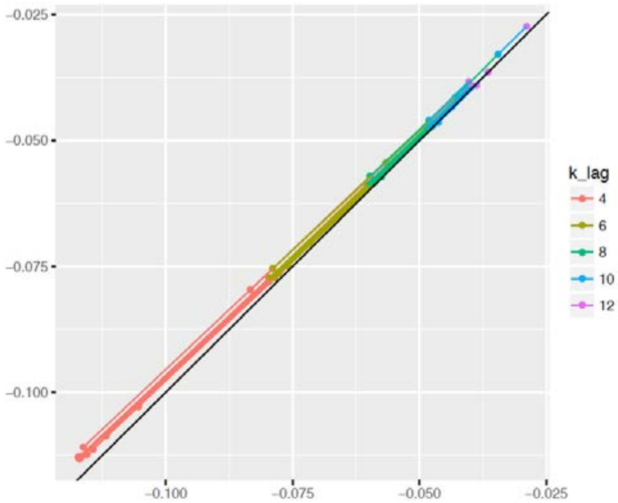


Figure 3.5. For β ranging from 1 to 10, and for $k = 4$, simulations consisting of one million realizations of $\Delta DJIA_t$ and X_t were drawn. The empirical correlation of $\mathbb{Z}_{X_t}$ and $\Delta DJIA_{t+1}$ was computed, and is plotted against the theoretical correlation between $\mathbb{C}_{X_t}$ and $\Delta DJIA_{t+1}$. The different lines represent different values of β. We fix $\sigma^2 = 1$.

To test whether the correlation between $\mathbb{Z}_{X_t}$ and $\Delta DJIA_u$ is equal to that between $\mathbb{C}_{X_t}$ and $\Delta DJIA_u$, we perform Monte Carlo simulations. For β taking integer values from 1 to 10, and for k taking values 4, 6, 8, 10, and 12, we compute a series of normally distributed $\Delta DJIA_t$ of length 1,000,000 and a dependent series of X_t with $\sigma^2 = 1$ and ϵ_t normally distributed. We compute the correlation $\rho(\mathbb{Z}_{X_t}, \Delta DJIA_{t+1})$ and compare it to the value of $\rho(\mathbb{C}_{X_t}, \Delta DJIA_u)$ given in Equation 5. Correlations $\rho(\mathbb{Z}_{X_t}, \Delta DJIA_{t+1})$ and $\rho(\mathbb{C}_{X_t}, \Delta DJIA_u)$ are plotted against each other in Figure 3.5, with one line for each value of β. We see good agreement, and although the empirical correlation is somewhat higher (though still negative) for small values of β than suggested by Equation 1, we still see significant negative correlation for Z-scored values even though the causality is broken.

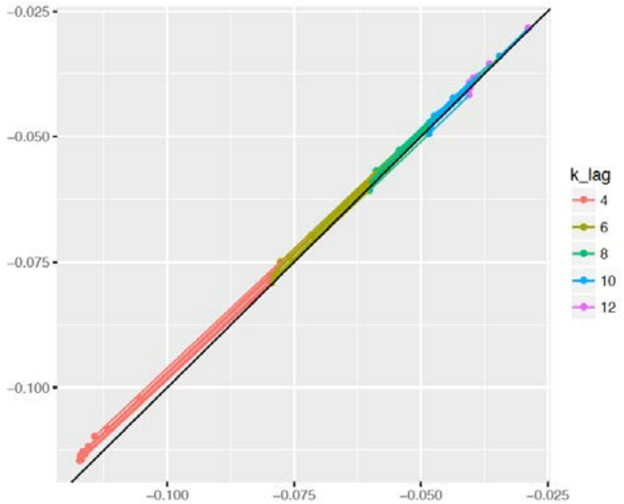


Figure 3.6. For ranging from 1 to 10, and for $k = 4$, simulations consisting of one million realizations of $\Delta DJIA_t$ and X_t were drawn, with innovations drawn from a t-distribution with 4 degrees of freedom. The empirical correlation of $\mathbb{Z}_{X_t}$ and $\Delta DJIA_{t+1}$ was computed, and is plotted against the theoretical correlation between $\mathbb{C}_{X_t}$ and $\Delta DJIA_{t+1}$. The different lines represent different values of β. We fix $\sigma^2 = 1$.

To test whether the quality of the approximation in Equation 11 is robust to distributional shape, we repeat the experiment above but draw ϵ_t i.i.d. from a t-distribution with 4 degrees of freedom, plotting the empirical correlation $\rho(\mathbb{Z}_{X_t}, \Delta DJIA_{t+1})$ against $\rho(\mathbb{C}_{X_t}, \Delta DJIA_{t+1})$, as given in Equation 11, in Figure 3.6. There is effectively little difference between this plot and Figure 3.5, indicating the approximation is robust against mild kurtosis. This experiment shows how BMZ's local normalization procedure can reverse causality between the underlying CALM series and DJIA differences, due entirely to it using leading data to define the mean for centering X_t. We believe that this may have contaminated the linear test results BMZ present in Section 2.4 of their paper, but our $\mathbb{Z}_{X_t^{tone}}$ time series did not Granger-cause DJIA differences. To test whether the DJIA contains information about the future content of Twitter mood, we run the regression suggested by equation (10) as well as the regression of Twitter mood on itself and lagged DJIA variables and present the p-values of our coefficients in Table 3.5. While the results in column 1 suggest that the Granger-causality reversal mechanism we present may explain the results in TMP, without access to the parameter *k* that BMZ used to normalize their underlying mood time series[34] we cannot further explore the possibility. Our last two columns' p-values suggest that (10) does not hold for our Composed-Anxious time series, and so any reverse Granger-causality mechanism at work must necessary be more complex than those we explore here.

TABLE 3.5. P-values from a series of regressions of X_t^{tone} estimated on $\Delta DJIA_t^{min}$ in our full sample and BMZ's subsample

Full sample: July 19, 2007 to November 3, 2008				
Lag Length	$\Delta DJIA_t^{drop}$	$\Delta DJIA_t^{min}$	$\Delta DJIA_t^{drop}$	$\Delta DJIA_t^{min}$
1	0.599	0.941	0.908	0.908
2	0.144	0.564	0.346	0.347
3	0.098	0.703	0.580	0.581
4	0.279	0.851	0.974	0.974
5	0.343	0.378	0.177	0.177
TMP subsample: February 28, 2008 to November 3, 2008				
Lag Length	$\Delta DJIA_t^{drop}$	$\Delta DJIA_t^{min}$	$\Delta DJIA_t^{drop}$	$\Delta DJIA_t^{min}$
1	0.911	0.972	0.228	0.466
2	**0.026**	0.911	0.401	0.674
3	**0.016**	0.888	0.018	0.147
4	**0.031**	0.968	0.511	0.377
5	**0.043**	0.519	0.151	0.444
Tone included	Yes	Yes	No	No

Notes: Our first two columns report the p-values of nested ARDL regressions tests with response variable X_t^{tone} and including an equal number of lags of X_t^{tone} and $\Delta DJIA_t$ so that the p-values reflect the additional information content of $\Delta DJIA_t$. For $\Delta DJIA_t^{drop}$, we simply dropped weekends and holiday values of our mood series from the dataset. For $\Delta DJIA_t^{min}$, we assigned the last trading day's mood series value before a weekend or holiday to the minimum across that trading day, weekend, or holiday. Our second two columns report the p-values of a t-test applied to a regression of X_t^{tone} on lagged DJIA differences that does not include lagged X_t^{tone} tone terms. Thus, a lag length of 1 for these columns corresponds to a test of specification (10). These regressions show that the joint distribution of our DJIA differences and mood is sensitive to the researcher's choice of how missing data is handled.

[34] In particular, we are worried that that parameter *k* used to normalize the CALM series may not take on integer values, and so the visual comparison exercises we conducted for the DJIA would be prohibitively time-consuming to apply to our mood time series.

Nonparametric models for emotion-based stock prediction

To evaluate BMZ's claim that the relationship between collective mood and the DJIA is non-linear, we re-estimate our 5-lagged ARDL model using all available training data.

To capture any non-linearities, we use projection pursuit regression to fit a nonparametric model to our bivariate time series of DJIA differences and X_t^{tone}. Then we conduct an out-of-sample experiment in which we compare the ability of our fitted model and the best ARDL model to predict the up-down pattern of the DJIA. We also compare the out-of-sample MAPE and root mean squared error (RMSE) of each model.

We chose the projection pursuit regression for three reasons. First, our ARDL regression chose eleven parameters and we have only 174 days in BMZ's training set. Kernel regression and even semi-parametric spline regression methods can fall victim to the curse of dimensionality, in which our model gives poor estimates for variable combinations that it has never seen before yet are likely to occur in this context.[35] Projection pursuit was designed to fight the curse of dimensionality by first reducing a p-dimensional explanatory variable to a one-dimensional space. It does so by estimating vectors $\mathbf{a}_j$ and projecting our explanatory variables onto the space spanned by these vectors, then fitting ridge functions φ_j onto the projected explanatory variables. Specifically, projection pursuit regression estimates the function

$$\varphi(x) = \alpha + \sum_{j=1}^{m} \phi_j(\mathbf{a}_j \cdot \mathbf{x}),$$

where we use bold to indicate that our explanatory variable $\mathbf{x}$ is a vector. At the same time, like many nonparametric techniques, projection pursuit regression can (in theory) fit arbitrary continuous functions. That makes it a natural choice to discover any non-linearity or lack thereof in the joint distribution of Twitter mood and DJIA differences.[36]

The second reason we chose projection pursuit regression is that it is well understood. Since its development in 1974 it has been successfully used in a wide range of financial econometric and asset pricing problems, including option pricing, high frequency trading, and risk analysis, though as far as we know it has not

[35] We have eleven variables that correspond to five lagged DJIA differences, five mood scores, and a constant. In financial settings, nonparametric methods not designed to address the curse of dimensionality typically do very poorly out-of-sample (Carmona 2013).

[36] In practice of course, projection pursuit regression is computationally limited to a much smaller class of easily implemented 'super smoother' functions.

yet been applied to text-derived variables in an asset pricing context. Projection pursuit regressions are often used in these types of horseraces to test for nonlinearities in the joint distribution of equity market variables and potential predictors.

Third, unlike the SOFNN results presented in TMP, our findings are unlikely to be caused by the optimal selection of hyperparameters. Projection pursuit, in theory, has no hyperparameters and so there is no possibility of these hyperparameters being chosen to maximize test set performance. In practice, we must specify a minimal number of ridge functions to be fit, but the selection process is transparent and small changes in our ridge function parameters are unlikely to significantly affect the outcome of tests performed using projection pursuit.

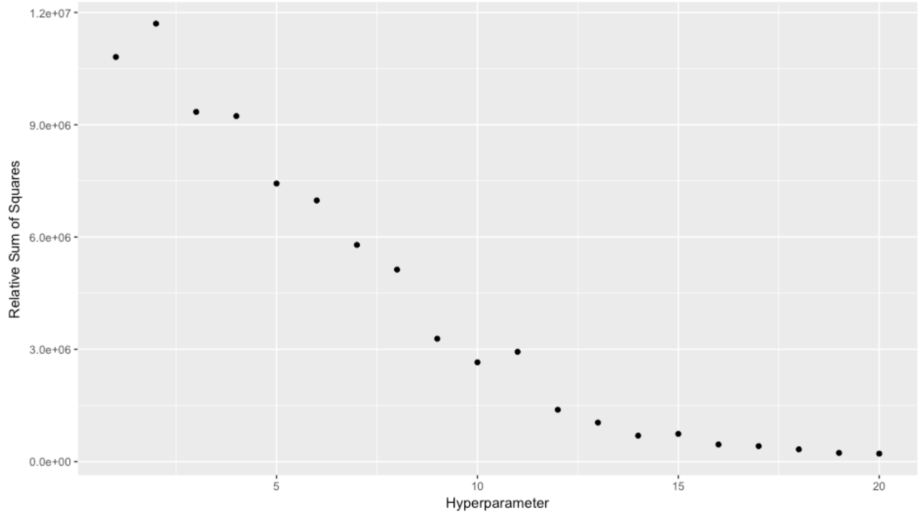


Figure 3.7. This figure gives our relative sum of squares, a measure analogous to the residual sum of squares in linear regression. Inflection points near local minima are often recommended as choices for the hyperparameter. We choose a value of 10 for our hyperparameter. Choosing 9 or 11 gave similar results.

Fitting a projection pursuit regression involves choosing one hyperparameter that determines model complexity. In Figure 3.7, we produce the figure of merit giving our 'relative sum of squares' as a function of the number of the minimal number ridge functions in our projection pursuit regression trained over the training set. It is typically recommended that one choose an inflection point close to the minimum value of the relative sum of squares. The largest drops in our relative sum of squares measure occur for 5, 7, 9, 10, and 14. Because we are concerned with optimal out-of-sample predictive power, it is typically recommended to err on the side of parsimony. Thus, we choose 10 as our hyperparameter for this exercise. Qualitatively similar results are found for other choices of our hyperparameter.

While our ARDL models are estimated through least squares, our projection pursuit regression is fit using a complex cost function not directly comparable to least squares. Therefore, we compare the performance of our non-linear algorithm to linear regression using three metrics: up-down predictability, MAPE, and RMSE.[37] In our case, the RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{40}\sum_{t=1}^{40}\left(\Delta DJIA_t - \Delta\widehat{DJIA}_t\right)^2}$$

where $\Delta\widehat{DJIA}_t$ are the predictions made by one of our models. Better models will have lower RMSE scores. The RMSE should favor the ARDL model because it was trained under a least-squares criterion, but neither MAPE nor up-down predictability should favor one model or the other a priori. To better identify the marginal predictive power added by Twitter mood rather than our choice of algorithm, for both simulations we fit a baseline five-lag ARDL and projection pursuit model. Unlike BMZ's test of non-linearity, we fit our models to the same left-hand side variables. Specifically, we fit

$$\Delta DJIA_t = a_0 + \sum_{j=1}^{m} \phi_j\left(\left(\Delta DJIA_{t-1}, \ldots, \Delta DJIA_{t-5}, X_{t-1}, \ldots, X_{t-5}\right)\left(a_1, \ldots, a_5, a_6, \ldots, a_{10}\right)^T\right), \tag{12}$$

where $\left(a_1, \ldots, a_{10}\right)$ must be estimated from our training set and ϕ_j is the ridge function (defined in Carmona 2013). For our baseline function, we fit

$$\Delta DJIA_t = a_0 + \sum_{j=1}^{m} \phi_j\left(\left(\Delta DJIA_{t-1}, \ldots, \Delta DJIA_{t-5}\right)\left(a_1, \ldots, a_5\right)^T\right) \tag{13}$$

where all parameter values take on the usual meanings. Using a variety of metrics, applied to both models out-of-sample in order to assess the linearity or non-linearity of Twitter mood, addresses the critiques made on pages 321–322. In Figures 3.8 and 3.9, we visualize the outcomes from our first simulation. In the first simulation, which replicates the approach taken in TMP, our model is estimated once using the full training set and parameters are fixed throughout the testing period. We then plot the predictions made by our model against the realized DJIA differences and use the RMSE, MAPE, and up-down forecast accuracy to assess

[37] Our projection pursuit algorithm, in its typical presentation, does not come with a finite dimensional likelihood and so we cannot effectively compare the in-sample performance. The residual sum of squares for nonparametric, non-linear functions is generally lower than those obtained from linear models, and that was also the case for our projection pursuit regression.

our models in Table 3.6.

Twitter mood did not predict the stock market

The results in Table 3.6's first four rows show that, using BMZ's testing strategy, we have no evidence that Twitter mood predicts the stock market. We conduct two-sample nonparametric chi-squared tests[38] on our up-down accuracy scores to assess whether or not the proportion correctly predicted increased when we incorporated Twitter mood into our models. Taking Twitter mood as our treatment, we find that for none of the four paired samples is there outperformance that is statistically significant at any standard level. Although there are no formal hypothesis tests associated with our RMSE and MAPE criterion, both show the kinds of ambiguities we might expect if our Twitter mood proxies were noise. For our linear models estimated over the full sample, incorporating Twitter mood improved all three metrics, but the improvements are not economically significant. For our nonlinear fitted models and linear fitted models within TMP's subsample, different criteria give different results as to whether including Twitter mood improves accuracy or not. Within BMZ's subsample, Twitter mood improves both our models unambiguously, but even with Twitter mood it is the poorest performing model among those we test. For our linear parametric model, we observe no improvement in the ability to forecast up-down accuracy and no economically significant improvement in our RMSE or MAPE scores when Twitter mood is incorporated into the forecast.

TABLE 3.6. Data corresponding to our simulation results shown in Figures 3.8 through 3.9

Model	Training set	RMSE	MAPE	Up-down accuracy
Linear parametric without collective mood	Full sample	273.50	106.21%	60.0%
Linear parametric with collective mood	Full sample	272.74	103.52%	62.5%
Non-linear nonparametric without collective mood	Full sample	331.13	145.87%	**67.5%**
Non-linear nonparametric with collective mood	Full sample	295.71	128.86%	57.5%
Linear parametric without collective mood	TMP subsample	273.25	112.49%	60.0%
Linear parametric with collective mood	TMP subsample	264.28	112.25%	60.0%
Non-linear nonparametric without collective mood	TMP subsample	440.53	243.22%	50.0%
Non-linear nonparametric with collective mood	TMP subsample	335.46	164.40%	55.0%

Notes: We have bolded up-down accuracy scores which violate the random walk hypothesis with 95 percent confidence. While not a statistically significant violation of the random walk hypothesis, even the 60 percent predictability result would violate good deal bounds by the reasoning in Lachanski (2015) if it held across all time periods. Notice however, that none of our models effectively forecast the magnitude of market movements. Even the best models' accuracy is off by approximately the value of a daily market movement. The TMP subsample models were estimated with 7 hyperparameters by the reasoning illustrated in in Figure 3.7.

[38] Parametric Z-score tests give similar results.

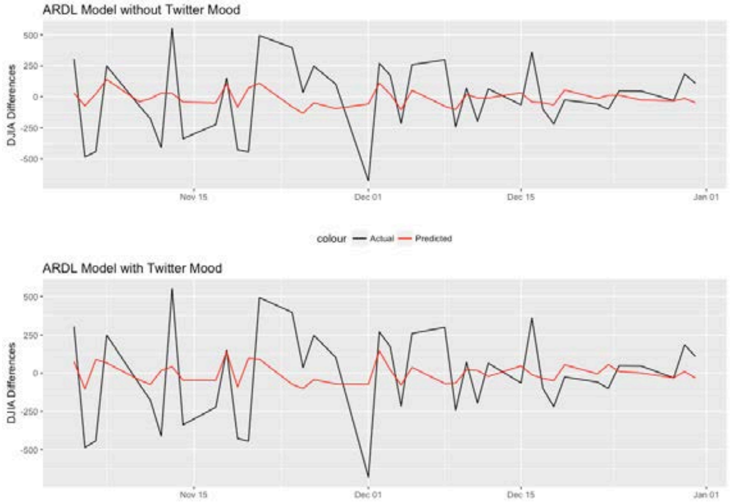


Figure 3.8. This plot shows $\Delta\widehat{DJIA}_t$ as predicted each day by our ARDL(5,0) model, which does not incorporate Twitter mood, against the realized $\Delta DJIA_t$ in the top panel. The bottom panel shows $\Delta\widehat{DJIA}_t$ as predicted each day by our ARDL(5,5) model, which does incorporate Twitter mood, against the realized $\Delta DJIA_t$. Both models were estimated using the full training set. Visually, there is no compelling evidence of outperformance by the model incorporating Twitter mood.

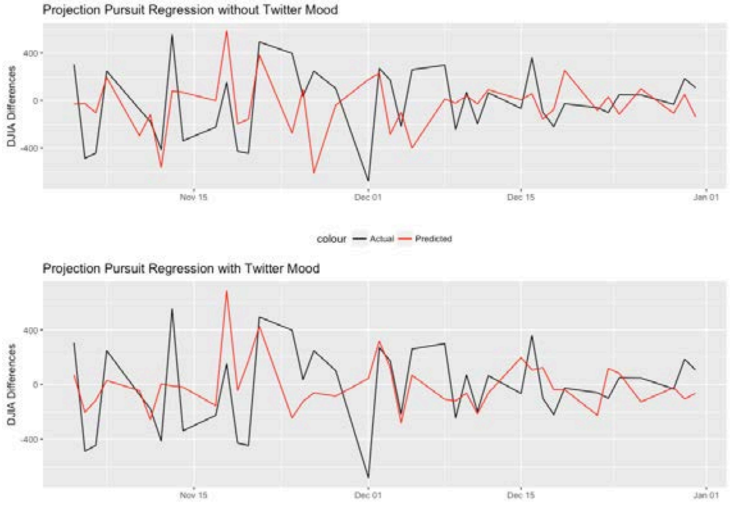


Figure 3.9. This plot shows $\Delta\widehat{DJIA}_t$ as predicted each day by the projection pursuit model described in (13), which does not incorporate Twitter mood, against the realized $\Delta DJIA_t$ in the top panel. The bottom panel shows $\Delta\widehat{DJIA}_t$ as predicted each day by the projection pursuit model described in (12), which does incorporate Twitter mood, against the realized $\Delta DJIA_t$. Both models were estimated using the full training set. Visually, there is no compelling evidence of outperformance by the model incorporating Twitter mood. We can see that both projection pursuit models make large out-of-sample errors.

Concluding remarks

> Given my results, the answer to the question of whether Twitter can predict the stock market is currently "no." … The methodology problems of Bollen et al., and the fact that several groups were unable to replicate their accuracy raise serious concerns about the validity of these authors' results. (Kuleshov 2011)

The results reported by BMZ are compatible with neither the information-theory justification nor the investor-sentiment justification for text mining. Even in the machine-learning literature of which TMP is a part, it is an outlier. We constructed a mood time series that shares five local extreme values with the time series shown in TMP. Despite looking similar to the time series shown in TMP, we found some in-sample but almost no out-of-sample evidence that this measure contains information about the DJIA.

Throughout this study, we have been limited primarily by a lack of access to deleted tweets. It is possible, but very unlikely, that the 10 percent of tweets deleted since BMZ took their sample positively covary with the linguistic sentiment factors that actually predict the stock market, but we have no means to investigate such a hypothesis. At the same time, a lack of tweets extending beyond 2009 prevented us from investigating the possibility that models fitted to linguistic sentiment variables will generally perform worse outside of recessions. Diego García (2013) suggests that most of the predictive power of linguistic sentiment comes from its ability to forecast asset price movements during recessions. We were unable to investigate whether his finding, conducted using measures on *New York Times* articles over the past century, carries over to Twitter.

We show that the out-of-sample failures by other researchers to find a Twitter-mood effect using off-the-shelf techniques also hold in the time frame covered by BMZ's original sample. By limiting our extended sample to that used by BMZ, we were able to obtain statistically significant relationships between one of our measures of Twitter mood and the stock market, but did not find any out-of-sample predictive power of this series. Given that the time series we select for in our main empirical tests was selected for its visual similarity to the BMZ time series, we believe that the deviations from our off-the-shelf techniques and BMZ's mood time series are small. Since BMZ obtain strong statistical relationships between the DJIA difference series and their CALM time series while, outside of our subsample matching theirs, we do not, we conclude that it is likely idiosyncratic features of their mood construction algorithm and the use of particular windows for data analysis that explain the low p-values obtained in their linear tests and the high predictive power in their non-linear tests. Perhaps it was data snooping that drove BMZ's in-sample results. Our opinion, based on our search of the literature and practitioner experience, is that there exists no credible evidence that one can use the collective mood content of raw Twitter text data from the universe of tweets to forecast index activity at the daily time scale.

Given that Bollen and Mao (2011, slide 19) claim that the GPOMS tool has gone through many versions, it is not inconceivable that, inadvertently, BMZ chose algorithms giving rise to co-occurrence weights that led to the overfit we conjecture exists in TMP. Indeed, given the large number of parameters that go into the mood time series construction algorithm, it is surprising that BMZ did not report more significant results. The issue of whether words should be unit-weighted or weighted according to another procedure, like in TMP, is currently controversial in text analysis for finance applications: Many researchers claim that these procedures allow them to achieve more intuitive and reasonable sentiment and mood scores while others claim that it gives researchers too many degrees of freedom to overfit (Loughran and McDonald 2016).

Economics and finance have long profited from the importation of methods and techniques from natural and computational sciences. Questions like whether one should term weight and problems associated with linguistic sentiment analysis in general cannot be resolved by strictly empirical work because even (reported) out-of-sample performance in the big-data regime could be the result of data snooping and publication bias. The taxonomy of the potential relationships between linguistic sentiments and asset price movements provided by Tetlock (2007) is, to us, a step in the right direction: It helped identify TMP as likely to be fundamentally flawed.

In 2006, the U.S. financial services sector accounted for 8.3 percent of GDP and has hovered around 8 percent into the 2010s. This represents an increase in size of nearly 200 percent from 1950 according to Robin Greenwood and David Scharfstein (2011), who point out that nearly a third of this increase comes from asset management fees. The representative asset manager does not outperform a low-cost cap-weighted index and the marginal asset manager is not likely to contribute to market efficiency (Malkiel 2013). While Derwent Capital Markets, no exception to the rule, only drew several million dollars of resources to itself, TMP has been cited by many small financial-service startups as evidence of the efficacy of specific kinds of large-scale mood analysis for stock market prediction, and of sentiment analysis in general.[39] If TMP's results are incorrect, then, to the extent that BMZ and their financial media enablers have drawn resources from

[39] See, e.g., the now-defunct website SNTMNT.com (**link**).

passive indices to actively managed funds searching for non-existent arbitrage opportunities, TMP has contributed to a growing deadweight loss in finance.

Appendices and code

Extensive appendices providing details on the data used and the mood analysis conducted for this article, as well as code files and the daily mood time series used to conduct the analysis, can be downloaded from the journal website **here**. The authors also make material related to the article available through GitHub (**link**). The authors' contract with Twitter prevents them from making the raw Twitter data publicly available. Researchers interested in accessing the aforementioned Twitter data for the purpose of reproducing results from this article or for use in other projects are encouraged to email Michael Lachanski at the address provided in his biography below.

References

Baker, Malcolm, and Jeffrey Wurgler. 2007. Investor Sentiment in the Stock Market. *Journal of Economic Perspectives* 21(2): 129–152.

Bollen, Johan, and Huina Mao. 2011. Twitter Mood Predicts the Stock Market. Presentation at Social Mood Conference (Atlanta), April 9. **Link**

Bollen, Johan, Huina Mao, and Alberto Pepe. 2011. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 450–453. Menlo Park, Calif.: AAAI Press. **Link**

Bollen, Johan, Huina Mao, and Xiaojun Zeng (BMZ). 2011 (TMP). Twitter Mood Predicts the Stock Market. *Journal of Computational Science* 2(1): 1–8.

Carmona, René. 2013. *Statistical Analysis of Financial Data in* R, 2nd ed. New York: Springer.

Chen, Ray, and Marius Lazer. 2011. Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement. Working paper. **Link**

Conway, Drew. 2010. The Data Science Venn Diagram. DrewConway.com, September 30. **Link**

Das, Sanjiv Ranjan. 2014. Text and Context: Language Analytics in Finance. *Foundations and Trends in Finance* (Now Publishers Inc., Boston) 8(3): 145–261.

DeLong, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. Noise Trader Risk in Financial Markets. *Journal of Political Economy* 98(4): 703–738.

Dew-Jones, Steve. 2013. DCM Capital Goes Under the Hammer. *WatersTechnology* (Infopro Digital Risk Ltd., London), February 5. **Link**

Edmans, Alex, Diego García, and Øyvind Norli. 2007. Sports Sentiment and Stock

Returns. *Journal of Finance* 62(4): 1967–1998.

Fama, Eugene F. 1970. Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance* 25(2): 383–417.

Farmer, Roger E. A., and Jang-Ting Guo. 1994. Real Business Cycles and the Animal Spirits Hypothesis. *Journal of Economic Theory* 63(1): 42–72.

García, Diego. 2013. Sentiment During Recessions. *Journal of Finance* 68(3): 1267–1300.

Gilbert, Eric, and Karrie Karahalios. 2010. Widespread Worry and the Stock Market. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 58–64. Menlo Park, Calif.: AAAI Press. **Link**

Greenwood, Robin, and David Scharfstein. 2011. The Growth of Finance. *Journal of Economic Perspectives* 27(2): 3–28. **Link**

Han, Zhichao. 2012. *Data and Text Mining of Financial Markets Using News and Social Media*. M.S. diss., University of Manchester. **Link**

Henry, Elaine. 2008. Are Investors Influenced by How Earnings Press Releases Are Written? *Journal of Business Communication* 45(4): 363–407.

Jordan, Jack. 2010. Hedge Fund Will Track Twitter to Predict Stock Moves. *Bloomberg News*, December 22. **Link**

Karabulut, Yigitcan. 2013. Can Facebook Predict Stock Market Activity? Working paper. **Link**

Kearney, Colm, and Sha Liu. 2014. Textual Sentiment in Finance: A Survey of Methods and Models. *International Review of Financial Analysis* 33: 171–185.

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Kuleshov, Volodymyr. 2011. Can Twitter Predict the Stock Market? Working paper. **Link**

Knyazkov, Konstantin V., Sergey V. Kovalchuk, Timofey N. Tchurov, Sergey V. Maryin, and Alexander V. Boukhanovsky. 2012. CLAVIRE: E-Science Infrastructure for Data-Driven Computing. *Journal of Computational Science* 3(6): 504–510.

Lachanski, Michael. 2014. Did Twitter "Calm"-ness Really Predict the DJIA? *Journal of Undergraduate Research in Finance* 4(1). **Link**

Lachanski, Michael. 2015. Not Another Market Timing Scheme!: Detecting Type I Errors with "Good Deal" Bounds. *Journal of Undergraduate Research in Finance* 5(1). **Link**

Lachanski, Michael. 2016. Did Twitter Mood Really Predict the DJIA?: Misadventures in Big Data for Finance. *Penn Journal of Economics* (PennJOE.com) 1(2): 8–48. **Link**

Lo, Andrew W., and A. Craig MacKinlay. 1988. Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test. *Review of Financial Studies* 1(1): 41–66.

Logunov, Anatoly. 2011. A Tweet in Time: Can Twitter Sentiment Analysis Improve Economic Indicator Estimation and Predict Market Returns? Bachelor's thesis, University of New South Wales. **Link**

Logunov, Anatoly, and Valentyn Panchenko. 2011. Characteristics and Predictability of Twitter Sentiment Series. In *MODSIM 2011: 19th International Congress on Modelling and Simulation*, eds. Felix Chan, Dora Marinova, and R. S. Anderssen, 1617–1623. Canberra: Modelling and Simulation Society of Australia and New Zealand. **Link**

Lorr, Maurice, and Douglas McNair. 1984. *Manual for the Profile of Mood States Bipolar Form*. San Diego: Educational and Industrial Testing Service.

Loughran, Tim, and Bill McDonald. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research* 54(4): 1187–1230.

Mackintosh, James. 2012. Last Tweet for Derwent's Absolute Return. *Financial Times*, May 24. **Link**

Malkiel, Burton G. 2013. Asset Management Fees and the Growth of Finance. *Journal of Economic Perspectives* 27(2): 97–108. **Link**

Manning, Christopher D., and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Mao, Huina, Scott Counts, and Johan Bollen. 2015. Quantifying the Effects of Online Bullishness on International Financial Markets. *Statistics Paper Series* 9. European Central Bank (Frankfurt). **Link**

Milnes, Paul. 2014. Whatever Happened to the "Twitter Fund"? HedgeThink.com, April 24. **Link**

Mitra, Gautam, and Leela Mitra, eds. 2011. *The Handbook of News Analytics in Finance*. Chichester, UK: Wiley.

Nassirtoussi, Arman Khadjeh, Saeed Reza Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Text Mining for Market Prediction: A Systematic Review. *Expert Systems with Applications* 41(16): 7653–7670.

Niederhoffer, Victor, and M. F. M. Osborne. 1966. Market Making and Reversal on the Stock Exchange. *Journal of the American Statistical Association* 61(316): 897–916.

O'Connor, Brendan T. 2014. *Statistical Text Analysis for Social Science*. Ph.D. diss., Carnegie Mellon University. **Link**

Pav, Steven. 2012a. The Junk Science Behind the "Twitter Hedge Fund." Sellthenews.tumblr.com, April 14. **Link**

Pav, Steven. 2012b. Derwent Closes Shop. Sellthenews.tumblr.com, June 22. **Link**

Pav, Steven. 2013. Bollen: 1; Worm: 0. Sellthenews.tumblr.com, October 21. **Link**

Porter, Alan L., and Ismael Rafols. 2009. Is Science Becoming More Interdisciplinary? Measuring and Mapping Six Research Fields Over Time. *Scientometrics* 81(3): 719–744.

Shen, Pu. 2003. Market Timing Strategies That Worked. *Journal of Portfolio Management* 29(2): 57–68.

Sprenger, Timm O., Philipp G. Sandner, Andranik Tumasjan, and Isabell M. Welpe. 2014a. News or Noise? Using Twitter to Identify and Understand Company-Specific News Flow. *Journal of Business Finance & Accounting* 41(7–8): 791–830.

Sprenger, Timm O., Andranik Tumasjan, Philipp G. Sandner, and Isabell M. Welpe. 2014b. Tweets and Trades: The Information Content of Stock Microblogs. *European Financial Management* 20(5): 926–957.

Tetlock, Paul C. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance* 62(3): 1139–1168.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More Than Words: Quantifying Language to Measure Firms' Fundamentals. *Journal of Finance* 63(3): 1437–1467.

Watson, Mark W., and James H. Stock. 2010. *Introduction to Econometrics*, 3rd ed. Boston:

Addison-Wesley.

Wong, Felix Ming Fai, Soumya Sen, and Mung Chiang. 2012. Why Watching Movie Tweets Won't Tell the Whole Story? In *Proceedings of the 2012 ACM Workshop on Online Social Networks*, 61–66. New York: Association for Computing Machinery. **Link**

Yates, Jonathan. 2011. Tweet Investing. TradersLog.com (Barcelona), May 13. **Link**

Zhang, Xue, Hauke Fuehres, and Peter A. Gloor. 2011. Predicting Stock Market Indicators Through Twitter: "I Hope It Is Not as Bad as I Fear." Presented at the 2nd Collaborative Innovation Networks Conference, Savannah, Ga. *Procedia – Social and Behavioral Sciences* (Elsevier Ltd., Amsterdam) 26: 55–62. **Link**

About the Authors

Michael Lachanski is an MPA and Office of Population Research certificate candidate at Princeton University's Woodrow Wilson School. He is currently a SINSI Fellow at the Department of Transportation, specializing in financial and regulatory policy analysis. He graduated in 2015 from Princeton University with an A.B. in Economics summa cum laude with certificates in engineering and management systems, applied and computational mathematics, statistics and machine learning, and finance. His email address is michael.stephen.lachanski@uwcim.net.

Steven Pav holds a Ph.D. in Mathematics from Carnegie Mellon University. He was formerly the Warcharski Visiting Assistant Professor at University of California, San Diego. His research is on quantifying trading strategies and the statistics of portfolio optimization. He lives and works in San Francisco, and blogs at **gilgamath.com**. His email address is steven@gilgamath.com.

Go to archive of Comments section
Go to September 2017 issue

Discuss this article at Journaltalk:
http://journaltalk.net/articles/5947