Volume 4 • Number 1
Fall 2014

JURF
Journal of Undergraduate
Research in Finance

Did Twitter "Calm"-ness Really Predict the DJIA?

Michael Lachanski[1]

Princeton University

Mark J. Bertus Prize Winner

In Bollen et al.'s 2011 article, "Twitter mood predicts the stock market", a proprietary measure of Twitter "calm"-ness was found to Granger cause increases in the DJIA from February 28, 2008 to November 3, 2008. We summarize several heretofore unpublished critiques of the statistical methods used in that article. We construct several alternative time series of Twitter "calm"-ness over November and December 2008; two months of data that Bollen et al. had collected but did not use for their Granger causality tests. Visual inspection suggests our measures of Twitter "calm"-ness replicate stylized features of Bollen et al.'s "calm"-ness measures. We find that our measures of Twitter "calm"-ness do not Granger cause either the level or change in the DJIA over this time frame.

JEL Classification Codes: G02, G12, G14

Keywords: stock market prediction, Twitter, mood analysis, text analytics, efficient market hypothesis

[1] Professors Rodriguez, Fan, Carmona, Horntrop and Liu taught me everything I know about hypothesis testing, financial econometrics, statistical analysis of financial data in R, probability theory and "big data"-style statistics, respectively. Professor Malkiel's early encouragement, feedback and funding recommendations made the paper possible. This project was financially supported by the Fred Fox '39 Fund and the great Class of '78 as well as Dean Poor and the faculty of the School of Engineering and Applied Science with the Keller Center's Eugene Wong '55 Fund for Engineering and Policy. The paper has benefit considerably from comments by Professor Fabozzi, Professor Bhatt, Professor Lu, Paul F. Schepel of Toussaint Capital Partners, Mark TenPas of Citadel Investments, Steven Pav, Professor Godbey, Gates Cambridge Fellow David Abugaber '14, Princeton Linguistics Club, Eric Taylor of OSEC, Ararat Gocmen, and of course, the anonymous referees of the *Journal of Undergraduate Research in Finance*. All errors are my own. Without a homework extension from Professor Zeng and travel funding from the Princeton University Department of Economics (particularly Professor Gul and Noelina Hall), I would not have been able to attend the Financial Management Association's 2014 Annual Meeting to present my work. Ani Deshmukh of Goldman Sachs initially referred me to JURF and so he has my infinite gratitude. I would like to thank Dean Avens for his timely and helpful advice. Last, but not least I would like to thank my parents without whom I would not be who I am today.

1. Introduction

Fan, Liu and Han (2014) point out that the so-called "big data" regime engendered by the increasing availability of "massive and high dimensional data" offers researchers a double-edged sword: new opportunities to discover "subtle population patterns and heterogeneities" on the one hand are balanced by the increased possibilities for spurious correlations and impossible-to-verify exogeneity assumptions on the other. In other words, even strong correlations in the big data regime, without cross-validation, could be suspect as spurious or the result of subtle errors in statistical technique. In their words, "high dimensionality introduces spurious correlations between response and unrelated covariates, which may lead to wrong statistical inference and false scientific conclusions."

While Fan et al.'s examples are taken from the world of genomics, they specifically highlight "unstructured text corpus" work as a domain in which the recent abundance of data has increased the danger of spurious correlation. No small part of this abundance has been the rise of Twitter, a microblogging platform that allows users to send "Tweets" (messages of 140 characters or less) to each other. Like most social media companies, Twitter generates revenue by selling users' data, but it also allows researchers and developers free access to an unprecedented volume of data generated on their site. Generating several terabytes of total data per day by the most recent estimates, Twitter's rise has been a boon for data-intensive research in computational and social sciences alike. By systematically sampling Tweets with some level of randomness, researchers can easily build bodies of natural language data (corpuses) in the billions of words. With high-dimensional text data like that obtained from Twitter, researchers can evade some of the challenges Fan et al. discuss by first applying dimensionality-reduction algorithms. Depending

on the domain, such techniques can have the added benefit of making the covariates under analysis much easier to interpret.

1.1 Killing Two Birds with One Stone: Moods Analysis

In computational psychology, "mood analysis" algorithms are one method of both reducing the dimensionality of and increasing the interpretability of the covariates under analysis. The most basic "bag-of-words"[2] approach to mood analysis supposes that humans have a fixed number of mood states and that every word we use can be associated with one of those mood states or a neutral category. As a purposely simplistic first example, we might suppose that humans have only two moods: happy and sad. In our example, the word "good" would be assigned to the mood state happy, the word "bad" would be assigned to the mood state sad and the word "I", which does not seem to be associated with either mood state a priori, would be assigned to neither. We could continue this process until we have obtained a dictionary structure that partitions[3] the English language into three categories "happy", "sad" or "neither". Our algorithm then takes in an expression, document or corpus and counts the number of words that our dictionary would assign to the category "happy" and the number of words our dictionary would assign to the category "sad." Then, a predefined kernel $f: \mathbb{R}^2 \to \mathbb{R}$ assigns the aforementioned counts of happy and sad words to a single number. A typical specification for our kernel might be:

[2] A bag-of-words is a document rendered as a set of ordered pairs $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where X_i is a unique word in the document and Y_i is the frequency with which that word occurs. For example, "Jimmy has a lovely cat named Jimmy" can be represented as the following bag-of-words {Jimmy: 2, has: 1, a: 1, lovely: 1, cat: 1, named: 1}. A bag-of-words technique is a technique, which operates on this data structure and does not need the original sentence to generate output. So a bag-of-words approach to sentiment analysis might use the fact that lovely is usually an affectionate description of something to suggest that Jimmy has positive sentiment about his cat. A bag-of-words approach to mood analysis might only notice the word lovely and suggest that Jimmy feels good.

[3] More sophisticated approaches will not only assign each word to a single mood state, but assign higher weights to words thought to more intensely express that mood state. For instance, "joyful" might be assigned a weight of 2 so that in our example, one instance of the word "joyful" contributes as much to our count of "happy" words as two instances of the word "good."

$$\frac{\#\{\text{words in expression associated with "happy"}\}}{\#\{\text{words in expression associated with "sad"}\}+1}=M \tag{1}$$

After applying this function, our mood analysis is complete. We have reduced our bag-of-words text data from dimension #{unique words in text} to a single number that could, for an appropriate dictionary and specification of f, be reasonably interpreted as the happiness or sadness of the author. For specification (1), the higher the value of M, the happier we might expect the author of the analyzed text to be.[4] The partition of all words into two categories (or three if one counts the third, neutral category), happy and sad, is called a naïve partition.

For a concrete application, consider the statement:

"I feel good."[5] (2)

Expression (2) has bag-of-words dimension three corresponding with the number of unique words in the expression. Our aforementioned mood analysis might assign "good" to the happy mood state but assign "I" and "feel" to neither happy nor sad mood state. If (1) is the specification chosen, then our mood analysis algorithm will assign expression (2) to 1. Now, consider the statement:

"I feel bad today." (3)

[4] If we wished to be pedantic about ensuring that M could take on the range of values in $\mathbb{R}$ we could simply take the log of specification (1).

[5] Punctuation is usually ignored (and we remove all punctuation from our text data except where explicitly specified).

Intuitively, we expect that the author of expression (3) is likelier to be in a sad mood state than the author of expression (2). Our mood analysis validates our intuition: assigning "bad" to the sad category and assigning "I", "feel" and "today" to neither the happy nor the sad category. For expression (3) and specification (1), $M = 0$, providing evidence for our claim that specification (1) assigns text expressing happiness on the part of the author to higher numbers and text expressing sadness to lower numbers.

Notice that for (1), $f(x, y)$ is decreasing in one of its arguments and increasing in the other. We call any mood analysis in which the kernel f has this property *bipolar.*[6] Lorr and Shea (1979) suggest that many human mood states are, in fact, bipolar in the sense[7] that if someone feels more of one mood state then they necessarily feel less of the other. In general, mood analyses can involve classification of moods into more complex categories than the bipolar, but unidimensional happy versus sad. For example, Lorr and Shea (1979) provide experimental evidence that humans have at least three bipolar mood dimensions[8] which Lorr and Shea name composed-anxious, energetic-tired and agreeable-grouchy. On the other hand, Lorr, McNair and

[6] An example of a non-bipolar mood analysis might use $f_t(x, y) = \frac{x+y}{t}$ where x and y are defined as the total number of terms with happy and sad emotional content respectively and t is the total number of words in the document. We might interpret such an expression as a measure of the emotional content of the text where higher values of f_t indicate higher levels of emotional content in the text.

[7] Notice that the use of the term bipolar by Lorr and Shea (1979) refers to an empirical fact about human mood states, while our definition of bipolar refers to the functional specification applied to the counts of the total number of words within a text that fall into predefined categories. One might argue that if mood states are actually bipolar with mood state X at one end of the spectrum and mood state Y at the other then in a mood analysis it is sufficient to work with the count of the words that we classify as belonging to the category X since higher counts on this "half-scale" measurement will indicate high expectations for the author's identification with the mood X and low expectations for the author's identification with mood Y and vice-versa for low counts. This reasoning is correct for text analysis applications, but Lorr, McNair and Fisher (1982) suggest that this reasoning is misplaced in the psychometric literature because of so-called "extreme response bias." Extreme response bias refers to the fact that certain individuals simply mark the maximum or minimum value on a survey even though their true value might be approximately neutral. To see how a monopolar survey design can cause problems if moods are actually bipolar, consider the case in which an individual with extreme response bias fills out a survey with only three half-scales (using Lorr and Shea's 1979 conception of human mood states): composed, agreeable, and energetic. For an individual with a neutral mood state and the largest extreme response bias (i.e. they mark off the adjectives with the most extreme responses possible) for all moods, moving from a monopolar measure to a bipolar measure will increase the accuracy of the test from 0% (maximum deviation from the true mood state) to 100%. For individuals without extreme response bias, the bipolar structure of the questionnaire should give results equal to the monopolar structure, ceteris paribus. Text-based mood analysis algorithms, like those proposed by Bollen, Pepe and Mao (2011), borrow liberally from the psychometric literature and we attempt to follow the established precedents used by psychometric literature even when it appears that the original justification (extreme response bias on surveys) for those precedents (bipolar specifications of f) do not necessarily apply.

[8] These three dimensions are, in theory, independent of each other, but experimental evidence suggests that the measurements on dimensions typically give small but statistically significant correlations.

Fisher (1982) suggest that humans have five bipolar mood dimensions that they term composed versus anxious, agreeable versus hostile, energetic versus fatigued, elated versus depressed and clear-thinking versus confused.

1.2 Sentiment Analysis

Despite the centrality of collective mood shifts to many behavioral theories of asset pricing and business cycles (Keynes 1936; Shiller and Akerlof 2010), C. Monica Capra (2004) notes that "mood is almost entirely unexplored in economics." Although mood analysis algorithms have found little application in economics until recently, Kearney and Liu (2014) suggest that there is now a rapidly growing empirical asset pricing literature that uses elements of mood analysis in what is called "sentiment analysis" or "opinion mining." The survey by Liu (2012) defines a sentiment implicitly as follows:

> An opinion consists of two key components: a target *g* and a sentiment *s* on the target, i.e., (g, s), where g can be any entity or aspect of the entity about which an opinion has been expressed, and s is a positive, negative, or neutral sentiment, or a numeric rating score expressing the strength/intensity of the sentiment (e.g., 1 to stars). Positive, negative and neutral are called sentiment (or opinion) orientations (or polarities).

We can see that a sentiment must be paired with a target object; much of the work in this area has simply applied mood analysis algorithms to the text surrounding a target of interest to estimate the sentiment that the author of the text has about that object. Since sentiments can be thought of as a summary statistic for an individual's stated preferences, it is not difficult to see

how the concept might find use in economics. Kearney and Liu's (2014) survey summarizes how sentiment analysis has been applied to several media sources for empirical asset pricing.

An example of a sentiment analysis using our previous set-up is:

"I feel bad today about Obama." (4)

"I feel good about Obama." (5)

In statement (4), the author expresses a negative sentiment about President Obama and in statement (5) the author expresses a positive sentiment about President Obama. To see how we can simply re-apply our work on mood analyses to sentiment analysis, let all the words we assigned to the categories "happy", "sad" and "neither" be, respectively, assigned to the sentiments "positive", "negative" and "neutral." Noting that "about" is neutral allows us to simply re-use all of our mood analysis calculations here. In this case, using specification (1) and letting *g* be "Obama", we can see that our sentiment analysis algorithm gives us opinions of (0, Obama) for statement (4) and (1, Obama) for statement (5), validating our expectation that the author of statement (5) has a higher opinion of President Obama than the author of statement (4).

1.3 Motivation

Bollen, Mao, and Zeng's (BMZ hereafter)[9] 2011 *Journal of Computational Science* article, "Twitter Mood Predicts the Stock Market", provides one of the first scale uses of mood analysis for empirical asset pricing. BMZ claim 86.7% accuracy in predicting the up-down sequence of the Dow Jones Industrial Average (DJIA hereafter) using a mood analysis algorithm applied to

[9] In response to numerous requests, Appendix E contains an abbreviation reference for all abbreviations used in this article.

Twitter text.[10] This finding suggested that Twitter mood could be used to devise lucrative trading strategies. To exploit this finding, BMZ teamed up with the hedge fund Derwent Capital Markets (DCM hereafter) and raised tens of millions of dollars in capital. Unfortunately, what seemed like a sure-fire moneymaker proceeded to go catastrophically awry as less than one year after being founded, DCM shut down and auctioned off its assets for approximately 120,000 GBP[11], far below the break-even point of 350,000 GBP and the guidance price of 5 million GBP.[12]

In this paper (DTMP hereafter), we make three original contributions. First we provide a concise, but complete documentation of the ambiguities in the methodology of "Twitter Mood Predicts the Stock Market" (TMP hereafter) that have prevented other researchers from replicating it. Second, we aggregate and summarize heretofore unpublished critiques of the statistical techniques used in TMP. Finally, we conduct an empirical investigation and try to rediscover a Twitter "calm"-ness effect in a sample of Twitter and DJIA data subsequent to that used in TMP. This entails constructing our own mood time series, providing enough detail to ensure reproducibility and performing the same hypothesis tests as in TMP.

Our ambiguity documentation reflects the fact that several key procedures used in TMP are only loosely described. The website cited in TMP that is supposed to contain all of the code and data used in the study, https://terramood.soic.indiana.edu/data currently yields a 404 error. Furthermore, the archives provided by The Internet Archive, a 501(c)(3) non-profit that takes snapshots of the web suggests that no material was ever available at the URL provided by

[10] The DJIA is a price-weighted stock market index of 30 large U.S. stocks.

[11] http://sellthenews.tumblr.com/post/45150391415/did-you-ever-think-your-tweets-might-predict-the, accessed on 2013/12/01.

[12] http://sellthenews.tumblr.com/post/42530922672/the-sentiment-trading-platform-is-for-sale, accessed on 2013/12/01.

BMZ.[13] By investigating the effect found for the months that BMZ left out of their Granger causality analysis[14], we aim to illuminate the strengths and weaknesses of the mood analysis and empirical asset pricing techniques found in TMP.[15]

The critiques we summarize, drawn from both unpublished notes and the empirical asset pricing blogosphere, motivate a robustness check on TMP that entails estimating Twitter "calm"-ness using the universe of Tweets from the months that BMZ leave out of their Granger causality tests: November and December 2008. This robustness check is valuable for three reasons. First, TMP is the most frequently cited article[16] in Elsevier's *Journal of Computational Science*. Had DCM succeeded, the Twitter mood effect would have joined the set of robust, actively traded financial market anomalies. Frequently cited results demand replication for robustness checks, and TMP particularly so because the poor out-of-sample performance of DCM raises questions about the existence of the Twitter mood-equity price relationship and consequently the profitability of any trading strategy it motivates. While precise replication is impossible because of ambiguities in the procedure, we can use standard mood analysis techniques outlined above to try to detect "calm"-ness effects.

Finally, our summary of critiques and empirical findings make a contribution to the literature on stock market predictability. If BMZ's result is correct, then they have provided strong evidence

[13] https://web.archive.org/web/20101030141043/http://terramood.soic.indiana.edu/data, accessed on 2014/11/06.

[14] BMZ write: "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 lagged values of X will exhibit a statistically significant correlation with Y. Correlation however does not prove causation. We...are not testing actual causation but whether one time series has predictive information about the other or not." We use the same approach and give the same interpretation to Granger causality analysis.

[15] All code and time series are available on request. While the DJIA data is publicly available, at this time I am unfortunately legally unable to share raw Twitter data without explicit permission from Twitter.

[16] As of the date of submission, TMP has over 900 citations on Google scholar in less than five years after publication.

against the random walk hypothesis (RWH hereafter). Because the RWH is considered a reasonable approximation of short-term asset price movements (Malkiel 2012), the results in TMP are highly improbable a priori.[17] TMP's claim that Twitter mood can predict the up-down sequence of the DJIA with 86.7% accuracy would, if correct, necessitate a rethinking of the RWH as even a short-run approximation to asset price behavior because the RWH suggests that asset price movements cannot be predicted with more than 50% accuracy. If we are able to find a Twitter "calm"-ness effect similar to that found by BMZ, this strengthens the evidence against RWH. Additionally, because we choose a sample of Twitter and DJIA data prior to the original upload of TMP to arXiv, finding a Twitter "calm"-ness effect will provide evidence that arbitrageurs eliminated the arbitrage opportunities described in TMP at some point after its initial publication. On the other hand, if we are unable to find a Twitter "calm"-ness effect in November and December 2008, then researchers will have another plausible explanation for DCM's failure: the mood effect demonstrated in TMP was a spurious correlation rather than a robust arbitrage opportunity.[18]

2. Literature Review

In this review, we first elucidate the methodology and findings of TMP, taking care to document the methodology ambiguities that prevent complete replication of TMP. Next, we compare those

[17] Lucas (1978) and Niederhoffer and Osborne (1966) provide theoretical and empirical evidence, respectively, that equity markets need not follow random walks over any time frame.

[18] It is also possible that the effect discovered in TMP existed but was non-robust in the sense that its appearance was conditional on particular economic conditions. In particular, Eric Taylor of OSEC has noted much of the American media's collective attention was on the stock market during the 2008 recession. It is not unreasonable to believe that the collective mood (as measured on social media) might co-move with or even Granger cause stock market changes under these conditions. As the financial crisis abated and collective attention moved away from the stock market, we might expect any statistical regularities in the joint distribution of collective mood and financial market behavior identified during the crisis to break down. Unfortunately, we do not have the data to test distinguish between this hypothesis and spurious correlation so we use the descriptors spurious and non-robust interchangeably.

findings with other effects from the behavioral finance literature, and finally we summarize subsequent critiques and partial replications.

2.1 TMP's Methodology

The authors characterize their work as an attempt to evaluate how psychological factors like mood impact equity prices. Before they can do this, they must devise a method for classifying and quantifying moods. BMZ write:

> To capture…public mood we created a second mood analysis tools [sic], labeled GPOMS, that can measure human mood states in terms of 6 different mood dimensions, namely *Calm, Alert, Sure, Vital, Kind and Happy*. GPOMS' mood dimensions and lexicon are derived from an existing and well-vetted psychometric instrument, namely the Profile of Mood States (POMS-bi).

The "POMS-bi" refers to a psychometric instrument called the Profile of Mood States-Bipolar Edition (typically referred to as the POMS-BI or POMS-Bi rather than POMS-bi). The psychologists Lorr and McNair (1984) who devised the test called it "bipolar" because it identifies moods as falling onto six spectra with both positive and negative axes according to Figure 1.

Figure 1.

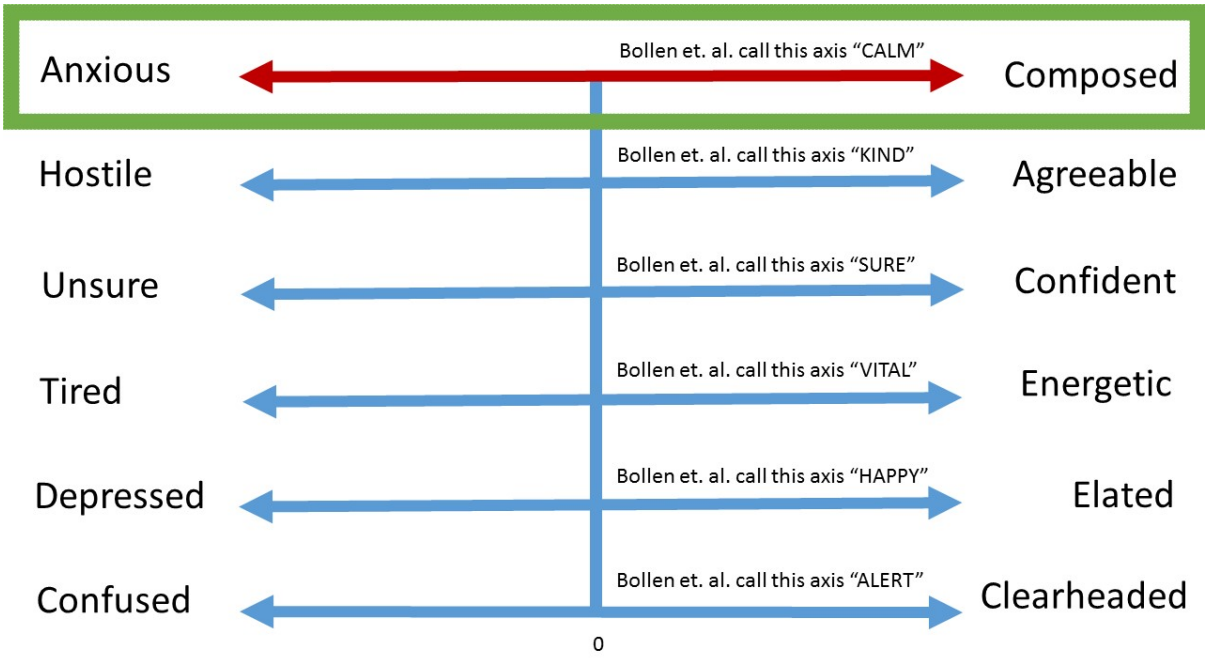

Figure 1: BMZ test for the presence of six moods effects. Only "calm"-ness (framed in green and highlighted in red above) was found to Granger cause DJIA increases.

The spectra used in the POMS-Bi are not *Calm, Alert, Sure, Vital, Kind and Happy,*[19] as a naïve reading of TMP would suggest, but composed-anxious, clearheaded-confused, confident-unsure, energetic-tired, agreeable-hostile, and elated-depressed respectively. In each spectrum, the first word in the pair represents the positive end of one mood spectrum and the second word in the pair represents the negative end of the same mood spectrum.

The POMS-Bi associates each positive and negative mood state with six words. For example, the POMS-Bi positive mood "composed" is associated with the words: "composed", "untroubled", "peaceful", "serene", "calm" and "relaxed". The POMS-Bi negative mood "anxious" is associated with: "tense", "nervous", "jittery", "shaky", "anxious" and "uneasy". We call these groupings of words lexicons. The composed-anxious spectrum (what BMZ call *Calm*) consists of the twelve-word union of the composed and anxious lexicons, each of which contains six words. The entire POMS-Bi lexicon is seventy-two words, twelve words for each of the six spectra with six words assigned to each end of the spectrum.

In a 2011 presentation sponsored by the Socionomics Institute for Advancing the Study of Social Mood and Social Action, Bollen and Mao explain how they mapped the POMS-Bi characterization of mood to the spectra used in the GPOMS. We reproduce the relevant slide in Figure 2. Since BMZ only find that their mood *Calm* has reasonable predictive power for changes in the DJIA, we only concern ourselves with tests of the effects of this particular mood, which we will generically refer to as Twitter mood, "calm"-ness or the composed-anxious spectrum.

[19] The characterization of moods presented in TMP should be called *unipolar* rather than bipolar because each positive mood is not paired with a negative mood. This is not to say that the GPOMS uses a non-bipolar kernel; because the GPOMS is a proprietary tool, we simply cannot know.

Figure 2.

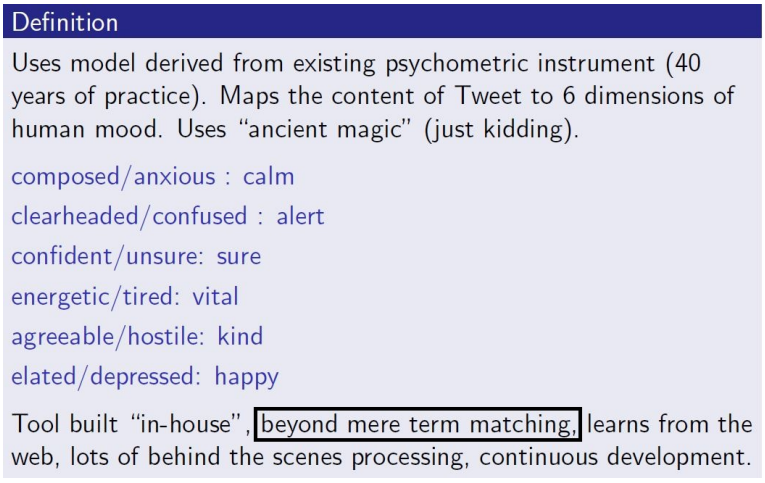

Figure 2: This slide relates the original POMS-Bi moods to the GPOMS tool built by BMZ. The box around "beyond term matching" above has been added to the original slide. This figure corresponds with slide 19 out of 46 in Bollen and Mao's 2011 presentation.

2.1.1 Methodology Ambiguity 1

To measure mood, BMZ expand the POMS-Bi lexicon to analyze Twitter data, doing so in order to "capture a wider variety of naturally occurring mood terms in Tweets and map them back to their respective POMS mood dimensions." BMZ claims that they:

> Created a second mood analysis tools [sic] labeled GPOMS, which can measure human mood states in terms of 6 different mood dimensions, namely *Calm, Alert, Sure, Vital, Kind* and *Happy*… To make it 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[20] computed by Google in 2006 from approximately 1 trillion word tokens observed in publicly accessible Webpages.

BMZ offer no further explanation as to how they constructed the GPOMS tool. This is problematic because there exists a large informatics literature that provides many methods to "analyze co-occurrences"[21] for the purposes of expanding a lexicon. Researchers hoping to understand, extend or replicate TMP must address this methodological ambiguity.

2.1.2 Methodology Ambiguity 2

The second methodological ambiguity relates to how BMZ use their GPOMS tool to derive mood time series from the Tweets. BMZ write:

[20] An n-gram is a set of words found in a certain order. For example, if the phrase "we are the robots" was commonly found on public websites in 2006, this phrase might have made its way into the Google N-gram corpus as a 4-gram.

[21] A co-occurrence is simply the presence of two words next to each other.

> We match the terms used in each tweet against this lexicon. Each tweet term that matches an *n*-gram term is mapped back to the original POMS terms (in accordance with its co-occurrence weight) and via the POMS scoring table to its respective POMS dimension.

TMP seems to count the number of matches between terms in the GPOMS and each Tweet. Since every term in GPOMS was derived from a term in one of the six POMS-Bi moods, BMZ suggest that one can multiply each GPOMS term-Tweet match by a "co-occurrence weight" to get a number on the POMS-Bi mood scale. Consider the Tweet: "I actually feel sick about going back to sixth form… So anxious and want to just skip my whole life." Our Google N-grams data might contain the n-grams "I actually feel sick" and "anxious skip my whole life". Since "anxious" is in the POMS-Bi lexicon, it is possible that the subsequent 4-gram "skip my whole life" was added to the GPOMS for the composed-anxious axis and given a co-occurrence weight of 0.5. The 4-gram expression "I actually feel sick" might also map to anxious half-scale[22] and be given a co-occurrence weight of 1.5. Suppose that anxious has a co-occurrence weight of 1; then the example Tweet would be given a score of negative 3.0 on the composed-anxious axis since the term "anxious" occurs once and has a co-occurrence weight of 1 with anxious (the negative part of the composed-anxious axis), the expression "skip my whole life" occurs once and has a co-occurrence weight of 0.5 and the expression "I actually feel sick" occurs once and has a co-occurrence of 1.5.[23]

[22] The presence of a 5-gram like "I actually feel sick anxious" or "anxious I actually feel sick" in our Google N-gram data would allow us to assign a co-occurrence weight to the 4-gram "I actually feel sick."

[23] In this case, our bipolar kernel is a simple sum of composed co-occurrence scores subtracted from the sum of anxious co-occurrence scores.

BMZ provide no explanation for how they created their GPOMS lexicon or calculated the co-occurrence weights. Additionally, notice that the algorithm proposed above relies on term matches between the Tweet text and the GPOMS lexicon. However, in the slide excerpted in Figure 2, Bollen and Mao suggest that the GPOMS tool calculates collective mood changes using a technique that goes "beyond mere term matching". Without additional information about how the GPOMS works, it is difficult to pursue this possible contradiction any further.

2.1.3 Methodology Ambiguity 3

Since TMP includes a number of dimensionless mood time series, BMZ normalize their moods to compare the effects of changes in two different moods, for instance *Calm* and *Happy*, on stock prices in terms of changes in standard deviation. They write the following:

> We normalize them[24] 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:
>
> $$\mathbb{Z}_{X_t} = \frac{X_t - \bar{x}(\mathrm{X_{t\pm k}})}{\sigma(X_{t\pm k})} \quad (6)$$
>
> where $\bar{x}(\mathrm{X}_{t\pm k})$ 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.

BMZ do not indicate what *k* is or where it comes from. The fact that *k* is not fixed suggests that it

[24] Here, "them" refers to the mood time series calculated in TMP.

varies through time, but BMZ do not provide a procedure to allow a researcher interested in replicating this work to calculate *k*.

2.1.4 Statistical Evaluation of the Effect of Collective Mood Changes on the DJIA

To answer the question of whether or not Twitter mood predicts the stock market, BMZ follow the procedure presented in Stock and Watson (2007). They obtain DJIA data from Yahoo! Finance and fit autoregressive distributed lag (ARDL hereafter) models to changes in stock prices from February 28th, 2008 to November 3rd, 2008, removing weekends from their data. BMZ write:

> We are concerned with question [sic] 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…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*: $D_t = DJIA_t - DJIA_{t-1}$.[i] 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 28th to November 3, 2008 to exclude the exceptional public mood response to the Presidential election and Thanksgiving from the comparison.

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

$$L_1: D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-i} + \epsilon_t \tag{7}$$

$$L_2: D_t = \alpha + \sum_{i=1}^{n} \beta_i D_{t-i} + \sum_{i=1}^{n} \gamma_i X_{t-i} + \epsilon_t \tag{8}$$

In these models, $\alpha, \beta_{i,i\in\{1,\dots,n\}}$ and $\gamma_{i,i\in\{1,\dots,n\}}$ are parameters estimated from the data, D_t defined above by BMZ, and ϵ_t only requires the property that: $E_t(\epsilon_t | Y_{t-1}, Y_{t-2}, \dots, X_{t-1}, X_{t-2}, \dots) = 0$. It is not difficult to see that these 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. This nested structure allows us to conduct Granger causality tests. Stock and Watson (2007) define a Granger causality statistic as "the F-statistic testing the hypothesis that the coefficients on all the values of one of the variables" $(X_{t-1}, X_{t-2}, \dots, X_{t-q})$ are zero. 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 $n \in \{1, \dots, 7\}$. For each n, the null and alternative hypotheses for the Granger causality tests are given by:

$$H_0\colon\ \gamma_1 = \gamma_2 = \cdots = \gamma_n = 0 \tag{9}$$

$$H_A\colon\ \gamma_1 \neq 0 \vee \gamma_2 \neq 0 \vee \cdots \vee \gamma_n \neq 0 \tag{10}$$

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 results with F-statistics associated with p-values less than 10% significant.

2.2 TMP's Contribution to the Behavioral Finance Literature

TMP purports to measure stock prices' sensitivity to mood changes with a finer granularity than previously used. BMZ write:

> 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.

The first technique for measuring mood to determine its impact on stock prices involves measuring factors that are correlated with mood; if any event is known to correlate with a change in collective mood, then one can use these events in a standard event study framework to estimate the impact of collective mood changes on equity prices. For example, using the well-known relationship between sports victories and collective mood changes, Edmans et al. (2007) document a next-day increase in stock prices from international soccer victories, and a decline after losses. This effect does not persist for more than one day. Additionally, Pettengill (2003) points out a "Monday effect" on stock prices: on average, stock prices decrease on Monday, and more significantly if they fell on the previous Friday. A common explanation for this well-documented day-of-the-week effect is that traders are depressed at the beginning of their work week. Finally, weather can be used as a proxy for mood. Hirshleifer and Shumway (2003) discover that morning sunshine increases the return for that day and the next day. All of these effects on stock prices respond immediately to mood changes, and vanish rapidly, typically by the next day.[25]

The second technique for measuring mood involves evaluating consumer confidence surveys. Today, consumer confidence reports are sold to algorithmic traders seconds before they are released so that they can trade on the mood information before it reaches the market. However, the association between these surveys and stock price performance is unclear at daily time scales. Both Jansen and Nahuis (2002) and Statman and Fisher (2003) find that high consumer confidence has an insignificant negative effect on equity returns; however, past high returns predict high consumer confidence over one to two weeks. Lemmon and Portniaguina (2006) find

[25] Kamstra, Kramer and Levi (2003) document that returns are lower in the winter and fall and provide geographic evidence that declining sunlight is the key driver of the decreased returns. This may provide an example in which mood shocks persist, but because the onset of the mood shocks are not precisely identified, we cannot be certain.

that only certain portions of the consumer confidence index have predictive power for a subset of small-cap stocks on a quarterly time scale. BMZ point out that surveys are not conducted at intervals frequent enough to be comparable to TMP.

Finally, a recent innovation to measure mood utilizes natural language processing (NLP) techniques to conduct mood analysis. Antweiler and Murray (2004) use NLP applied to Yahoo! Finance message boards, and discover that neither the content nor volume of messages carry significant information about the direction of equity returns, although the volume of messages predicts volatility. Tetlock (2007) matches a list of words from the Harvard General Inquirer against a *Wall Street Journal* column to derive a measure of investor pessimism. He finds a significant negative effect on the DJIA from pessimism the previous day. The effect becomes insignificant on day two and declines to zero after five days. Tetlock finds significant support that investor sentiment about the stock market interacts with the stock market in a complex and non-linear, but predictable fashion. These studies disagree on the magnitude of collective mood shifts' effect on stock prices, but agree that the first order effects of mood should occur immediately following mood changes, and vanish quickly.

2.3 BMZ Contra the Behavioral Finance Literature: Mood Effects Do Not Appear Instantaneously and Persist

Because BMZ do not provide the coefficients of their estimated ARDL models, we can only conduct a qualitative robustness check of their results. We compare the time-to-appear and persistence of Twitter mood on stock prices with those in the behavioral finance literature. BMZ's results for their Granger causality test for nested models L_1 and L_2 are shown in Figure 3.

Figure 3.

Statistical significance (p-values) of bivariate Granger-causality correlation between moods and DJIA in period February 28, 2008 to November 3, 2008.

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 3: This figure corresponds with Table 2 in TMP. We have added the blue and red boxes around the results of interest. OF is an abbreviation BMZ use for the "Opinion Finder" tool. The Opinion Finder is a standard computational linguistic tool for scoring phrases and documents according to their positive and negative sentiment. BMZ find significance for lags 2 through 6 of their ARDL model using a measure of Twitter "calm"-ness. However, standard results in the behavioral finance literature suggest that the effect of collective mood shifts on equity markets should appear instantaneously and vanish rapidly. In this respect, the results BMZ obtain using the Opinion Finder tool (boxed in blue above) are more in line with standard models of large-scale sentiment changes.

All of the major behavioral finance findings cited above suggest that collective mood changes impact the stock market instantaneously. Yet, the p-values in Figure 3 suggest that changes in Twitter "calm"-ness take two days to have a statistically significant impact on equity prices. This suggests that, if the findings are not spurious, that the Twitter "calm"-ness effect is not analogous to previous behavioral finance research on collective mood changes.

BMZ find that statistically significant effects of collective mood shifts persist for several days. While Tetlock (2007) finds similar persistence in collective mood impacts on stock prices, all research in the event study framework suggests that such mood effects should vanish quickly. In fact, the results BMZ obtain using their Opinion Finder tool (boxed in blue in Figure 3 above) to measure Twitter sentiment are much more analogous in both time-to-appear and persistence to those obtained in the event study literature.

2.4 Critiques from the Blogosphere and Two Unpublished Notes

Given the divergence between BMZ's findings and the behavioral finance canon, it is not surprising that the finding has generated controversy in the blogosphere. Hedge fund principal and blogger Steven Pav, in his comments on TMP, went so far as to write that BMZ:

> …exhibit a level of sloppiness that taints the integrity of the results. From basic accounting mistakes to appalling methodological flaws, these errors call into question whether *any* of their results can be trusted…Because the paper gives only loose technical details, and the data used are not widely available (collecting *all* Twitter feeds over a 1 year period is a technically challenging feat),

> it is impossible to definitively refute the claims; rather they can only be cast into serious doubt. [26]

However, this criticism appears to have completely evaded academia, as we were unable to find a single published criticism of TMP. In this section, we summarize the contents of comments, blogs and unpublished papers that provide substantive critiques of TMP.

We order these critiques by validity and importance. Our first critique appears fatal for the significance of the results shown in Figure 3. Our second critique and third critique, if addressed, may prove similarly deleterious. Finally, our fourth and fifth critiques, while valid in our opinion, only reflect the possibility of error or are otherwise unlikely to undermine the core conclusions of TMP by themselves.

2.4.1 Multiple Hypothesis Testing Biases

Stanford Computer Science doctoral candidate Volodymyr Kuleshov (2011) points out in an unpublished note that because BMZ test each of the six GPOMS moods and a seventh sentiment generating tool separately, they implicitly conduct a multiple hypothesis test, but fail to adjust their p-values upward using the Bonferroni correction. Blogger Steven Pav at Sell the News, Buy the Hype points out that the Bonferroni corrected p-values are not significant at any standard level.[27]

[26] http://sellthenews.tumblr.com/post/21067996377/noitdoesnot, accessed on 2014/11/14.

[27] http://sellthenews.tumblr.com/post/21067996377/noitdoesnot, accessed last on 2014/11/13.

2.4.2. Selection Bias and No Cross-Validation

Ben Gimpert points out that the reported p-values of the Granger causality analysis in TMP are biased upward because of the selection bias associated with the choice of the endpoint (November 3rd, 2008). Gimpert notes that BMZ:

> … exclude "exceptional" sub-periods from the sample, around the Thanksgiving holiday and the U.S. presidential election. This has no economic justification, since any predictive information from tweets should persist over these outlier periods.

In fact, we add that Gimpert understates the magnitude of the selection bias problem. In their 2011 presentation, Bollen and Mao confirm that they had obtained to the universe of Tweets from 2006 to 2008 but TMP only looks at the data from February 29, 2008 onwards. In other words, BMZ's choice for not just the endpoint, but also the start date of the Granger causality analysis appears to be arbitrary at best, and data mined at worst.

Petzoldt points out BMZ do not cross-validate any of their models ex-sample.[28] They use the same dataset (of 2008 DJIA prices from February 29 to November 3rd) for suggesting that "calm"-ness is the best predictor of the DJIA and for evaluating the effectiveness of "calm"-ness in their Granger causality tests. Figure 4 provides details. BMZ train a neural net using data from February 28 to November 30, 2008 which they test using data from December, 2008 (see SOFNN training in Figure 4), but do not cross-validate this model either. This curious omission motivates the Granger causality tests we provide in Section 4.

[28] http://petzoldt.tumblr.com/post/3236488086/statistical-flaws-in-twitter-mood-predicts-the-stock, accessed last on 2014/11/14.

Figure 4.

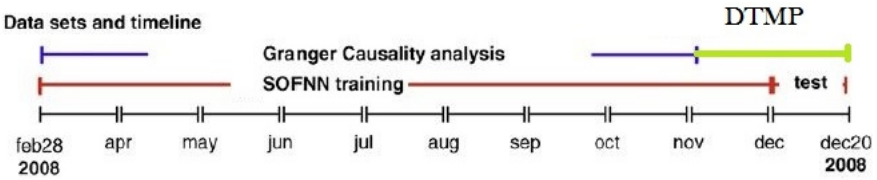

Figure 4: This figure corresponds with Figure 1 in TMP. Standard cross-validation of the models BMZ fit for their Granger causality analysis would have tested their models, fitted using data from February 28th to November 3rd, with data from November and December 2008. BMZ conduct no such ex-sample validation of their ARDL models in TMP. Our Granger causality tests use data from November and December 2008.

2.4.3 Normalization Procedure Takes in Data from the Future

Kuleshov (2011) points out that the mood normalization procedure in equation (6) takes in data from the future and biases the coefficients. BMZ appear to be aware that the "forward-looking" normalization procedure in equation (6) will introduce what they call "in-sample bias"[29], 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.

Because of the ambiguity in the choice of k (see section **2.1.3** for details), we are unable to estimate the size of the bias that normalization formula (6) induces. Given that BMZ appear aware of this, it is curious that they choose to use (6) anyway for their Granger causality analysis. There exist a number of available alternative formulas that would enable the interpretability of the raw mood time series without inducing in-sample bias. We present such a formula in (11) below:

$$\mathbb{Z}^{*}_{X_t} = \frac{X_t - \bar{x}(\mathrm{X_{t-k}})}{\sigma(X_{t-k})} \tag{11}$$

[29] The precise meaning of "in-sample" bias is unclear here. We take this 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 parsing 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 A. they test a number of moods and B. the neural net's training time-period overlaps with the Granger causality test time 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 A. and B. would be to simply validate both the neural net and the ARDL models ex-sample.

where $\bar{x}(\mathrm{X}_{t-k})$ and $\sigma(X_{t-k})$ 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*. Note that, unlike BMZ, we are only interested in evaluating the Twitter "calm"-ness effect and do not wish to compare multiple mood time series. Therefore, we do not actually use (11) in any of the calculations below, but simply present it as a standardization that would not have contaminated the coefficients in the ARDL model with information from the future.

2.4.4 Calculation Errors, Inappropriate Use of Significant Figures and Evidence of Possible Data Mismanagement

BMZ state that their self-organizing neural net successfully predicts 13/15 days of stock prices movements in December 2007. This is 86.7% accuracy, but blogger Steven Pav points that elsewhere in TMP, BMZ erroneously claim to have reached 87.6% accuracy using their neural net.[30] Steven Pav also points out that 86.7% accuracy figure is reported with too many significant digits, writing that:

> If the model had correctly predicted 12 or 14 days' directions, instead of the 13 it did, the number would change by plus or minus 7 percent. For the technically minded, the standard error on the accuracy figure is around 9%, and a 95% lower confidence interval on the accuracy figure is 72%. For the layman, the upshot is that it is not inconceivable that the accuracy of the system is as small as 72%, but it looked better in this experiment simply due to random luck. In all, reporting two

[30] http://sellthenews.tumblr.com/post/21067996377/noitdoesnot, accessed last on 2014/11/13.

> significant figures is unwarranted, much less three. The effect is perhaps minor, but it does not instill confidence in the authors' attention to detail.

Quant and finance blogger Ben Gimpert notes that there is evidence of possible data mismanagement on the part of BMZ because they report that their Granger causality tests were conducted on "a time series of 64 days" but the Yahoo! Finance DJIA time series from February 29th, 2008 to November 3rd, 2008, which BMZ claim to use, includes 173 trading days.

2.4.5 Results Are a Priori Implausible and Ex-Post Fail-to-Replicate

Steven Pav and Ben Gimpert point out that the results in TMP are unlikely to hold in general because all of the tests BMZ report were conducted during a bear market.[31] Steven Pav uses a Monte Carlo simulation to show that if the 86.7% accuracy in predicting the DJIA held for all time-periods, then the strategy of timing the market using the techniques in TMP would have a Sharpe Ratio greater than 8, making it the greatest market timing strategy ever discovered. Because the implied performance is so far out of the range of published market timing strategies (Shen 2002), the accuracy figure is a priori unlikely.[32]

We found several credible ex-post replications. First, Sharma and Vyas (2012), like BMZ, extend the POMS-Bi to generate a mood lexicon and then used Twitter data from January 2009 to March 2010 to generate mood indices. Sharma and Vyas conducted Granger causality tests on

[31] http://blog.someben.com/2011/05/sour-grapes-seven-reasons-why-that-twitter-prediction-model-is-cooked/, accessed last on 2014/11/14.

[32] http://sellthenews.tumblr.com/post/21067996377/noitdoesnot, accessed last on 2014/11/14.

ARDL models that indicated that the mood "Clearheaded-Confused" was significant at the 1% level, but Composed-Anxious was not significant at any standard level.[33]

In 2011, Stanford Computer Science Professor Andrew Ng gave his students the final project of replicating TMP or using a similar methodology to investigate the relationship between measures of Twitter mood and the stock market.[34] Since, as we have already shown, TMP is not replicable from publicly available data, seven projects tried variations on methodology of TMP. Of the seven projects, none was able to achieve a prediction accuracy for the stock market (the DJIA, Nasdaq and other sets of equities were considered) above 80%. All but one could not achieve a prediction accuracy above 70%. Stanford computer science Ph.D. candidate Vladimir Kuleshov using a different mood lexicon and different neural net algorithm, but otherwise following the procedure of TMP exactly, concluded his unpublished note on TMP by writing:

> 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. Given the boldness of their claims, I believe they ought to either publish their methods and their code, or withdraw these claims.

3. Data and Methodology

The criticisms above motivate our own investigation of TMP in which we attempt to follow the procedures of BMZ as closely as possible. Our chief innovations over previous replication attempts is that we use the universe, rather than a subsample, of data matching BMZ's

[33] Sharma and Vyas do not correct their p-values for multiple hypothesis testing. This means that, assuming that both the methodologies of BMZ and Sharma and Vyas yield valid measurements of the mood of individuals on Twitter, these contradictory findings are most consistent with spurious results on the part of one or both research groups.

[34] http://micarum.blogspot.com/2012/02/rethinking-of-sentiment-analysis.html, accessed last on 2014/11/14.

specification and we choose a time frame within TMP's sample. To attempt to overcome our uncertainty about exactly how GPOMS generates a "calm"-ness time series, we generate two mood lexicons and provide four mood kernels, giving us a total of eight "calm"-ness time series, each of which we subject to the same Granger causality analysis as BMZ.

3.1 Tweet Cleaning and Data Management

Following BMZ, we are concerned with Tweets that contain explicit words that represent the author's mood. We purchased the universe of Tweets matching the expressions "I feel", "I am feeling", "I don't feel", "I'm", "Im", "I am", and "makes me", for November and December from Twitter. By using these expressions to filter our Tweets, BMZ selects Tweets both more likely to be written in English and to have a higher emotional content than would be obtained from a purely random sample. BMZ do not perform inference on the mood time series they construct, but this sampling scheme, a kind of importance sampling, should ensure that the variance of their (and our) mood time series estimators is smaller than would be obtained from a purely random sample.

Our dataset consists of 3,510,351 Tweets divided into 8,757 files, each having a ten-minute duration. Since our investigation is for the U.S., we recode our Tweets, originally in Greenwich Mean Time (GMT), into Eastern Standard Time (EST). To focus our analysis, we isolate the text and user-id of each Tweet, and ignore the additional associated data, including location, and re-Tweets. This simplifies our data management, and these factors are not directly related to the mood expressed in a Tweet. Our Tweets are from 2008, when most Tweets written in English were from users in the U.S. However, from mid-2009 onward, a substantial and increasing

proportion of Tweets, even those written in English, were produced by users outside the U.S. Therefore, using Tweets from 2010 onward without considering geographic information would not accurately produce a collective mood index for the U.S.

3.2 Constructing the Mood Lexicons

To construct our mood index, we begin with the composed-anxious lexicon of POMS-Bi, consisting of six composed terms and six anxious terms, used by BMZ. We follow BMZ by augmenting the POMS-Bi lexicon using Google N-grams data. In January 2006, Google "took a picture" of the public web and extracted the text of publicly available English-language webpage. It provided this data in the form of the Google N-gram corpus[35] and temporarily made it publicly available.[36] The corpus consists of 1 trillion words organized into 1-grams, 2-grams, 3-grams, 4-grams and 5-grams. Following BMZ, we only use the 4-grams and 5-grams to augment our lexicon. We have a total of 1,313,818,354 4-grams and 1,176,470,663 5-grams for a total of 2490289017 4 and 5-grams. The Google N-gram corpus is a good choice for augmenting the POMS-Bi because it has been used by Google in thousands of other NLP projects, and we are chiefly interested in how words from the POMS-Bi lexicon are used in natural language.

Ultimately, because we are uncertain of how exactly the GPOMS uses co-occurrences between the POMS-Bi terms and phrases in the Google N-gram corpus (see section **2.1.1**), three Princeton University undergraduates created a composed-anxious lexicon by manually analyzing the Google N-gram corpus. Details are provided in section **3.2.1**. Because of the inherent subjectivity of this analysis, we validate the conclusions drawn from this composed-anxious

[35] NB: an N-gram is an ordered n-length list of words. "We are the robots", for instance, could be a 4-gram.
[36] The full explanation Google N-gram dataset, samples and a link to where the corpus can be purchased may be found here: http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

lexicon by building a second composed-anxious lexicon mechanically. This second lexicon expands the POMS-Bi lexicon by adding terms from Roget's 1911 Thesaurus that match the original POMS-Bi terms in connotation. Details are provided in section **3.2.2**.

3.2.1 Lexicon Extension by Inspection

To generate an extended POMS-Bi composed-anxious lexicon, we isolated every 4-gram and 5-gram database in the Google N-gram database containing one of the composed-anxious POMS-Bi terms using regular expressions. This yielded 227,037 4-grams and 205,945 5-grams that contain a "composed" term and 109,668 4-grams and 103,449 5-grams that contain an "anxious" term respectively. Then, three Princeton University undergraduate students inspected each of the 646,099 total 4 and 5 grams. Each undergraduate assigned adjectives, adjective phrases and past-participles that they identified as having a composed and anxious connotation to a new list of composed and anxious terms, respectively. The final composed lexicon, termed C_I^* , is an 82 term list obtained from the union of each of the three undergraduate's new composed term lists, as well the original six terms from POMS-Bi composed half-scale. The final anxious lexicon, termed A_I^*, is an 80 term list obtained from the union of each of the three undergraduate's new anxious term lists, as well the original six terms from POMS-Bi anxious half-scale. Like BMZ, we extend our POMS-Bi lexicon using only co-occurrences between Google N-grams database terms with terms already in the POMS-Bi composed-anxious lexicon.

Appendices A and B, respectively, provide the full composed term list, C_I^*, and anxious term list, A_I^*, that were obtained with the extension by inspection technique.

3.2.2 Thesaurus Augmentation Method

To validate the inherently subjective lexicon extension method provided in section **3.2.1,** we also provide an automatic extension process using Roget's 1911 Thesaurus.[37] The augmentation procedure was particularly simple. To generate C_T^*, the final composed term list generated with the extension by thesaurus augmentation method, any time a word in the original POMS-Bi half-scale for composed was found in a section of the thesaurus with heading "adjective" all phrases and words in that thesaurus section (but not those in the "related words" section denoted by hashtag) were added to C_T^*. The analogous procedure generated A_T^*.

Appendix C and D, respectively, provide the full 140 term composed list and 148 term anxious list we obtain via the extension by thesaurus augmentation method.

3.3 Mood Time Series Kernel Specifications

After mood lexicons have been generated, each Tweet is scored according to the number of regular expression matches with terms in our lists: C_I^* and A_I^* as well as C_T^* and A_T^*. To convert these into a final daily mood score, we must specify a kernel function f that maps the two counts for composed-ness and anxious-ness into a single real number. Besides equation (6), BMZ do not provide enough details about how the GPOMS generates a daily "calm"-ness score in TMP to replicate their analysis (see section **2.1.2** for more details). Therefore, in this section we present a variety of kernel functions and levels of aggregation for the application of our kernel functions. Section **3.3.1**'s kernel functions are applied to the collection of Tweets treated as a well of text. Section **3.3.2**'s kernel functions are applied to each Tweet, then averaged to make a daily

[37] NB: We used the 1911 edition primarily because it comes in an easily parsed XML format. The thesaurus is available here, https://code.google.com/p/roget1911/wiki/RogetXml, last accessed on 2014/11/15.

"calm"-ness score. In the section that follows, we generically refer to C^* and A^* but we test all kernel functions against both sets of mood lexicons (see section **4.1** for more details).

3.3.1 Mood Time Series Specifications: Aggregate Daily Twitter Activity

Using C^* and A^*, we estimate Twitter calm-ness as follows: read in all of the day t's Tweets, and count the number of words in the Tweets that match terms in C^*. Let the number of matches be denoted as C_t for date t. Similarly, we count the number of words in the Tweets that match terms in A^*, denoted A_t for date t. We calculate our composed-anxious score for date t to generate the mood time series:

$$CA_t^1 := C_t - A_t \tag{12}$$

Because of specification uncertainty, we present the following alternate aggregate specification of "calm"-ness.

$$CA_t^2 = \frac{C_t}{A_t + 1} \tag{13}$$

3.3.2 Mood Time Series Specifications: Individual Tweet Activity

During their 2011 Socionomics presentation, Bollen and Mao suggest that they use the GPOMS to score each Tweet individually, rather than using the aggregate daily Tweet text. This motivates the kernels we present in (14) and (15).

Using C^* and A^*, we estimate twitter mood as follows: read in the Tweet, and count the number of words in the Tweet that match terms in C^*, denoted by C_t for date t. Similarly, we count the number of words in the Tweet that match terms in A^*, denoted A_t for date t. We calculate our composed-anxious score for date t to generate the mood time series:

$$CA_t^3 := \frac{1}{n_t}\sum_{i=1}^{n_t}\left[C_t^i - A_t^i\right] \tag{14}$$

In the above formula, n_t is the number of Tweets that day and i indexes the Tweet for that day. This is specification is unsatisfactory because all it does is scale (12) by the number of Tweets per day, while this can potentially address the problem of unit roots from Twitter's growing user-base, it does little to mitigate heterogeneity in emotional content across Tweets and allows a potentially small number of users to have outsize effect on our mood daily mood index. Fortunately, applying the rescaling in (14) to the specification in (13) yields:

$$CA_t^4 := \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{C_t^i}{A_t^i + 1} \tag{15}$$

This form has many of the useful properties of (14) in that it can potentially eliminate unit roots. More importantly, some types of individual Tweet heterogeneity in total mood content is controlled because Tweets with high emotional content (i.e. a large number of matches for terms in both C^* and A^*) have their contribution to the "calm"-ness score is scaled down by their "anxious"-ness.

These four kernel functions, varying by functional form and level of aggregation for the Tweets and our two lexicons, provide us with a total of eight "calm"-ness time series. These time series, by construction, are strongly correlated and so adjusting for the multiple hypothesis tests we will

perform in the next section are non-trivial in the sense that the Bonferroni test is likely to generate false negatives. Details on the time series and hypothesis test results are provided throughout Section **4**. Fortunately, even the most liberal multiple hypothesis corrections give the same results as the Bonferroni correction in this case (see Section **4.2** for details).

3.4 Financial Time Series

Like BMZ, we obtain our DJIA adjusted closing price series using Yahoo! Finance from October 31st, 2008 to December 31st, 2008. We follow BMZ in deleting all weekends from both our financial and "calm"-ness time series. This gives us 41 trading days. We provide the relevant summary statistics for our financial data in Section **4.1**.

3.5 Functional Form Specifications

Following BMZ[38], our first model looks at the first difference of DJIA values, defined as:

$$D_t \coloneqq DJIA_t - DJIA_{t-1} \tag{16}$$

For $n \in \{1, \dots, 5\}$ we estimate the following nested ARDL models[39]:

$$L_1 \colon D_t = \alpha + \sum_{i=1}^{n} \beta_i \, D_{t-i} + \epsilon_t \tag{17}$$

$$L_2 \colon D_t = \alpha + \sum_{i=1}^{n} \beta_i \, D_{t-i} + \sum_{i=1}^{n} \gamma_i \, CA_{t-i} + \epsilon_t \tag{18}$$

[38] 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 regression's left hand side as a log-return. However, because the time period under investigation is so short, the outcomes of our hypothesis tests do not hinge on whether we take first differences in the level or the first difference in the log-level of the DJIA: our qualitative Granger causality results are the same for either choice of left hand side. For simplicity, we follow BMZ's approach and only present Granger causality results for level-differences.

[39] While BMZ test 7 lags in TMP, they only find significance at the 5% level on the first 5 lags.

where the error term is ϵ_t, and α is a constant. CA_t is defined above. We perform the F-test for Granger causality of mood. The hypotheses are:

$$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_n = 0 \tag{19}$$

$$H_A: \gamma_1 \neq 0 \vee \gamma_2 \neq 0 \vee \cdots \vee \gamma_n \neq 0 \tag{20}$$

If our F-statistic is significant at the 5% level, we can conclude that Twitter mood, as stated by BMZ, Granger causes changes in the DJIA. Choosing our test interval to border the range of BMZ's, our ex--sample tests are more powerful than other replications. It is possible that after TMP was published, traders attempted to utilize the results Twitter mood in their strategies, if Twitter "calm"-ness represented an arbitrage opportunity, then this trading should remove the predictive power of Twitter "calm"-ness from subsequent data. This would explain the failure of DCM and the inability to replicate BMZ's finding. Since TMP was released on October 14, 2010, we expect to discover the arbitrage opportunity via the Twitter mood effect in our data.

Because there is no theoretical basis for believing that BMZ's proposed functional form of DJIA-Twitter "calm"-ness relationship is correct, we test a number of alternate functional forms. Since the time frame is small, trends are unlikely to be detectable over this time frame. To validate this intuition, we perform Dickey-Fuller tests on both our DJIA index and Twitter mood series.

$$L_3 : DJIA_t = \alpha + \sum_{i=1}^{n} \beta_i \, DJIA_{t-i} + \epsilon_t \tag{21}$$

$$L_4 : DJIA_t = \alpha + \sum_{i=1}^{n} \beta_i \, DJIA_{t-i} + \sum_{i=1}^{n} \gamma_i \, CA_{t-i} + \epsilon_t \tag{22}$$

L_3 and L_4 differ from L_1 and L_2 in that they use levels while the previous models use first-differences of the DJIA. We perform an F-test on our coefficients.

Finally, we point out that measures given by (12) and (13) of Twitter mood do not take into account Twitter's rapidly growing user base. Failing to do so can cause a trend to be present in our mood measurements using these specifications, resulting in non-stationarity and biased parameter estimates. Anticipating this, we specify L_5 and L_6 taking the first-difference of Twitter mood. This gives us the models (with $n \in \{1,2,\dots,5\}$) L_5, L_6, L_7, and L_8:

$$L_5 : D_t = \alpha + \sum_{i=1}^{n} \beta_i \, DJIA_{t-i} + \epsilon_t \tag{23}$$

$$L_6 : D_t = \alpha + \sum_{i=1}^{n} \beta_i \, DJIA_{t-i} + \sum_{i=1}^{n} \gamma_i \, \Delta CA_{t-i} + \epsilon_t \tag{24}$$

We conduct Granger causality tests as above, via the F-statistic, to measure the impact of mood first differences on the DJIA. Models L_1, L_3, L_5, and L_7 are estimated only for the purpose of conducting likelihood ratio tests for L_2, L_4, L_6, and L_8, the latter of which includes Twitter mood (either in its level or first-difference form) in addition to lags of the DJIA (again, in either level or first-difference form) as predictors.

$$L_7 : DJIA_t = \alpha + \sum_{i=1}^{n} \beta_i \, DJIA_{t-i} + \epsilon_t \tag{25}$$

$$L_8 : DJIA_t = \alpha + \sum_{i=1}^{n} \beta_i \, DJIA_{t-i} + \sum_{i=1}^{n} \gamma_i \, \Delta CA_{t-i} + \epsilon_t \tag{26}$$

4. Results and Discussion

In section **4.1**, we present the summary statistics for each of the time series generated above. We also provide evidence that the time period under consideration exhibited highly irregular financial market behavior and thus we should be careful not to generalize our findings to other

time periods without further data. In section **4.2**, we present the results of our F-statistics and Granger causality tests. In section **4.3**, we discuss major sources of error in our mood analysis and "calm"-ness time series. In section **4.4**, we discuss our results in the context of the big data regime and future planned work in this area.

4.1 Summary Statistics and Exploratory Analysis

In Table 1 we present the summary statistics for the DJIA and our eight mood time series from October 31st to December 31st, 2008. In section **4.1.1**, we conduct elementary exploratory analysis on our "calm"-ness time series and visually confirm that the stylized features of these time series match features of the standardized GPOMS "calm"-ness measure provided in TMP.

Table 1.

Variable	Mean	Median	SD	Min	Max	1st Quartile	3st Quartile
$DJIA_t$	8622	8598	364	7552	9625	8450	8812
D_t	-13	-5	285	-680	553	-215	184
CA_t^1 (I)	65	141	524	-2642	707	21	271
CA_t^2 (I)	1.02	1.02	0.0759	0.689	1.20	1.00	1.05
CA_t^3 (I)	0.000856	0.000193	0.00643	-0.0311	0.0132	0.000193	0.00386
CA_t^4 (I)	$8.11(10^{-6})$	$1.84(10^{-5})$	$8.96(10^{-6})$	$8.11(10^{-6})$	$8.05(10^{-5})$	$1.52(10^{-6})$	$2.06(10^{-5})$
CA_t^1 (A)	825	842	281	-223	1487	659.8	992
CA_t^2 (A)	1.18	1.17	0.0669	0.974	1.47	1.16	1.22
CA_t^3 (A)	0.0150	0.0145	0.00455	-0.00262	0.0314	0.0131	0.0175
CA_t^4 (A)	$2.36(10^{-5})$	$2.10(10^{-5})$	$1.38(10^{-5})$	$1.15(10^{-5})$	$1.21(10^{-4})$	$1.74(10^{-5})$	$2.49(10^{-5})$
ΔCA_t^1 (I)	21	59	500	-1628	2315	-126	156
ΔCA_t^2 (I)	0.00602	0.00585	0.0595	-0.166	0.253	-0.0283	0.0377
ΔCA_t^3 (I)	0.000452	0.000955	0.00505	-0.0127	0.0259	-0.00208	0.00272
ΔCA_t^4 (I)	$-1.00(10^{-5})$	$-2.00(10^{-7})$	$8.02(10^{-6})$	$-5.32(10^{-5})$	$9.60(10^{-6})$	$-1.70(10^{-6})$	$1.80(10^{-6})$
ΔCA_t^1 (A)	14	21	254	-501	1258	-116	113
ΔCA_t^2 (A)	0.0000256	0.00173	0.0627	-0.217	0.214	-0.0338	0.0300
ΔCA_t^3 (A)	0.00000821	0.000128	0.00446	-0.0119	0.0192	-0.00198	0.00215
ΔCA_t^4 (A)	$-1.60(10^{-6})$	$-3.00(10^{-7})$	$1.21(10^{-5})$	$8.40(10^{-5})$	$1.64(10^{-5})$	$-2.00(10^{-6})$	$1.60(10^{-6})$

Table 1: This table contains summary statistics for all of our time series. (I) indicates that the measure was obtained by inspection and (A) indicates that the measure was obtained with thesaurus augmentation. With the exception of "calm"-ness measures 2 and 4, all values are rounded to the nearest unit. Specifications 2 and 4 are presented to 3 digits accuracy after leading zeros. All differencing takes place after weekends have been removed.

Notice that in Table 1 above, for each specification, our thesaurus generated daily "calm"-ness measures are greater than the corresponding inspection generated "calm"-ness measure; we attribute this to the higher number of "composed" words in the 1911 Thesaurus still in use relative to the number of "anxious" words in the 1911 Thesaurus still in use. In our financial summary statistics, we can observe three irregularities over these forty days that suggest our parameter estimates will not be valid for other time periods. First, our sample occurred in the same quarter as the financial crisis and subsequent Troubled Asset Relief Program (TARP) bailouts. Second, it features a historically atypical downward trend in DJIA values. Third, our sample includes several of the largest single-day increases in DJIA history. Nonetheless, our statistical tests are valuable in the sense that, if we fail to find a Twitter "calm"-ness effect, it is more likely that the original findings were spurious or held conditional on particular unobserved economic conditions (non-robust).

4.1.1 Some Elementary Exploratory Analysis

In Table 2, we conduct Augmented Dickey-Fuller tests to find unit roots in our time series. Unit roots bias the coefficients in our time series regressions downward.

Table 2.

Variable	ADF p-value (1 lag)
$DJIA_t$	0.267
D_t	**0.000**
CA_t^1 (I)	**0.011**
CA_t^2 (I)	**0.000**
CA_t^3 (I)	0.067
CA_t^4 (I)	0.065
CA_t^1 (A)	**0.000**
CA_t^2 (A)	**0.000**
CA_t^3 (A)	**0.000**
CA_t^4 (A)	**0.000**
ΔCA_t^1 (I)	**0.000**
ΔCA_t^2 (I)	**0.000**
ΔCA_t^3 (I)	**0.000**
ΔCA_t^4 (I)	**0.000**
ΔCA_t^1 (A)	**0.000**
ΔCA_t^2 (A)	**0.000**
ΔCA_t^3 (A)	**0.000**
ΔCA_t^4 (A)	**0.000**

Table 2: This table contains the p-values for all our hypothesis tests. Results significant at the 5% level are bolded. All results are reported to three decimals. In this case, significance means that we have rejected the hypothesis of a stochastic unit root using the ADF test.

As we suggested earlier, the DJIA and two of our mood measures contain a unit root. Surprisingly, it is the mood measures that scale by the number of Tweets containing a unit root. As we point out in Section **3.5**, it is a priori likely that our other collective mood time series contain unit roots, particular specifications (12) and (13), however we do not detect any unit roots over the time frame. Nonetheless, none of our differenced mood measures contain a unit root and so we can be certain that the tests on model L_6 will not suffer from downward biased coefficients.

Figures 5 and 6 contain pairwise comparison graphs. We note that, by construction, all of our mood measurements exhibit pairwise linear correlation except mood measure 15, which appears to be linearly independent. It also appears that while there are significant differences between the lexicons obtained by inspection and thesaurus augmentation (see Appendices A-D), the time series generated by counting words matching each of these lexicons exhibit significant positive linear correlation, providing evidence for the construct validity of our lexicons obtained by inspection.

Figure 5.

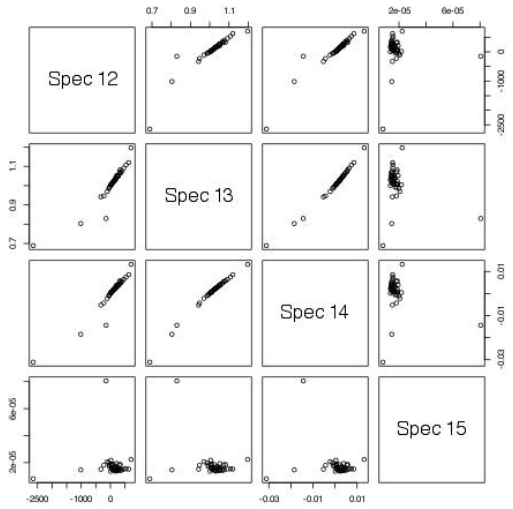

Figure 5: This figure contains the pairwise plots of the time series of all our specifications with lexicons obtained by inspection. All of our time series are strongly linearly correlated with the exception of Specification 15, which appears approximately linearly independent from the other three specifications. The same results hold for the time series derived from lexicons obtained by thesaurus augmentation.

Figure 6.

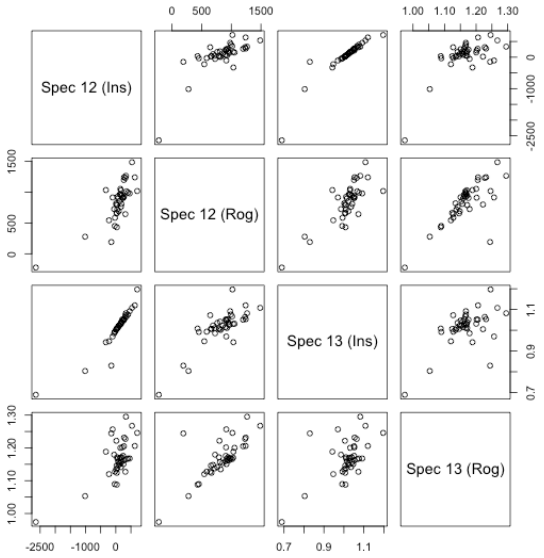

Figure 6: This figure contains the pairwise plots of the time series Specifications (12) and (13), comparing the time series values for lexicons obtained by inspection versus those obtained by thesaurus augmentation. They are strikingly similar, suggesting that the time series derived from the lexicon obtained by inspection provides a valid measure of "calm"-ness.

Finally, Figure 7 provides a comparison of one of our Twitter "calm"-ness measures side-by-side with BMZ's GPOMS CALM measure.[40] Our Twitter "calm"-ness measure replicates many of the salient features of BMZ's "calm"-ness including the sudden drop on November 4th, 2008 (coinciding with the election of Barack Obama) and the immediate post-election recovery in the "calm"-ness time series. The normalization procedures used in TMP, as well as the coarse-ness of TMP's Figure 2 prevent us from making a more detailed comparison between the GPOMS CALM time series and our own simple "calm"-ness time series.

[40] In fact, all of our time series specification-lexicon pairs match the "CALM" measure provided by BMZ in the sense that they reproduce the pre-election increase in anxiety and post-election decrease in anxiety.

Figure 7.

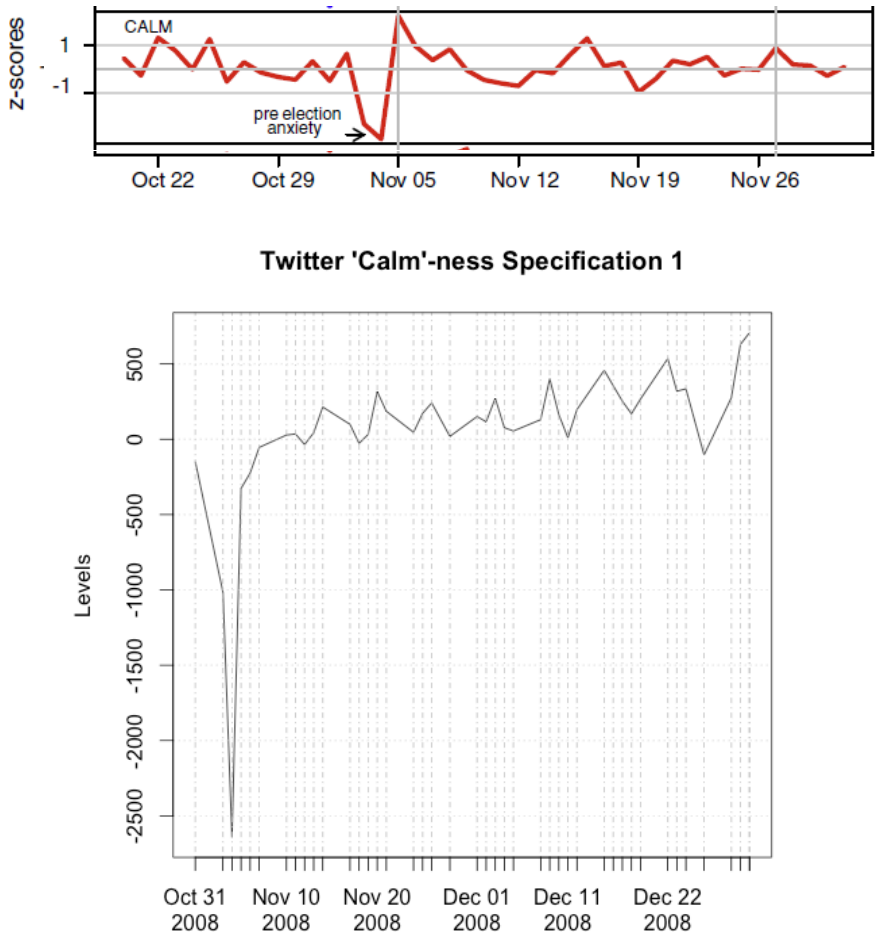

Figure 7: The top graphic is taken from Figure 2 in TMP. The bottom graphic is taken from our first "calm"-ness time series specification with a lexicon obtained by inspection. Both show the variation in our respective "calm"-ness measures against time. BMZ's proprietary GPOMS tool weights each "composed" or "anxious" term according to a proprietary system, while our measures simply count the number of "composed" words each day and subtract that number from the number of "anxious" words. Nonetheless, our simple measure visually reproduces many of the increases and declines of TMP's CALM measure including what BMZ term "pre-election anxiety" and the subsequent post-election recovery.

4.2 Inferential Statistics

For each pair of models, we conduct 40 hypothesis tests because we have five possible lags and eight "calm"-ness time series. Since we have four pairs of models, we are conducting a total 160 hypothesis tests. Because we are testing a large number of hypotheses, we expect several p-values to appear significant by chance. Using any standard multiple hypothesis test correction method[41], none of the p-values we find are significant at the 10% level. For the remainder of this discussion, we treat the single-hypothesis tests as valid. In Table 3, we present the (flawed) single-hypothesis test p-values from both TMP along-side our own.

If TMP's result is robust, we expect to find lags 2 through 5 are significant and that Twitter "calm"-ness increases Granger cause stock market increases (i.e. positive coefficients). Overall, 17 single-hypothesis tests find that Twitter "calm"-ness increases the stock market. We present the original tests from TMP below, along-side our tests, in Table 2.

None of the specification-lexicon pairs exhibit the same spectacular Granger causality pattern found in TMP. In Section **2.3**, we presented two features of the collective mood-DJIA relationship identified in TMP that do not concord with the behavioral finance literature. First, the "calm"-ness effect only appears from the second lag onwards. Second, it persists for several days. None of our model-specification-lexicon triples exhibit both features. In particular, none of the tests with insignificant first lags and significant second lags on the "calm"-ness effect exhibit persistence.[42] The converse is also true: none of the model-specification-lexicon triples with

[41] In this case, we used the Bonferroni, Holm, Hommel and Hochberg multiple hypothesis test corrections using p.adjust implementation in R's stats package. To err on the side of liberalism, we conducted all corrections on a model-by-model basis (i.e. taking n to be 40, rather than 160, for each correction). Our insignificance result is independent of which particular method we select (Bonferroni being the most conservative and Hommel being the least conservative) we used and holds for all n>30.

[42] Of course, the coefficient of the first difference of collective mood contains information about two days; therefore, one could argue that the four specification-lexicon pairs with significant first lags in Model 8 actually reflect information about the significance of the second day's "calm"-ness information and exhibit persistence in the "calm"-ness effect. The coefficients in Table 4 provide support for this hypothesis:

multiple significant lags have insignificant first lags. In other words, all of the model-specification-lexicon triples with persistence of any kind do not have the same time-to-appear characteristic in TMP and all of the model-specification-lexicon triples with the same time-to-appear characteristic as TMP do not exhibit the same persistence as in TMP.

n Table 4, we present coefficients for selected models. We note that none of the coefficients are large enough to suggest the presence of an arbitrage opportunity.[43] Furthermore, many of the most significant p-values are associated with models that, counter-intuitively and contra the findings in TMP[44], assign large negative coefficients to "calm"-ness.[45] Because of the inconsistency in the sign of the effect of collective mood changes across models (but within specification-lexicon pairs) and the overall insignificance of the results, contra TMP, we cannot say that the predictive power inherent in our collective mood change dataset provides evidence against the RWH.

---

relatively high "calm"-ness scores two days ago correspond with a negative first difference yesterday. The first lag's first difference has a significant negative coefficient, indicating higher than average "calm"-ness scores two days ago have a positive effect on the change in the DJIA today. Unfortunately, persistence of any kind only occurs in those time series derived from lexicons obtained by thesaurus augmentation and still does not match the persistence pattern discovered by BMZ. Contra the p-values reported in TMP, the third and fourth day's information does not register as significant for any specification-lexicon pairs in Model 8. In this case, the relevant question is: why would differencing preserve the second lag of the daily mood score's significance but not the third or fourth lag of the daily mood score's significance, since the p-values for all three lags in TMP are of similar size?

[43] To see this, compare the summary statistics in Table 1 with the coefficients in Table 4.

[44] In fact, BMZ do not report coefficients of their ARDL models so we cannot be *certain* that they do not find negative signed mood coefficients, but we assume they would have reported counter-intuitive negative coefficients for collective "calm"-ness had they occurred. Furthermore, in his media appearances, Bollen suggests that the more "calm"-ness forecasts a higher stock market. Relatively large, significant and negative coefficients for any lag would surely have warranted a mention. For several of the model-specification-lexicon triples we estimate, increases in "calm"-ness have *net* negative effects on the DJIA.

[45] In fact, a very curious finding is that eight of the seventeen significance results are in the five lag model. Our five lag models typically have coefficient sign-reversals on collective mood changes similar to those found in Tetlock (2007).

Table 3.

Model 2	TMP	CA_t^1 (I)	CA_t^1 (A)	CA_t^2 (I)	CA_t^2 (A)	CA_t^3 (I)	CA_t^3 (A)	CA_t^4 (I)	CA_t^4 (A)
1 Lag	0.272	0.145	0.534	0.185	0.974	0.200	0.946	0.754	0.400
2 Lags	**0.013**	0.056	0.219	**0.049**	0.258	0.052	0.241	0.448	0.346
3 Lags	**0.022**	0.250	0.757	0.275	0.546	0.240	0.556	0.137	0.099
4 Lags	**0.030**	0.754	0.923	0.764	0.988	0.756	0.993	0.478	0.413
5 Lags	**0.036**	0.779	0.566	0.782	0.963	0.792	0.966	0.284	0.359
Model 4	TMP	CA_t^1 (I)	CA_t^1 (A)	CA_t^2 (I)	CA_t^2 (A)	CA_t^3 (I)	CA_t^3 (A)	CA_t^4 (I)	CA_t^4 (A)
1 Lag	0.272	0.975	0.675	0.506	0.589	0.456	0.551	0.516	0.464
2 Lags	**0.013**	0.694	0.910	0.911	0.110	0.855	0.142	**0.012**	**0.009**
3 Lags	**0.022**	0.738	0.723	0.668	0.775	0.683	0.778	0.163	0.183
4 Lags	**0.030**	0.441	0.886	0.581	0.952	0.511	0.950	0.215	0.254
5 Lags	**0.036**	0.579	0.910	0.664	0.989	0.618	0.990	0.251	0.285
Model 6	TMP	CA_t^1 (I)	CA_t^1 (A)	CA_t^2 (I)	CA_t^2 (A)	CA_t^3 (I)	CA_t^3 (A)	CA_t^4 (I)	CA_t^4 (A)
1 Lag	0.272	0.589	0.508	0.760	0.090	0.589	0.099	0.176	0.176
2 Lags	**0.272**	0.406	0.847	0.752	0.300	0.672	0.338	**0.032**	**0.028**
3 Lags	**0.013**	0.795	0.882	0.942	0.556	0.876	0.566	0.202	0.176
4 Lags	**0.022**	0.848	0.948	0.858	0.986	0.859	0.993	0.578	0.540
5 Lags	**0.030**	**0.039**	**0.005**	0.056	**0.039**	**0.045**	**0.013**	0.373	0.458
Model 8	TMP	CA_t^1 (I)	CA_t^1 (A)	CA_t^2 (I)	CA_t^2 (A)	CA_t^3 (I)	CA_t^3 (A)	CA_t^4 (I)	CA_t^4 (A)
1 Lag	0.272	0.467	0.672	0.946	**0.033**	0.734	**0.045**	**0.007**	**0.006**
2 Lags	**0.013**	0.324	0.575	0.275	0.538	0.264	0.507	0.489	0.383
3 Lags	**0.022**	0.580	0.857	0.647	0.785	0.621	0.786	0.553	0.481
4 Lags	**0.030**	0.509	0.849	0.578	0.968	0.545	0.969	0.538	0.496
5 Lags	**0.036**	0.536	**0.048**	0.617	**0.030**	0.568	**0.020**	0.637	0.523

Table 3: This table contains the p-values for all our hypothesis tests. Results significant at the 5% level are bolded. All results are reported to three decimals. All of these results come from individual hypothesis tests and have not been corrected for the possibility of multiple comparison bias. Making said correction for any reasonable parameter choice and method (Bonferroni, Holm, etc.) would render all results in this table insignificant.

Table 4.

Model 2	CA_t^2					**Model 4**	CA_t^4				
1 Lag	827	1034	1316	1750	1532	1 Lag	2681912	-33545906	**-48915638**	-42061637	-44890913
2 Lags	N/A	656	667	-272	-635	2 Lags	N/A	**10156090**	17657919	28305942	2325459
3 Lags	N/A	N/A	-218	-188	187	3 Lags	N/A	N/A	-7110796	-33106438	-8423733
4 Lags	N/A	N/A	N/A	144	278	4 Lags	N/A	N/A	N/A	1029635	-28076529
5 Lags	N/A	N/A	N/A	N/A	286	5 Lags	N/A	N/A	N/A	N/A	5216206
Model 6	CA_t^1					**Model 6**	CA_t^2				
1 Lag	-0.053	-0.031	-0.064	0.220	0.126	1 Lag	-259	187	-230	803	644
2 Lags	N/A	0.131	0.061	-0.061	-0.252	2 Lags	N/A	652	276	-159	-1525
3 Lags	N/A	N/A	0.007	-0.069	**-0.331**	3 Lags	N/A	N/A	-265	-525	-1705
4 Lags	N/A	N/A	N/A	-0.054	**-0.264**	4 Lags	N/A	N/A	N/A	-574	-1727
5 Lags	N/A	N/A	N/A	N/A	**-0.328**	5 Lags	N/A	N/A	N/A	N/A	**-268**
Model 6	CA_t^3					**Model 6**	CA_t^4				
1 Lag	-5000	810	-5527	11778	10385	1 Lag	-5889865	**-39320612**	-37223489	-42181737	-49111202
2 Lags	N/A	8226	3066	-2602	-14663	2 Lags	N/A	**10064580**	6912706	-3826306	-10662276
3 Lags	N/A	N/A	-2394	-6384	-18308	3 Lags	N/A	N/A	927812	-15411115	-27823688
4 Lags	N/A	N/A	N/A	-5871	-18230	4 Lags	N/A	N/A	N/A	-1992605	-29681683
5 Lags	N/A	N/A	N/A	N/A	**-28489**	5 Lags	N/A	N/A	N/A	N/A	-1723519
Model 8	CA_t^1					**Model 8**	CA_t^4				
1 Lag	-0.061	0.028	0.131	0.337	0.323	1 Lag	**-10679760**	-22440426	-26779467	-30280330	-39517511
2 Lags	N/A	0.125	0.164	0.202	0.086	2 Lags	N/A	3144895	13024961	4945875	-6261621
3 Lags	N/A	N/A	0.027	0.104	0.055	3 Lags	N/A	N/A	-2322772	3026038	-15943958
84 Lags	N/A	N/A	N/A	0.035	-0.006	4 Lags	N/A	N/A	N/A	-6075857	-7980925
5 Lags	N/A	N/A	N/A	N/A	-0.120	5 Lags	N/A	N/A	N/A	N/A	-6281115

Table 4: This table contains the coefficients for selected models. Only coefficients from models with lexicons obtained by inspection are reported. Coefficients significant at the 5% level are bolded. All results are rounded to whole numbers or reported to three decimals. All of these results come from individual hypothesis tests and have not been corrected for the possibility of multiple comparisons. Because our Granger causality tests in Table 3 evaluate the significance of the coefficients collectively, significant coefficients in this table do not necessarily correspond with significant results in Table 3.

4.3 Major Sources of Error[46]

Our first source of error stems from our potentially biased coefficient estimates. Both our explanatory variables and our response variables contain unit roots. Stock and Watson (2007) point out that non-stationary time series will lead to ARDL regression coefficients being biased towards zero. In our case, the unit root in mood biases the coefficients of L_2 and L_4, towards zero, while the potential unit root in the DJIA may bias the coefficients of L_8 towards zero.

Our second major source of error comes from large sample variance in our estimators. The high variance in our estimators is one probable cause for the sign changes across models in Table 4. Resource constraints limited the amount of Twitter data we could acquire to 42 trading days. Given the atypical behavior of the DJIA over this time-period, we can only use our ARDL models in the context of evaluating TMP and cannot generalize our findings from these models by themselves to other time periods.

Potentially the largest problem is the measurement error in our moods. Lorr and McNair (1984) introduced the POMS-Bi as an update to the POMS in part because the lexicons that people used to describe their moods had changed since the publication of the POMS. We cannot be sure that the language we use to describe moods has not changed significantly since then. We chose our algorithm because it was the simplest and most computationally tractable technique available for measuring moods. Agarwal et al.'s 2011 analysis of Twitter sentiment analysis algorithms

[46] Several commentators have objected to TMP because the sample of Twitter users is not representative of the U.S. population, based on characteristics including age, income, and sociological factors. Many of the users of Twitter in 2008 were early adopters, and likely to be display a much younger and wealthier profile than the average individual across the United States as a whole. We do not find these criticisms warranted for econometric and philosophical reasons. Econometrically, if the aforementioned sources of mood error are unbiased (i.e. have mean zero for each day), then our estimated mood coefficients are biased downward. With more accurate mood measurements we are likely to find a larger effect, not a smaller one. Philosophically, it may simply be the case that Twitter users' collective mood (i.e. Twitter mood), rather than the collective mood of the nation, predicts the stock market. We do not attempt to, in any sense, correct for the skewed demographic of Twitter.

suggested an upper bound of 76% accuracy for algorithms working with naïve partitions. It is likely our algorithm for detecting moods, which is based on a more complicated partition but uses much less sophisticated term-matching, achieved an accuracy at understanding moods below 76%. Our bag-of-words technique discards all grammatical information and in the worst case, our mood measurement technique gets the sign for mood wrong. To give a concrete example, consider the word that characterized the positive axis for “composed”. Figure 8 depicts how "composed" can be used on Twitter. In the sample we present, it is most commonly used to depict the action of creating music rather than the user's mood. Derick Woodward writes, "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." In this context, it is suggested that the user is clearly anxious although his Tweet is associated with "composed." Woodward's Tweet would have increased our composed-anxious score, when it really should have caused a reduction in our daily composed-anxious score.

Figure 8.

Figure 8: This snapshot provides examples of how the word "composed" is actually used on Twitter. Notice that four of the six Tweets use "composed" differently than the POMS-Bi uses the term.

4.4 Spurious Correlation in the Age of Big Data: a Context for Future Work

Many of the statistical problems caused by big data, in particular spurious findings, can be cured with big ex-sample validation. Indeed, virtually all of the problems we have identified with our own findings could be remedied by sufficiently increasing the sample size. The variance of our estimators and the bias in our parameter estimates for the models L_2 and L_4 would be reduced with a large enough sample size. Furthermore, as our sample grows larger, the impact of the financial crisis on our parameter estimates should shrink. Specifically, if we had access to the universe of Twitter data, from 2008 we could conduct a direct comparison of the ARDL p-values that we obtain from estimating the model using data from February 28 to November 3, 2008 with those BMZ obtain for the same time period.[47]

Recent work in so-called "Deep Learning" algorithms has enabled breakthroughs in measuring sentiment in online movie reviews that achieve accuracy of over 85%; this technique can be adapted to measure mood. In other behavioral finance literature, Baker and Wurgler (2007) suggest that we would have better luck discovering mood effects in small, illiquid stocks rather than the DJIA.[48] Statman and Fisher (2002) suggest we should try to measure investor sentiment, rather than the sentiment of everyone on Twitter to predict the stock market. Tetlock (2007) suggests that our mood measurement technique would produce higher quality output with the inclusion of media sources besides Twitter and that we should test non-linear functional form relationships between Twitter mood and the stock market. Antweiler and Murray (2004) motivate an alternate approach of using Twitter mood to predict DJIA price volatility rather than the actual change in DJIA price.

[47] In fact, we have already obtained this data and will be presenting our results in "Two Essays in Empirical Asset Pricing" (forthcoming).

[48] In fact, much of the literature suggests that mood effects of the kind identified in TMP are unlikely to affect large-cap, closely watched equities like those comprising the DJIA. There are, however, several credible studies finding that mood and sentiment effects do affect the prices of large-cap equities so we do not identify this as a discrepancy between TMP and the behavioral finance canon.

5. Conclusions

While BMZ do not provide enough information to replicate or fail-to-replicate their main empirical findings, it is certain that changes in Twitter "calm"-ness under a wide variety of specifications did not predict the DJIA in November and December of 2008. Because TMP[49] has had a dramatic effect on both the literature of behavioral finance and the actual investment practices of hedge funds, we hope that this work that can bring further scrutiny to TMP. We believe our findings increase the likelihood that the results presented in TMP were spurious or otherwise non-robust but, because the time series we worked with was so short, more research is necessary before any kind of strong conclusions can be drawn.

[49] http://sellthenews.tumblr.com/post/26542555585/thanks-for-your-response-question-why-did-you-address, last accessed on 2014/11/28.

Bibliography

Agarwal, A., Xie, B., Vovsha, I. Rambow, O., & Passonneau, R. (2011)**.** "Sentiment Analysis of Twitter Data", *Proceedings of the Workshop on Languages in Social Media* (30-38). Association for Computational Linguistics.

Akerlof, G. A., & Shiller, R. J. (2010). *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism.* Princeton University Press.

Antweiler, Werner, and Murray Z. Frank. 2004. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", *Journal of Finance*, 59(3): 1259-1294.

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

Bollen, Johan and Alberto Pepe and Huina Mao. 2011. "Modeling Public Mood and Emotion: Twitter Sentiment and Socio-economic Phenomena", *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011: 450-453.

Bollen, Johan and Huina Mao. 2011. "Twitter Mood Predicts the Stock Market", *2011 Socionomics Summit*.

Baker, Malcom and Jeffrey Wurgler. 2007. "Investor Sentiment in the Stock Market", *Journal of Economic Perspectives*, 21(2): 129-151.
Capra, C. Monica. (2004). Mood-Driven Behavior in Strategic Interactions. *American Economic Review*, 367-372.

Edmans, Alex and Diego Garcia and Øyvind Norli. 2007. "Sports Sentiment and Stock Returns", *Journal of Finance*, 62(4): 1967-1998.

Fan, Jianqing and Fang Han and Han Liu. 2014. "Challenges of Big Data Analysis", *National Science Review*, 00: 1-22.

Hirshleifer, David and Tyler Shumway. 2003. "Good Day Sunshine: Stock Returns and the Weather", *Journal of Finance*, 58(3): 1009-1032.

Jansen, **W. Jos and Niek J. Nahuis.** 2003. "The stock market and consumer confidence: European evidence", *Economics Letters,* 79(2003): 89-98.

Kamstra, Mark, A. Kramer and Maurice D. Levi. 2003. "Winter Blues: Seasonal Affective Disorder (SAD) and Stock Market Returns", *American Economic Review*, 93 (1), 32.

Keynes, J. M. (1936). *General Theory of Employment, Interest and Money*. Atlantic Publishers & Dist.

Kuleshov, Volodomyr. 2011. "Can Twitter Predict the Stock Market?"

Lemmon, Michael and Evgenia Portniaguina. 2006. "Consumer Confidence and Asset Prices: Some Empirical Evidence", *Review of Financial Studies*, 19(4):1499-1529.

Liu, Bing. 2012. "Sentiment Analysis and Opinion Mining." *Synthesis Lectures on Human Language Technologies*, 5(1), 1-167.

Lorr, Maurice and Douglas McNair. 1984. "Profile of Mood States, Bipolar Form (POMS-BI)", Education and Testing Service, San Diego.

Lucas, Robert. 1978. "Asset Prices in an Exchange Economy", *Econometrica*, 46(6): 1429-1445.

Malkiel, Burton. 2012. *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*. WW Norton & Company.

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

Pettengill, Glenn N. 2003. "A Survey of the Monday Effect Literature", *Quarterly Journal of Business and Economics*, 42(3/4): 3-27.

Sharma, Jayant and Aniruddh Vyas. 2012. "Twitter Sentiment Analysis."

Shen, **Pu**. 2002. "Market-Timing Strategies that Worked."

Shiller, Robert. 1981. "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?", *American Economic Review*, 71(3): 421-436.

Statman, Meir and Kenneth Fisher. 2003. "Consumer Confidence and Stock Returns", *Journal of Portfolio Management*, 30(1): 115-127.

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

Watson, Mark W. and James Stock. 2007. *Introduction to Econometrics*. New York: Pearson Education.

Appendix A "Calm" Words Obtained by Inspection

1. calm
2. refreshed
3. pure
4. tranquil
5. quiet
6. confident
7. collected
8. comfortable
9. serene
10. composed
11. behaved
12. soothed
13. soothing
14. established
15. untroubled
16. balanced
17. uncrowded
18. pleasant
19. reassured
20. rejuvenated
21. stable
22. relaxed
23. peaceful
24. rested
25. controlled
26. measured
27. steady
28. thoughtful
29. prosperous
30. settled
31. even tempered
32. even-tempered
33. undisturbed
34. laid back
35. laid-back
36. unflappable
37. at ease
38. nourished
39. harmonious
40. secure
41. safe
42. patient
43. resolute
44. pampered
45. orderly
46. affluent
47. bountiful
48. unruffled
49. civil
50. carefree
51. mellow
52. blissful
53. still
54. placid
55. cozy
56. cosy
57. cool
58. credible
59. dignified
60. idyllic
61. restful
62. restored
63. sheltered
64. sound
65. successful
66. sustainable
67. gentle
68. tolerant
69. together
70. phlegmatic
71. honest
72. organized
73. productive
74. unperturbed
75. content
76. centered
77. fulfilled
78. profitable
79. professional
80. satisfied
81. welcomed
82. healthy

Appendix B "Anxious" Words Obtained by Inspection

1. anxious
2. bewildered
3. excited
4. nervous
5. restless
6. confused
7. helpless
8. fearful
9. isolated
10. troubled
11. uncertain
12. depressed
13. searching
14. afraid
15. aggressive
16. agitated
17. annoyed
18. apprehensive
19. dizzy
20. embarrassed
21. bored
22. concerned
23. distressed
24. empty
25. exhausted
26. frustrated
27. high strung
28. impatient
29. insecure
30. irritable
31. irritated
32. listless
33. jittery
34. miserable
35. overwhelmed
36. paranoid
37. preoccupied
38. reactive
39. reluctant
40. scared
41. stressed
42. stressed out
43. suspicious
44. desperate
45. tired
46. uncomfortable
47. under stress
48. uneasy
49. unhappy
50. unsettled
51. unsure
52. upset
53. vulnerable
54. withdrawn
55. worried
56. eager
57. disturbed
58. flustered
59. guilty
60. panicky
61. tense
62. fretful
63. edgy
64. antsy
65. keyed up
66. worked up
67. jumpy
68. uptight
69. twitchy
70. intense
71. overwrought
72. negative
73. frightened
74. watchful
75. disgruntled
76. skittish
77. neurotic
78. timid
79. trepidation
80. shaky

Appendix C "Calm" Words Obtained by Roget's 1911 Thesaurus

1. nonadhesive
2. immiscible
3. incoherent
4. detached
5. loose
6. baggy
7. slack
8. lax
9. relaxed
10. flapping
11. streaming
12. disheveled
13. segregated
14. unconsolidated
15. like grains of sand
16. uncombined
17. weak
18. feeble
19. impotent
20. unnerved
21. sapless
22. powerless
23. weakly
24. unstrung
25. flaccid
26. nervous
27. moderate
28. lenient
29. gentle
30. mild
31. mellow
32. cool
33. sober
34. temperate
35. reasonable
36. measured

37. tempered
38. calm
39. unruffled
40. quiet
41. tranquil
42. still
43. slow
44. smooth
45. untroubled
46. tame
47. peaceful
48. peaceable
49. pacific
50. halcyon
51. horizontal
52. level
53. even
54. plane
55. flat
56. flat as a billiard table
57. flat as a bowling green
58. alluvial
59. calm as a mill pond
60. smooth as glass
61. quiescent
62. motionless
63. moveless
64. fixed
65. stationary
66. immotile
67. at rest at a stand
68. at a standstill
69. at anchor
70. stock
71. standing still
72. sedentary
73. untraveled
74. stay-at-home
75. becalmed
76. stagnant
77. unmoved
78. undisturbed
79. restful
80. cataleptic
81. immovable
82. stable
83. sleeping
84. inactive
85. silent
86. still as a statue
87. still as a post
88. still as a mouse
89. still as death
90. vegetative
91. vegetating
92. transparent
93. pellucid
94. lucid
95. diaphanous
96. translucent
97. limpid
98. clear
99. serene
100. crystalline
101. clear as crystal
102. vitreous
103. glassy
104. hyaline
105. hyaloid
106. leisure
107. leisurely
108. deliberate
109. at leisure
110. at ones ease
111. at loose ends
112. at a loose end
113. reposing
114. unstrained
115. bloodless
116. neutral
117. licensed
118. unbridled
119. anarchical
120. unauthorized
121. unwarranted
122. easy-going
123. placid
124. quiet as a mouse
125. cool as a cucumber
126. cool as a custard
127. undemonstrative
128. composed
129. collected
130. unexcited
131. unstirred
132. unperturbed
133. unimpassioned
134. unresisting
135. unafflicted
136. unmolested
137. at rest
138. snug
139. comfortable
140. in ones element

Appendix D "Anxious" Words Obtained by Roget's 1911 Thesaurus

1. weak
2. feeble
3. impotent
4. relaxed
5. unnerved
6. sapless
7. powerless
8. weakly
9. unstrung
10. flaccid
11. nervous
12. broken
13. lame
14. withered
15. shattered
16. shaken
17. crazy
18. shaky
19. palsied
20. decrepit
21. hard
22. rigid
23. stubborn
24. stiff
25. firm
26. starch
27. starched
28. stark
29. unbending
30. unlimber
31. unyielding
32. inflexible
33. tense
34. indurate
35. indurated
36. gritty
37. proof
38. vigorous
39. powerful
40. forcible
41. trenchant
42. incisive
43. impressive
44. sensational
45. tottering
46. unstable
47. unsteady
48. top-heavy
49. tumbledown
50. ramshackle
51. crumbling
52. waterlogged
53. helpless
54. in a bad way
55. reduced to the last extremity
56. at the last extremity
57. trembling in the balance
58. nodding to its fall
59. destruction
60. threatening
61. ominous
62. illomened
63. alarming
64. fear
65. explosive
66. excited
67. wrought up
68. astir
69. sparkling
70. in a quiver
71. in a fever
72. in a ferment
73. in a blaze
74. in a state of excitement
75. in hysterics
76. black in the face
77. overwrought
78. taught
79. on a razors edge
80. hot
81. red-hot
82. flushed
83. feverish
84. all of a twitter
85. in a pucker
86. with quivering lips
87. with tears in ones eyes
88. uncomfortable
89. uneasy
90. ill at ease
91. in a taking
92. in a way
93. disturbed
94. discontented
95. out of humor
96. weary

97. afraid
98. fearful
99. timid
100. timorous
101. diffident
102. coy
103. faint-hearted
104. tremulous
105. afraid of ones shadow
106. apprehensive
107. shadow
108. restless
109. fidgety
110. more frightened than hurt
111. desirous
112. desiring
113. inclined
114. willing
115. partial to
116. fain
117. wishful
118. anxious
119. wistful
120. curious
121. at a loss for
122. sedulous
123. solicitous
124. craving
125. hungry
126. sharp-set
127. ravening
128. with an empty stomach
129. thirsty
130. athirst
131. parched with thirst
132. pinched with hunger
133. famished
134. dry
135. hungry as a hunter
136. hungry as a hawk
137. hungry as a horse
138. hungry as a church mouse
139. hungry as a bear
140. modest
141. humble
142. bashful
143. shy
144. skittish
145. sheepish
146. shamefaced
147. blushing
148. overmodest

Appendix E Abbreviations Used in This Article (Alphabetized)

ARDL: auto-regressive distributed lag models

BMZ: Johan Bollen, Huina Mao and Xiao-Jun Zeng

DCM: Derwent Capital Markets/Derwent Capital Management (referred to in news media under both names)

DJIA: Dow Jones Industrial Average

DTMP: Did Twitter "Calm"-ness Really Predict the DJIA?

GPOMS: Google Profile of Mood States

GBP: British pounds, the official currency of the UK

NLP: natural language processing

OF: Opinion Finder 2.0

POMS-Bi: Profile of Mood States (Bipolar Edition)

RWH: Random Walk Hypothesis

TARP: Troubled Asset Relief Program

TMP: Twitter Mood Predicts the Stock Market

URL: Uniform Resource Locator

XML: Extensible Markup Language