Appendix II **Mood Analysis Details**

> The question is no longer whether investor sentiment affects stock prices, but how to measure investor sentiment and quantify its effects. One approach is "bottom up," using biases in individual investor psychology, such as overconfidence, representativeness, and conservatism, to explain how individual investors underreact or overreact to past returns or fundamentals. The investor sentiment approach that we develop ... is, by contrast, distinctly "top down" and macroeconomic: we take the origin of investor sentiment as exogenous and focus on its empirical effects. We show that it is quite possible to measure investor sentiment and that waves of sentiment have clearly discernible, important, and regular effects on individual firms and on the stock market as a whole.
>
> *Investor Sentiment in the Stock Market*
>
> MALCOM BAKER AND JEFFERY WURGLER (2007)

In Appendix II, we review the theory and empirics of text analysis in finance and summarize key concepts from computational linguistics, covering the development of the theory of financial sentiments with a review of the DeLong et al. (1990) model and its equilibrium. Appendix II frames the sharp predictions of the theory of financial sentiments, discusses recent empirical findings related to the theory of investor sentiment, and the tenuous link between investor sentiments and textual sentiment. The ultimate purpose of this discussion is to derive Table 2.1 in the main text. Next, it discusses text analysis in finance, the n-gram bag-of-words model, mood analysis and the econometric evaluation of linkages between financial time series and linguistic sentiment time series. It concludes with a description of how our mood time series was constructed and selected visualizations.

1. Why Investor and Linguistic Sentiments Matter for Asset Pricing

The theory of financial sentiments arose in response to flawed arguments for market efficiency. Market efficiency, as defined by Fama (1976), means that asset prices reflect all available information so that the excess returns obtained from holding any bundle of assets will equal compensation for riskiness of the assets. It is often argued that in an efficient market, assets cannot deviate from their fundamental value, the value of the discounted cash flows from holding the asset, in a predictable way because doing so would enable arbitrageurs to profit by purchasing or selling the asset until it was priced at its fundamental value.[1] Friedman (1953) gave three arguments that (appear to) rely on progressively weaker assumptions against deviations of equities from their fundamental value:

1. If investors rationally value securities, there will be no deviation from the fundamental value of the security.
2. If investors are irrational and make symmetric errors in their security valuation, those errors will cancel for large numbers of investors and so there should be no large deviation from the fundamental value of the security.
3. Even if some investors are irrational and make asymmetric errors, a small number of arbitrageurs should be able to make money betting against those irrational investors. Those irrational investors will lose money, leave the market and there should be no large deviation from the fundamental value of the security in the long-run.

The first and second argument are valid but these arguments' premises of investor rationality or, at least, no systematic deviation from rationality, have been increasingly undermined by both experimental evidence of systematic deviations from individual rationality (Kahneman and Tversky, 1979; Smith and Smith, 2003) and the behavioral finance revolution (Barberis and Thaler, 2003). The third argument, which appears to rely on the weakest assumptions, however, is invalid. DeLong et al. (1990) present a model (which we call the DSSW model) in which there are two representative investors, both with constant absolute risk aversion with parameter γ utility and two assets, both paying the same dividend. The first asset has a perfectly elastic supply so it is risk-free and its price is always 1. The second asset however, has a fixed supply (normalized to 1). Even though the asset has no fundamental risk, we will call it the "risky asset" for reasons that will soon become clear. Since both assets pay the same dividend every period, they both share a fundamental value of 1. As usual, the risk-free asset pays r each period. The first investor is a rational arbitrageur who comprises fraction 1-μ of the market and has a correct expectation of the next period price.[2] The second investor, however, is a noise trader comprising fraction μ of the market who misperceives next period's price by a value $\boldsymbol{\rho_t}$ drawn from a normal distribution:

$$\boldsymbol{\rho_t \sim N\left(\rho^*, \sigma_\rho^2\right)} \qquad (1)$$

[1] One formulation of the no arbitrage condition in the single-period model is that all securities are priced as the expectation of a discounted sum of cash flows derived from the ownership of that security. We call this discounted expectation of the sum of cash flows the "fundamental value."

[2] Both investors believe that the price process is stationary.

In the DSSW model, $\boldsymbol{\rho^*}$ is the average level of misperception ("bullishness") and $\boldsymbol{\sigma_\rho^2}$ is the variance of the noise traders' misperceptions of the expected return per unit of the risky-asset. Given the aforementioned assumptions, our arbitrageur will not drive the price to its fundamental value, 1. Instead the equilibrium price in each period *t* will be:

$$\boldsymbol{p_t} = \underbrace{\mathbf{1}}_{\boldsymbol{fundamental\ value}} + \frac{\overbrace{\mu(\rho_t-\rho^*)}^{\boldsymbol{opinion\ shift}}}{1+r} + \frac{\mu\rho^*}{\underbrace{r}_{\text{"}\boldsymbol{price\ pressure}\text{"}}} - \frac{\overbrace{(2\gamma)\mu^2\sigma_\rho^2}^{\text{"}\boldsymbol{create\ space}\text{"}}}{r(1+r)^2} \qquad (2)$$

From (2) we can see that the generic[3] price of the risky asset in every period is not its fundamental value. The opinion shift effect in period *t* from noise traders having higher than average misperceptions varies from period to period and can push the price above or below its fundamental value. The fact that the noise traders have mean bias $\boldsymbol{\rho^*}$; will also act to push the current price of the asset above 1 for positive values and below 1 for negative values. Finally, the "create space" effect pushes the price down and results from the arbitrageur's purchasing less of the asset due to the increased return variability in the asset from noise traders. Friedman's argument might nonetheless hold if we could show that in equilibrium the arbitrageur outearns the noise trader. However, the equilibrium difference between the noise trader and the rational arbitrageur's expected returns is:

$$\boldsymbol{E[\Delta R_{n-i}] = \rho^*} - \frac{(1+r)^2(\rho^*)^2}{2\gamma\mu\sigma_\rho^2} - \frac{(1+r)^2}{2\gamma\mu} \qquad (3)$$

where we can see that the difference between the two returns is a concave down quadratic function in $\boldsymbol{\rho^*}$. This means that for values $\boldsymbol{\rho^*} > 0$ we can have parameters $(r, \gamma, \mu, \sigma_\rho^2)$ such that noise traders obtain higher returns in equilibrium. If the arbitrageurs copy the noise traders, they will have lower utility in equilibrium and so, given each trader's expectation, neither will imitate the other. If $\boldsymbol{\rho^*} < 0$, then our arbitrageur will hold more of the "risky asset" than the noise trader and outearn the noise trader in all scenarios. In the DSSW model, noise trader sentiments, or more precisely, their expectations about how the price of the financial instrument will deviate from its fundamental value, can affect the equilibrium price.

[3] By generic, we mean that the parameter space $(\mu, \rho_t, \rho^*, r, \sigma_\rho, \gamma)$ for which $p_t(\mu, \rho_t, \rho^*, r, \sigma_\rho, \gamma)$ = 1 is measure zero.

1.1 The Ambiguity of Theory

DSSW shows that investor sentiments, i.e. irrational beliefs about future cash flows, can affect the equilibrium price in both the short and the long-run. DSSW provides one of the most popular theoretical expositions of the importance of investor sentiments because it relies only on investor risk-aversion under a common utility assumption and noise trader dynamics, but today it is only one of many models in which investor sentiments can affect market outcomes. Shleifer and Vishny (1997) provide another canonical model in which the cost of betting against noise traders prevents arbitrageurs from pushing stocks to their fundamental values. Most noise trader models since Shleifer and Vishny (1997) share the key prediction that stocks which are difficult to arbitrage or value will experience the largest price variation due to sentiment. Furthermore, Baker and Wurgler (2007) note that "*in practice, the same securities that are difficult to value also tend to be difficult to arbitrage*." The fact that certain securities are more likely to meet the preconditions laid out by Shleifer and Vishny (1997) than others leads Baker and Wurgler to conjecture Figure II.1.1.

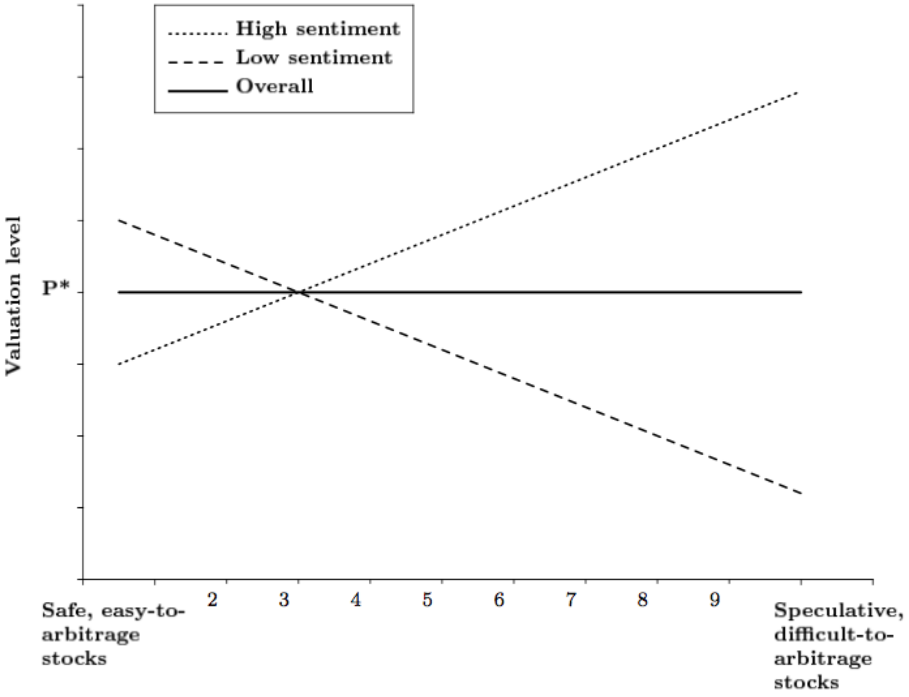

Figure II.1.1: Stocks that are speculative and difficult to value and arbitrage will have higher relative valuations when sentiment is high. This figure and caption corresponds with Figure 1 in Baker and Wurgler (2007).

P^* in the figure gives the fundamental value of the asset. On the right hand side of the cross point, it is easy to see that assets which are speculative (e.g. recent IPOs, firms experiencing extreme growth and small capitalization firms) should be more vulnerable to sentiment shocks than liquid, easy-to-arbitrage stocks (e.g. the firms comprising the DJIA). In practice, this means we should expect that higher levels of investor sentiment forecast more IPOs and and higher size premiums. As drawn, moving from the high sentiment curve to the low sentiment curve would lead to a "flight-to-quality" that boosts the relative value of less speculative, bond-like, firms.[4] Since these firms' market capitalization will increase when more speculative firms' market capitalization decrease, the net effect of sentiment changes on total capitalization is unclear. Thus, we can see that the sentiment theory does not necessarily lead to sharp predictions about the effects of investor sentiment on aggregate stock market. DSSW and in fact most theories of investor sentiment[5] correspond with the case in which the high sentiment curve is entirely above the fundamental value line and the low sentiment curve entirely below the fundamental value line; in these cases, high sentiment forecasts low aggregate returns and low sentiment forecasts high aggregate returns. While many of the econometric tests of investor sentiment proxies that we review in this section and perform in Section 4 are on the time series of differenced market prices, Figure II.1.1 should remind us that investor sentiment theory's most robust predictions are actually about the cross-section of asset returns.

1.2 The Empirics of Financial Sentiments

Under the EMH, we can decompose excess returns into compensation for risk and noise. In doing so, we are required to specify and estimate a model showing how investors in our sample were compensated for risk. Fama (1970) was the first to highlight the joint hypothesis problem. The joint hypothesis problem is that any test of the EMH is a test of the model specifying how investors were compensated as well as efficiency, individuals' ability to take advantage of profitable trading opportunities so that no such opportunities appear in equilibrium.[6] The EMH is empirically unassailable because any purported violation could be the result of model misspecification rather than a profitable trading opportunity. Theories of

[4] Baker and Wurgler (2006) suggests that the high "nifty fifty" stock valuations during the early 1970s, a time in which investor sentiment was low, are an example of flight-to-quality.

[5] Unfortunately, one key assumption driving the variability of DSSW's risky asset value was its inelasticity. Since the number of shares outstanding is elastic, the DSSW model does not necessarily help us understand the effects of sentiment on the level of the market. In general, strong assumptions are required for investor sentiment to have predictable effects on the time series of aggregate asset prices.

[6] A profitable trade is one that earns a higher excess return than our risk-compensation model would imply.

investor sentiment pose the same problem. Once a proxy for investor sentiment has been specified, any deviations from theory might just as well be the result of proxy misspecification rather than the incorrectness of the theory of investor sentiments. These empirical difficulties are further exacerbated by the fact that, even in theory, for many classes of equities, the presence of investor sentiments has ambiguous effects. These facts have led to the development of a number of ways of measuring investor sentiment in the hope that empirical work might be able to resolve theoretical ambiguity. In the next four subsubsections, we summarize four methods for measuring investor sentiments as well as the stylized facts derived from each approach.

1.2.1 Frankensentiments

The method pioneered by Baker and Wurgler (2006) measures sentiment by taking the first principal component of six proxies for investor sentiment, measured annually: the closed-end mutual fund discount, NYSE share turnover, the number and average first-day returns on IPOs, equity share in new issues and dividend premium. Their main findings are that their sentiment measure, when high (low), predicts low (high) subsequent returns for young, small, unprofitable, non-dividend-paying, high volatility, extreme growth and distressed stocks.

1.2.2 Exogenous Collective Mood Shocks

The second method for measuring investor sentiments is to analyze events that are correlated with collective mood changes but likely to be independent of corporate cash flows.[7] 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 returns. Further, if these events are associated with equity return increases or decreases but have no effect on corporate cash flows it is natural to suppose these events alter security prices through an investor sentiment channel. 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. Pettengill (2003) summarizes the "Monday effect"

[7] For instance, a natural disaster might alter collective mood by causing a panic, but it might also decrease future cash flows of corporations affected by the disaster. Since a rational investor would value the securities less to reflect the potential decrease in cash flows, we cannot easily disentangle the cash flow effects of the natural disaster from the sentiment effects of the natural disaster. Consequently, natural disasters are poor candidates for evaluating the effect of changes in investor sentiment.

literature. On average, stock prices decrease on Monday, and more significantly if they fell on the previous Friday. One 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. This literature suggests that stock prices respond immediately to events causing collective mood changes, and vanish rapidly, typically by the next day.[8]

1.2.3 Investor and Consumer Confidence Surveys

The third method for measuring mood involves evaluating consumer and investor confidence surveys. Investor confidence surveys are obvious candidates as proxies for investor sentiment. Although the case for consumer confidence surveys as investor sentiment proxies is less obvious, there is considerable overlap between the set of consumers and the set of investors (Fisher and Statman, 2003). Because survey data on consumers tends to be higher quality, there is more research on the relationship between consumer confidence and subsequent stock returns than investor confidence and subsequent stock returns. In fact, Wu et al. (2014) note that consumer confidence reports are sold to high-frequency traders seconds before they are released to the public so that traders can use the information before it reaches the market. Unfortunately, the association between consumer confidence surveys and equity returns is unclear at daily time scales. Nonetheless, we find that the survey evidence taken as a whole provides support for the cross-sectional theory of sentiments laid out by Baker and Wurgler (2007).

Fisher and Statman (2000) treat the survey of individual investors by *Investors Intelligence* as a proxy for investor sentiment and finds that the 1% increase in the individual investors' sentiment level is associated with a 0.1 percentage point decrease in S&P 500 returns over the following month. They find similar results for large investors' sentiment levels and subsequent S&P 500 returns. They find no relationship between changes in investor sentiment as measured by surveys and subsequent returns. They find no relationship between so-called medium sized investors (newsletter writers surveyed by *Investors*

[8] Kamstra et al. (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.

Intelligence) and subsequent equity returns over any time horizon.[9] Brown and Cliff (2004), using both *Investors Intelligence* and the American Association of Individual Investors Surveys find that investor sentiment has little predictive power at the monthly and weekly frequency across both equity indices and small-cap stocks. Consistent with Figure II.1.1, they find that higher levels of investor sentiment forecast increased IPO activity.

Christ and Bremmer (2003) find that the University of Michigan consumer sentiment survey does not predict the S&P500, DJIA or NASDAQ at the monthly frequency but rather that increases in any of the three equity indices forecast increases in consumer sentiment measures. Both Jansen and Nahuis (2003) and Fisher and Statman (2003) find that high consumer confidence has a statistically insignificant negative effect on short-term overall equity returns for the aggregate stock markets of selected European countries and the United States respectively; however, past high returns predict high consumer confidence over one to two weeks. Consistent with Figure II.1.1 however, both Lemmon and Portniaguina (2006) and Fisher and Statman (2003) find that certain parts of the consumer confidence index have predictive power for a subset of small-cap stocks, likely to be held by retail investors. Lemmon and Portniaguina (2006) find that on a quarterly time scale, higher consumer confidence forecasts lower small firm premiums. Fisher and Statman (2003) find that higher consumer confidence forecasts lower returns for small cap stocks over the next month. Fisher and Statman (2003) also find that high levels of consumer confidence predict statistically significant declines in the S&P500, and small-cap stocks over the six and twelve months. Lemmon and Portniaguina (2006) do not attempt to assess the economic significance of their findings and few if any of the consumer confidence driven declines in Fisher and Statman (2003) appear to be economically significant.[10] Consistent with Figure II.1.1, high consumer confidence appears to forecast lower relative and absolute returns for small cap stocks. In general, survey evidence suggests that changes in investor sentiments have no economically significant effect on the aggregate stock market or on large capitalization firms, like those comprising the DJIA.

[9] Even though the theory of investor sentiments, as conceived by Baker and Wurgler (2007), makes no sharp predictions about time series relationships between investor sentiments and equity returns, these findings are typically evaluated as if they are tests of the theory of investor sentiments (or, of course, as tests of surveys' ability to capture investor sentiment).

[10] In Exhibits 7, 8 and 9 of Fisher and Statman (2003), most of the coefficients regressing returns on confidence levels are small relative to the standard deviation of the consumer confidence index.

1.2.4 Linguistic Sentiments

> The connection between textual sentiment and investor sentiment is complex, and the extent to which they are causally related has not yet been thoroughly examined or understood. It is also unclear how investors interpret textual sentiment. The existing studies tend not to make assumptions about investor rationality, or about the relationship between textual sentiment and investor behavior. In this sense, they transcend the boundary between behavioral and traditional finance.
>
> *Textual sentiment in finance: A survey of methods and models*
>
> COLM KEARNEY AND SHA LIU (2014)

Our fourth method for measuring investor sentiments involves quantitative content analysis of text data.

Quantitative content analysis is defined by Riffe et al. (2005) as the:

> systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules and the analysis of relationships involving those values using statistical methods, to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption.

Kearney and Liu (2014) suggest that there is now a rapidly growing empirical asset pricing literature that relies on the analysis of soft, qualitative information as embedded in corporate filings (e.g. earnings press releases and annual reports), news media (e.g. news stories and analyst reports) and internet board postings.[11] Recent advances in NLP and the proliferation of large text corpora from the web have enabled widespread, systematic analysis of media content, in particular the tone, mood and sentiment of text media so that these can be used as covariates for financial analysis.[12] Automated content analysis, when applied to text, is called text mining, text analytics, or text data mining. Before discussing the results of text analytics for finance in depth, we believe that it is valuable to draw attention to a distinction, most forcefully illuminated by Sinha (2014), between investor sentiments as defined in financial economics and sentiments as defined in the text analytics, informatics, linguistics and computer science literatures. As noted above, investor sentiments refer to investors' irrational beliefs about future cash flows. Liu (2012) gives a partial definition of a sentiment, as the term is typically used outside of financial economics, 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 5 stars). Positive, negative and neutral are called sentiment (or opinion) orientations (or polarities).

[11] Despite the centrality of collective mood shifts to many behavioral theories of asset pricing and business cycles (Keynes, 1936; Shiller and Akerlof, 2010), even as of 2004 Capra noted that "mood is almost entirely unexplored in economics."

[12] A collection of text documents organized according to a fixed set of rules for the purpose of linguistic analysis is called a corpus.

We can see that, outside of financial economics, a sentiment is simply a representation, numerical or qualitative, of an author's opinion about something.[13] On the other hand, investor sentiments are *irrational expectations or beliefs* about future cash flows. Concretely, the statement:

$$\text{“I love AAPL! I think Apple will grow quickly for the indefinite future.”} \tag{4}$$

will always express positive sentiment about Apple Inc. in the computer science sense. But we cannot know whether or not the author of this sentence holds an *irrational expectation* about Apple Inc.'s cash flows without having other information about the state of the world. This statement might be perfectly justified on the basis of objective evaluation of Apple's fundamentals and management, in which case we would say that the author of this statement carries no financial sentiment about Apple Inc. We will call irrational beliefs about future cash flows *investor sentiments* or *financial sentiments* while referring to opinions about target objects as *linguistic sentiments, textual sentiments*, *polarities,* or just *sentiments*.

Sentiment analysis is the act of systematically extracting linguistic sentiments from a text about a target object. Linguistic sentiments are most commonly partitioned into positive, negative and neutral polarities (Liu, 2012) Given a document and a target in the document, there are two approaches to obtaining linguistic sentiments from that object (Kearney and Liu, 2014). The first approach is called the *dictionary-based approach*, sometimes called the *rules-based approach*, and we describe the most common variation on this approach extensively. Loughran and McDonald (2015) define a dictionary as a “tabulated collection of items, each with an associated attribute, as, for example, in its traditional form of a word and associated definition” and a lexicon as a dictionary “created for very specific purposes.” Within a dictionary, all words sharing the same attribute are assigned to the same “word list.” For example, given a dictionary that partitions every word into the polarities positive, neutral and negative, we have three word lists, one for each polarity. A dictionary-based approach on an Amazon.com review of a book might simply count the number of occurrences of words from the positive word list in the review and subtract from that the number of negative word list occurrences in the review to obtain a sentiment score for the book. The

[13] 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 and marketing.

second, less common method for sentiment analysis is the *machine learning approach* relying on statistical inference to classify documents. In this approach we split our documents into a training corpus and a testing corpus. We manually label each document in the training corpus with the polarities we are interested in and fit a statistical model, most typically naive Bayes (Mitra and Mitra, 2011), on the training corpus to learn classification rules that can be applied out-of-sample on the testing corpus.

Given that our measurements of linguistic sentiments are not necessarily equivalent to investor sentiments, how can we relate our text-derived time series to market behaviors? Early work on text analysis for finance ignored the theory of investor sentiments completely and focused on the information content of text. By definition, if markets are not semi-strong efficient, then fundamental analysis, which includes quantitative content analysis, could yield supernormal profits. Motivated by this possibility Antweiler and Frank (2004) used a machine learning approach applied to Yahoo! Finance message boards and discovered that neither the content of postings nor volume of messages carry significant information about the direction of equity returns, although the volume of messages predicts volatility. In a similar exercise on Yahoo! Finance message boards Das and Chen (2007) find that their machine learning approaches accurately measure linguistic sentiment, but that these constructed polarity measures have no relation with subsequent stock returns. Henry (2008) using a custom word list finds that, controlling for earnings, firms with more positive tone (tone is defined formally below) experience higher (but diminishing) abnormal returns and attributes this finding to investor irrationality.

Tetlock (2007), motivated by DeLong et al. (1990), was first to observe that linguistic sentiment time series should have systematically different effects on equity indices depending on whether those texts contain information, reflect previous investor sentiments or forecast investor sentiments. Concretely, 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 (analogous to negative polarity). The media pessimism factor can relate to the stock market returns (measured using the DJIA) in one of four ways. If Tetlock's measure of media pessimism forecasts investor sentiment, then high media pessimism should forecast low returns that are later offset by higher returns so that there is no net effect of pessimism on equity prices. If Tetlock's measure of media pessimism reflects previous investor sentiment, then high readings should follow low returns and predict high returns. Tetlock summarizes these two cases in a graph we reproduce in

Figure II.2.2. In the third case, in which media pessimism reflects new fundamental information, then increases in the pessimism index should forecast low short-run returns which are not offset by higher long-run returns. Finally, if the pessimism time series is noise or reflects stale information, there should be no systematic relationship between the pessimism time series and the DJIA.

Tetlock finds a significant negative effect on the DJIA from pessimism the previous day. The effect becomes insignificant on day two and, consistent with the hypothesis that his media pessimism factor forecasts investor sentiments, he finds evidence that low returns after high pessimism readings are later offset by higher returns so that the net effect of high pessimism, after one week, is statistically indistinguishable from zero. Tetlock's pessimism factor concords with the studies presented in above on collective mood shocks in that both find that the first order effects of collective mood shocks on the overall stock market (insofar as the DJIA proxies for this) should occur immediately following investor sentiment changes, and vanish quickly, usually within 24 hours. Consistent with Figure II.1.1, Tetlock finds large negative effects from increases in his pessimism factor on the small cap premium and finds no evidence of mean reversion in the small-cap premium over short time horizons.

2. Sentiment Analysis: Results and Fundamental Concepts

> Language is conceived in sin and science is its redemption.
>
> *The Roots of Reference: The Paul Carus Lectures*
>
> WILLARD VAN ORMAN QUINE (1974) AS QUOTED IN TETLOCK ET AL. (2008)

Since Tetlock (2007) illustrated how textual analytics can relate to equilibrium finance, work on textual analysis has (not so neatly) bifurcated into information-based approaches and attempts to capture investor sentiment. Both approaches use text-derived variables, like linguistic sentiment, as a proxy for information, investor sentiment or a mixture of both. Work based primarily on the information content of text is referred to as *news analytics* in part because it attempts to exploit information not completely reflected in asset prices, i.e. news. Mitra and Mitra (2011), in a survey of news analytics research, defines news analytics as the set of techniques “relating to the measurement of qualitative and quantitative attributes of textual news stories” so as to permit “the manipulation of everyday information in a mathematical and statistical way.” Most news analytics research focuses on company specific information. Early work in the news analytics tradition comes from Li (2006) who observes that the words “risk” and “uncertain” in firms' annual reports forecast low returns. Li (2006) shows that this information is economically significant by constructing arbitrage portfolios: annually buying stocks with small increases in the frequency of these words and shorting stocks with large increases in the frequency of these words, when matched on Fama-French and Fama-French-Carhart factors, generates an annual alpha of more than 10%.[14] Tetlock et al. (2008) analyzes 350,000 firm-specific news stories from the *Dow Jones Newswire* and the *Wall Street Journal*. They define a firm-specific news story as one that has at least 50 words, mentions the firm's official name within the first 25 words (including the headline), the firm's popular name at least twice, and has at least five total words designated “positive” or “negative” (with three of those five words being unique). They find that nearly all firm-specific information in articles is reflected in market prices within two days. Higher proportions of negative words forecast lower returns and a trader with no market impact or transaction costs using firm-specific information would earn 9.2 basis points over the Fama-French benchmark per day from 1980 to 1994; adjusting for realistic transaction costs (10 basis points per transaction), the news-based trading strategy earns no profits.

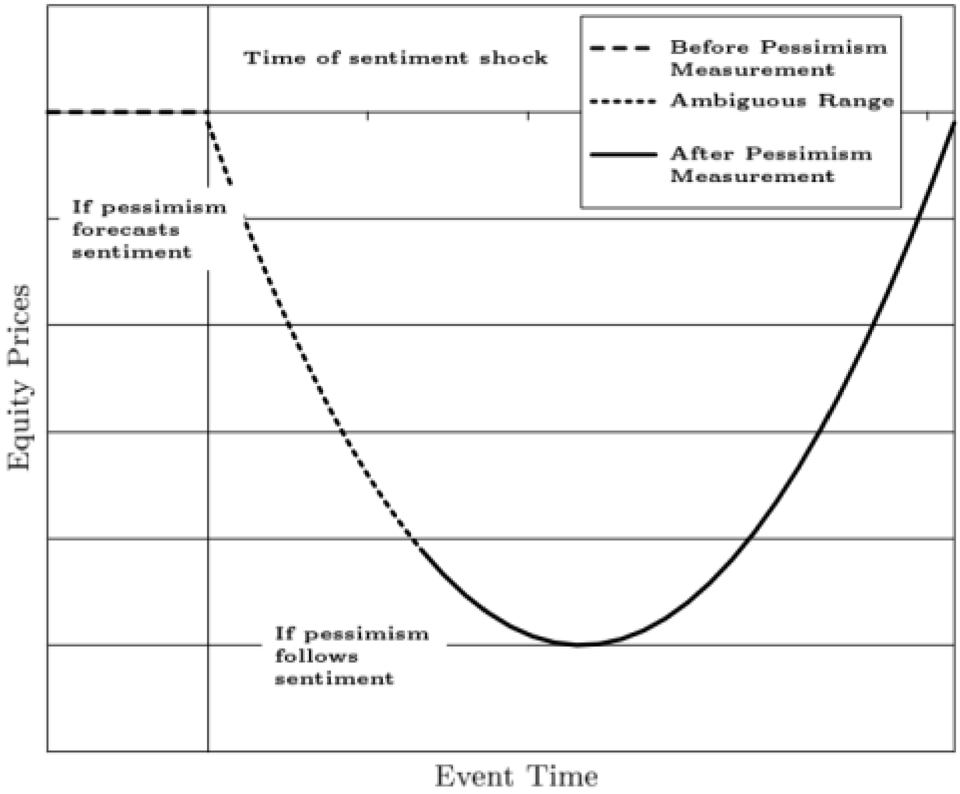

Figure II.2.2: The graph depicts the theoretical impact of a one-time increase in negative investor sentiment on equity prices. If the media pessimism measure is a predictor of investor sentiment, it will predict low short-horizon returns followed by high long-horizon returns of approximately equal magnitude. If the media pessimism measure follows past investor sentiment, it will predict low short-horizon returns followed by high long-horizon returns of greater magnitude than the short-horizon returns. This figure and caption corresponds with Figure 1 in Tetlock (2007).

[14] Assessment of the economic significance of results in news analytics studies appears to be more common than in work based on the theory of investor sentiment because news analytics focuses on the identification of arbitrage opportunities. Tests of investor sentiment often focus on identifying meaningful relationships between investor sentiment and financial and macroeconomic time series variables that, consistent with the limits to arbitrage literature, do not necessarily result in arbitrage opportunities.

In the investor sentiment tradition, Soo (2013) shows that the majority of geographic variation in house prices is explained by a sentiment index constructed from housing news in newspaper articles while Huang et al. (2014) show that abnormal positive tone in corporate filings predicts negative earnings surprises.[15]

2.1. The Bag-of-Words Model

> For many tasks, words and word combinations provide all the representational machinery we need to learn from text.
>
> *The Unreasonable Effectiveness of Data*
> ALON HALEVY, PETER NORVIG AND FERNANDO PEREIRA (2009)

Once we have established that linguistic sentiments, whether reflecting fundamental information or investor sentiments, matter for asset pricing, we are faced with the problem of what Loughran and McDonald (2015) call "qualitative analysis," or the mapping of text data onto interpretable numerical values that can be analyzed using statistical tools.[16] TMP presents itself as applying the n-gram bag-of-words approach (BOW hereafter) to extract a measure of collective mood states.[17] The n-gram and BOW are fundamental objects in computational linguistics. We review them below.

A BOW is a document rendered as a set of ordered pairs:

$$\{(X_1, Y_1)\ , (X_2, Y_2)\ , \cdots , (X_n, Y_n)\}$$

where X_i is a unique word, which we will call a unigram, in the document and Y_i is the frequency with which that word occurs. For example, the document consisting of the sentences "Jimmy has a black cat named Mouser. Jimmy loves Mouser." can be represented as the following BOW:

$$\{(Jimmy, 2), (has, 1), (a, 1), (black, 1), (cat, 1), (named, 1), (Mouse, 2), (loves, 1)\} \quad (5)$$

Notice that, because elements of a set have no order and all punctuation has been removed, this representation of the document destroys much of the syntactical information of the document. The BOW above has dimension eight, corresponding to $\#\{X_1, \cdots , X_n\}$. While in our toy example it may be possible for an individual to piece together the original sentences from the above representation, the space of possible sentences grows combinatorially as our documents grow in size and recovering the original documents from bag-of-words representations is considered an insoluble problem in computer science (Fillmore et al., 2009). We convert Tweets into BOWs. While BOWs may seem primitive, their tractability has made the

[15] The latter finding is particularly interesting because it suggests that corporate managers can exploit investor sentiments.

[16] Constructing microfounded structural models of verbal and written discourse is considered an open problem in linguistics, sociology and economics. All of the approaches to qualitative analysis we have found rely on econometric rather than microeconomic relationships.

[17] Bollen and Mao (2011b) suggests that the algorithm used in TMP is somehow more sophisticated than the n-gram bag-of-words approach. Every correspondent with whom we have discussed TMP reads the methodology presented therein as an n-gram bag-of-words approach.

BOW the standard data structure for work in textual analysis for accounting and finance (Loughran and McDonald, 2015; Kearney and Liu, 2014) while their simplicity has made them a standard for use in commercial NLP applications like spam filtering.

One way of augmenting the BOW model is to consider more complex features than the unigram. The most natural way of doing this involves a generalization of the unigram, the n-gram. An n-gram is a set of words found in a certain order. For example, the grammatically correct but semantically meaningless sentence “Colorless green ideas sleep furiously” is a 5-gram. The simplest n-gram extension of the unigram BOW model is the bigram BOW model (sometimes called bag of bigrams model) which considers all two word sequences. Our bigram BOW representation of (5) is:

```
{
   "Jimmy has": 1
   "has a" : 1
   "a black": 1
   "black cat": 1
   "cat named": 1
   "named Mouser": 1
   "Mouser Jimmy": 1
   "Jimmy loves": 1
   "loves Mouser" : 1
}
```

We can see that our bag of bigrams model has only marginally increased the dimension of our explanatory covariates to nine from eight in the unigram model. In exchange, we have an object that contains more word order information than the unigram BOW model. Intuitively, more information means that one would need less computation and information about the semantics of the English language to piece together the original document. This tradeoff holds in general. Higher order n-gram models contain more word order information, but also have higher dimension. The modal application of textual analysis in finance uses a variation on the unigram BOW model because of its transparency and ease of interpretation (Kearney and Liu, 2014; Loughran and McDonald, 2015). We use variations on the unigram and bigram BOW models.[18]

2.2 A Crash Course in Sentiment Analysis: Kernels and Aggregation

[18] Limitations on computing power as well the desire to limit model complexity motivate this choice. Before the mid-2000s, most commercial applications using BOWs used the trigram or bigram features. Miller and Leacock (2000) provide empirical evidence that suggests people are able to disambiguate different word senses with ±2 words of context; this finding suggests rapidly diminishing returns for n-gram models with n>4.

> Philosophy is a battle against the bewitchment of our intelligence by means of language.
>
> *Philosophical Investigations*
> LUDWIG WITTGENSTEIN (1953)

Once our document has been rendered as a BOW n-gram, we can extract linguistic sentiments from the document. The full definition of a linguistic sentiment is given by Liu (2012) as:

> a quadruple, (g, s, h, t) , where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion was expressed.

We must first define a target g, opinion holder h and time t for our sentiment analysis. In most cases, h and t will be obvious from context and we will not explicitly define them. The most basic BOW approach to sentiment analysis supposes that humans have a fixed number of possible textual sentiments about an object and that every word we use can be associated with one of those sentiments. As a purposely simplistic first example, we might suppose that humans have only three polarities: positive, negative and neutral. The partition of all sentiment states into these three polarities is sometimes called a naive partition in computational psychology but most academic and commercial applications rely on this assumption (Liu, 2012). We then define a dictionary that maps every n-gram in the English language to one of these three states. This gives us three n-gram lists[19], one list for each state. Then, given a set of documents $\mathcal{D}$ by a single author at a single point in time, we select every document by that author containing the target object of interest g; call these selected documents $\mathcal{D}_g$ g . Then, in our example we assign a 3-tuple (p, n, T) to each document where p is the number of positive words in the document, n is the number of negative words in the document and T is the total number of words in the document. Then, a predefined kernel $f\colon \mathbb{R}^3 \rightarrow \mathbb{R}$ assigns the aforementioned 3-tuple (p, n, T) to a single number. If our document size is fixed, Breen (2011) suggests a natural kernel choice might be:

$$M = f(p, n, T) := p - n \tag{6}$$

where p is a count of the words in document appearing on the "positive" n-gram list while n is a count of the words in the document appearing on "negative" n-gram list After applying this function to the document, we can average our kernel outputs across documents to obtain sentiment (g, s, h, t) and our analysis is complete.[20] The sentiment analysis technique just elucidated has reduced our bag-of-words text data from dimension # {unique words in text} to a single number that could, for an appropriate dictionary, specification of f, and document structure, be interpreted as a summary statistic telling us how the author feels about the target. For specification (6), the higher the value of M, the happier we might expect the author of the analyzed text to be.

[19] An n-gram list is the natural generalization of the word list concept.

[20] We will always take the mean and this is by far the most common approach in academic and commercial applications.

Consider two sets of documents (7) and (8) below, each consisting of a single Tweet at time t by different authors h_i and h_{i-1} respectively.[21] Let g = "Obama" and let us represent each document as a unigram BOW:

$$\text{“I feel bad today about Obama.”} \tag{7}$$

$$\text{“I feel good about Obama.”} \tag{8}$$

We can see that (7) appears to express a negative sentiment about g and (8) appears to express a positive sentiment about g. A reasonable dictionary would assign the word “good” to the positive polarity. The word “bad” would be assigned to the negative polarity and the rest of the words, which do not seem to be associated with either polarity a priori, would be assigned to a neutral polarity. Scoring each document gives us 3-tuple (0,1,6) for the document in (7) and 3-tuple (1,0,5) for the document in (8). Kernel (6) gives sentence (7) $a\ -1$ and sentence (8) a 1, validating our subjective expectations about how the authors of each set of documents feel about Obama and giving us sentiment 4-tuples (Obama , -1, h_i, t) and (Obama ,1, h_{i-1}, t).

Notice that for (6), $f(x, y)$ is decreasing in one of its arguments and increasing in the other. We call any sentiment analysis in which the document scoring kernel f has this property bipolar. For certain tasks, the monopolar kernels:

$$f\ (p, n, T) := p \tag{9}$$

or:

$$f(p, n, T) := n \tag{10}$$

will capture all information we need for sentiment analysis. Notice that in our toy examples in (7) and (8), simple kernels (9) and (10) will give the same rank ordering for the polarity of each author towards Obama as the more sophisticated (6). Other times, we may use non-bipolar kernels to capture specific features of the documents. A widely used measure of the subjectivity of a sentence is given by applying the following kernel to our 3-tuple:

$$f\ (p, n, T) := \frac{(p+n)}{T} \tag{11}$$

For this subjectivity measure, higher kernel values imply higher levels of subjectivity.

Kearney and Liu (2014) note that “standardization is necessary if the raw frequency of matched words in the total number of words is not stationary, which can happen when regime changes occur over time in the distribution of words in the text. This can happen, for example when the writing style changes with the author.” Because we are working with tweets of a fixed size, 140 characters, the error from assuming stationarity in the distribution of text content across Tweets is bounded (although it may be quite large) and so, with enough data, even simple kernels can be effective (Breen, 2011). One elementary way to

[21] We will treat single Tweets as documents throughout the main text. Treating each day’s Tweets as a single document did not result in outperformance in our main text in-sample significance tests.

partially adjust for different writing styles and completely address non-stationarity in document size is to control for word count. This suggests the following kernel:

$$f\ (p, n, T) := \frac{(p-n)}{T} \tag{12}$$

In empirical asset pricing and corporate finance applications, Loughran and McDonald (2015) point out that corporate managers use sophisticated syntax not detectable by our n-gram BOW models to disguise bad news using words that most researchers assign to the positive word list. To mitigate this they recommend ignoring positive word counts completely and using the following kernel:

$$f\ (p, n, T) := \frac{n}{T} \tag{13}$$

Tetlock et al. (2008), notes that empirically the kernel (13) carries approximately as much firm-specific information as (12).

Finally, we are faced with the problem of converting the distribution of sentiments across individuals into a single number. The correct aggregation depends on the application. Tetlock et al. (2008) suggests gathering all documents containing the target and treating them as one “composite” document, then simply using one of the kernels specified above. A more typical approach is to take a measure of central tendency (usually the mean) across individuals. A third approach suggested by Breen (2011) and now common for Twitter sentiment analysis applications, is to first note that sentiment scores from Tweets containing only one or two terms in our word lists are likely to be noise. Let x_j be the kernel output for Tweet j. We set a minimum threshold for Tweet sentiment using the following kernel:

$$f\ (p, n, T) := \ f(x) = \begin{cases} 1, & if\ p - n > 2 \\ -1, & if\ p - n < -2 \\ 0, & else \end{cases} \tag{14}$$

and aggregate across individual tweets $x_i,\ i\ \in\ \{1, \ldots\ , I\}$, with I the total number of Tweets, by taking the ratio of individuals expressing intense positive emotion about a topic to individuals expressing intense negative emotion:

$$S := \frac{\sum_{i=1}^{I} 1_{x_i=1}(x_i)}{\sum_{i=1}^{I} 1_{x_i=-1}\,(x_i)} \tag{15}$$

2.3 Mood Analysis: Fundamental Concepts and Results

> The limits of my language are the limits of my world.
>
> ---
>
> *Tractatus Logico-Philosophicus*

In computational psychology, "mood analysis" algorithms are one method of both reducing the dimensionality of text while increasing the interpretability of text covariates under analysis.[22] First, let us distinguish between sentiment and mood. A mood or emotion is simply a subjective feeling and thought.[23] Note that sentiment must have a target g, but no such restriction exists for defining a mood. In practice, mood analysis uses the exact same n-grams BOW model outlined above, but does not filter documents based on whether or not they contain a target object. Instead, mood analysis filters documents by author and returns a statistic that measures the intensity of the author's mood.

Mood analysis begins by supposing that humans have a fixed number of mood states and defines a dictionary by assuming that every n-gram we use can be associated with one of those mood states (e.g. happy or sad) or a neutral category. Unlike in sentiment analysis, where in practice naive partitions are the default, mood analysis often relies on more sophisticated models of human emotion. Taxonomizing mood states is an open problem in psychology (Liu, 2012).[24] Lorr and Shea (1979) suggests that many human mood states are bipolar in the sense that if someone feels more of one mood state then they necessarily feel less of the other. When mood states have an inverse relationship, they are said to be bipolar to each other and on the same mood dimension; we will write two moods m_1 and m_2 in a bipolar relationship to each other as $m_1 \intercal m_2$ or as $m_1 - m_2$.[25] For example, Lorr and Shea (1979) provide experimental evidence that humans have at least three bipolar mood dimensions labeled composed-anxious, energetic-tired and agreeable-hostile. Concretely, according to the bipolar theory of mood as presented in Lorr and Shea (1979) individuals who feel composed (anxious) necessarily feel less anxious (composed) while individuals who feel more tired (energetic) necessarily feel less energetic (tired) and so on. Lorr et al. (1982) suggest that humans have five bipolar mood dimensions that they term composed-anxious, agreeable-hostile, energetic-fatigued, elated-depressed and clear-thinking-confused. Given a bipolar model of mood specifying m bipolar mood dimensions $\mathcal{M} = \{1,2, \dots, m\}$ we will have 2m n-gram lists, i.e. one for each mood state. For example, Lorr, Maurice, and McNair (1984), based on the conjecture that humans have six bipolar mood dimensions (with 12 mood states: composed-anxious, clearheaded-confused, confident-unsure, energetic-tired, agreeable-hostile, and elated-depressed), present a psychometric test, the Profile of Mood States-Bipolar Edition (POMS-Bi hereafter) and a list of 72 terms, 6 for each mood state. The terms are comprised of unigrams (64), bigrams (5) and trigrams (3). The POMS-Bi exam gives participants the following instructions (all capitalization and emphasis from original):

[22] Similar concepts and terminology exist in computational linguistics, computational marketing and machine learning, but a full exploration of the nuances of each domain's approach to this topic would take us far afield.

[23] Kim et al. (2015) notes that "emotion, affect, and mood have distinguishable meanings in psychology" but that the literature on sentiment and mood analysis treats these as the same concept.

[24] Most work in computational psychology relies on the model in Parrott (2000); Liu (2012) in which humans have six emotions love, joy, surprise, anger, sadness and fear. In the full model, each of these moods has secondary and tertiary classifications, as well as variable levels of intensity.

[25] We use the former notation because the mood questionnaire of primary interest in this essay, the POMS-Bi, has a single mood state "clear-thinking" with a hyphen.

> Below are words that describe feelings and moods people have. Please read EVERY word carefully. Then fill in one space under the answer which best describes how you have been feeling DURING THE PAST WEEK INCLUDING TODAY. Suppose the word is *happy*.
>
> Mark the one answer which is closest to how you have been feeling DURING THE PAST WEEK INCLUDING TODAY.

Next to each of the 72 terms are four options, labeled 0,1,2 or 3 corresponding to "MUCH UNLIKE THIS", "SLIGHTLY UNLIKE THIS", "SLIGHTLY LIKE THIS" and "MUCH LIKE THIS" respectively. To return a 6-tuple mood score for someone who has taken the exam, for each mood state i we first sum the response across each term corresponding to that mood state as in (16):

$$s_i = \sum_{j=1}^{6} t_{ij} \qquad (16)$$

where t is a term and $t_{ij} \in \{0,1,2,3\}$ corresponds with each of the participant's four options above.[26] We compute 12 such sums corresponding with $s_i : i \in \{-6, -5, \ldots, -2, -1, 1, 2, \ldots, 5, 6\}$. Let $(s_1, s_{-1}, s_2, s_{-2}, \cdots, s_6, s_{-6},)$ correspond these sums such that mood state i is bipolar to mood state $-i$. The POMS-Bi is scored using the kernel specified in (6) so that each of the six mood dimensions has value:

$$m_i \intercal m_{-i} = s_i - s_{-i} \qquad (17)$$

Given the framework above, it is easy to adapt the POMS-Bi and its terms to naturally occurring text data for mood analysis. For an author who has written document d rendered as an n-gram BOW, we define a 13-tuple $\{w_1, w_{-1}, w_2, w_{-2}, \cdots, w_6, w_{-6}, T\}$ where T is the total number of words in d and w_i corresponds to the number of times n-grams from mood state i appear in the document. Because we believe that moods are bipolar, our kernel should be bipolar. However, in mood analysis applications, we usually want to compare the relative intensity of mood dimensions and so we need to account for how often mood n-grams from different dimensions are used in text. One way of doing this is to use what is sometimes called the tone kernel (Henry, 2008; Tetlock et al., 2008; Liu, 2012; Han, 2012). For each mood dimension $m_i - m_{-i}$ we would like to estimate, define our tone kernel $f\colon \mathbb{R}^2 \rightarrow \mathbb{R}$ as:

$$m_i \intercal m_{-i} = f(w_i - w_{-i}) = \begin{cases} \frac{w_i - w_{-i}}{w_i + w_{-i}}, & if\ w_i - w_{-i} > 0 \\ 0, & otherwise \end{cases} \qquad (19)$$

The tone kernel is bounded in $[-1,1]$ and allows us to compare mood dimensions in terms of the proportion of terms from a dimension drawn from one mood state. Kearney and Liu (2014) points out that using the tone kernel in sentiment analysis tasks naturally adjusts for different document sizes. We can use any of the three techniques outlined above for aggregation.

[26] Our index j in (16) runs from 1 to 6 because six terms in our dictionary are assigned to each mood state i so that $\sum_j \sum_i 1 = 72$.

As a first example of mood analysis, we might suppose that humans have only two moods in bipolar relation to each other: happy and sad. The following will be our dictionary[27], defined over bigrams:

{ I feel: neutral , feel good: happy , feel bad: sad , bad today: neutral } (19)

Now consider the statement:

"I feel good." (20)

This has BOW bigram representation:

{ I feel: 1 , feel good: 1 } (21)

One way of defining (6) is to replace happy, sad and neutral in (19) with the numerical values $(1, -1, 0)$ and rewrite our bigram BOW to include all the bigrams in our dictionary that occur 0 times in our statement, ordering the bigrams so that they appear in the same order as in (19) . We call this the vector representation of our BOW and we call our ordered dictionary the vector representation of the dictionary. We present both below:

B_{20} = (I feel: 1 , feel good: 1 , feel bad: 0 , bad today: 0)

D = (I feel: 0 , feel good: 1 , feel bad: -1 , bad today: 0)

Now, (6) is equivalent to the dot product[28] of our vector bigram BOW representation and our vector dictionary representation:

$$B_{20} \cdot D = 1$$

and this formula is the typical presentation of (6) in computational linguistics. Now consider the statement (22):

"I feel bad today." (22)

Repeating the steps above we calculate:

$$B_{22} \cdot D = -1$$

Intuitively, we expect that the author of expression (22) is sadder than the author of expression (20). Our mood analysis validates our intuition: scoring the author of (22) lower on the happy-sad

[27] The assignment of "bad today" to the neutral category is unintuitive, but necessary to avoid double counting. As one might expect, defining bigram dictionaries is generally trickier than defining unigram dictionaries.

[28] Just as one can generalize our dot product to the space of inner products, we can generalize (6) to a larger kernel space.

dimension than the author of (20). We now have all the tools we need to begin conducting mood analysis on text corpora.

2.4 Avoiding Confusion: Bipolar Kernels Versus Bipolar Mood States

> The same meaning can be expressed in many different ways, and the same expression can express many different meanings.
>
> *The Unreasonable Effectiveness of Data*
> ALON HALEVY, PETER NORVIG AND FERNANDO PEREIRA (2009)

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 in refers to the functional specification applied to the counts of the total number of n-grams 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 (i.e. using monopolar kernel) 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. Indeed, if one of our mood states is measured with more noise than the other, only using the less noisy mood state to measure a mood dimension can actually be advantageous.[29] Unfortunately, Lorr et al. (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 lead to incorrect estimates of individual mood states 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 the Lorr and Shea (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 maximum possible response) 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%. It can be shown that even for less pathological cases of extreme response bias, questionnaires featuring both half-scales will reduce the total bias. For individuals without extreme response bias, the bipolar structure of the questionnaire should give results equal to the monopolar structure, ceteris paribus.[30] Mood analysis algorithms, like those proposed by Bollen et al. (2011a), borrow liberally from the psychometric literature and we attempt to follow the established precedents used by the psychometric literature even when it

[29] Loughran and McDonald (2015) claim exactly this in the context of sentiment analysis (with positive and negative sentiment towards a target typically being assumed to be bipolar in the sense used by Lorr and Shea (1979)) for empirical asset pricing and corporate finance.

[30] This analysis relies on extreme response bias being *the only* bias among those surveyed. Notice that going from a half-scale questionnaire to a full-scale questionnaire doubles the length of the questionnaire. For individuals who demonstrate extreme response bias only when their attention is depleted (e.g. someone who marks off the maximum answers after they become bored with the test), a bipolar survey design could worsen extreme response bias.

appears that the original justification (extreme response bias on surveys) for those precedents (measuring word counts from both half-scales and using a bipolar specifications of f) do not apply.[31]

3. Constructing Explanatory Variables

> Because of a huge shared cognitive and cultural context, linguistic expression can be highly ambiguous and still often be understood correctly.
>
> *The Unreasonable Effectiveness of Data*
>
> ALON HALEVY, PETER NORVIG AND FERNANDO PEREIRA (2009)

In this section we discuss the variable selection process used in constructing our dictionaries and mood time series. Our main objective in this section is to produce a collective mood time series for bivariate analysis. Each time series is specified by a lexicon-kernel pair (ℓ, K) where ℓ is the lexicon and the kernel is K. We follow Bollen et al. (2011a) and simply take the mean across each day's individual Tweet scores to obtain the collective mood score for that day.[32]

3.1 Why Select At All?

> Ever think that your tweets might predict the future? An Indiana University professor has been putting your tweets into a powerful computing system to determine the general public mood and by following how we feel, using how you feel to predict the market and with shocking success! Tests from his first paper showed an 86.7% success rate against short-term ups and downs on the Dow.
>
> *Can Tweets Predict Stock Market's Future*
>
> FOX BUSINESS NEWS, 2013/03/01

We can more clearly explain why we conduct variable selection in this section by describing our objective. We aim to use the mood time series constructed in this section to find a Twitter mood effect in-sample. Since, we cannot construct exactly the same collective mood time series as BMZ, our tests in the main text are assess the *robustness* of the effect TMP purport to discover. This check helps us discriminate between explanations for the failure of the Twitter mood effect to materialize when BMZ attempted to exploit it at DCM.

[31] We suspect that this observation is original but we have not completely surveyed the mood analysis literature.

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

The first explanation is that BMZ discovered a real Twitter mood effect in the data which was later arbitraged out. In this case, we should in principle be able to find a Twitter mood effect in the data using standard methodologies, these effects should be robust to extending our sample into the past, and Twitter mood should forecast the stock market ex-sample, provided we do so before TMP's initial appearance in 2009.

The second explanation, for which Lachanski (2016) provides evidence, is that the Twitter mood effect present in TMP is simply a false positive resulting from multiple comparison bias. In this case, we expect that our linear hypothesis tests over the time period covered in TMP should be able to recover a Twitter mood effect even if that effect is simply the result of the data conspiring in BMZ's favor for that particular time period. However, we should not be able to recover significant effects using an extended sample. Twitter mood should also provide additional predictive power in the ex-sample used by TMP, but not in an extended ex-sample.

A third explanation is that the p-values reported in Section 2.4 of TMP are the result of the GPOMS overfitting in-sample, with the ex-sample performance being the result of luck. The latter is plausible because the ex-sample was not chosen randomly and consists of only 15 days. In this case, our mood analysis tools are unlikely to be able to recover a Twitter mood effect even in-sample.

We find low p-values in-sample (although they do not match the results presented TMP) and higher p-values in the extended sample, suggesting the second explanation as the dominant factor explaining both the in-sample performance of TMP and subsequent failure to replicate. We find little evidence that our Twitter mood constructs aid in predicting the stock market, but it is possible to choose hyperparameters for the non-linear algorithm to match the conclusions presented by TMP.

If we conduct many hypothesis tests then we must somehow restrict either the FDR or FWER to correctly estimate our p-values.[33] If many of (ℓ, K) parameters give *unreasonable* mood time series, then including them in our main text analysis will rig the game against TMP, biasing our p-values upward. This problem is exacerbated by the fact that sentiment analysis methodologies usually yield highly correlated

[33] The strategy pursued in Lachanski (2014a) is precisely this: estimating many time series out-of-sample (in November and December, 2008) and using the strongest possible FDR correction to find a Twitter mood effect. No Twitter mood effect was found. We opted for that strategy because only the November 4th, 2008 local minima was identifiable from TMP's graphics and all of our mood time series in that paper had local minima on that date. Because we have a longer time series with more large local minima, we can use more precise selection techniques on our collective mood time series.

time series. Using standard multiple comparison adjustments in this situation inflate our Type II errors. Another concern we have arises from the fact that none of the multiple hypothesis testing procedures used in Lachanski (2014) or Lachanski (2015) have been proven to have maximum power in their class of tests (e.g. FWER restricting, FDR restricting, FDR restricting if time series are independent, etc.). If we test many time series, we open the door to the possibility that a future multiple hypothesis test may be invented which finds a significant result among our time series. Therefore, by appropriately limiting the number of tests we conduct before investigating the DJIA data, we can give the result in TMP "the benefit of the doubt" while still assessing the validity of the Twitter mood effect in-sample. At the same time, we will see below that different data processing techniques give rise to different mood time series. Even though we use a hold out set for model selection, if we test every time series generated by every combination of data cleaning, parsing and standardization, we are likely to encounter mood time series with good ex-sample performance by chance, i.e. data dredging. Among a set of time series parameterized by (ℓ, K), we might imagine three criteria for pre-test selection:

1. Does the time series generated by scoring each Tweet using (ℓ, K) appear visually similar to BMZ's locally-normalized CALM time series?
2. Does (ℓ, K) accurately score individual Tweets?
3. Does the time series generated by (ℓ, K) give accurate collective mood scores?

In the following subsubsections we discuss these criterion. We find that the first criterion, properly operationalized, is stringent enough to select a single time series from the 18 we generate in this section.

3.2 Visual Comparison

BMZ display only two graphics containing their locally normalized CALM time series and, as discussed in the main text, do not present their normalization parameter k.[34] Therefore, even if our underlying collective mood time series was exactly the same as theirs, after normalization only the most salient features of our series would be identifiable with the CALM series they present. BMZ are also careful to show their locally normalized CALM time series over only a small fraction of the data which they analyze. This limits our ability to match features of the time series we construct in this section with their CALM time series. We

[34] The possibility of k taking on non-integer values dissuaded us from attempting to visually compare BMZ's time series with every time series we can derive from different combinations of lexicon-kernel-k triples.

present the two graphics in TMP that visually display the locally normalized CALM time series in Figures II.3.1 and II.3.2. Figure II.3.3 contains graphical information we need to more precisely relate the two curves.

BMZ's graphics are unintuitive and through close inspection of Figures II.3.1, II.3.2, and II.3.3 in this section, we will see that the most straightforward interpretation of TMP's Figure 3 is inaccurate. BMZ claim that the locally normalized CALM values in Figure 3 of their paper have simply been shifted back three days so that, for example, the height of the CALM curve over the horizontal axis point labeled "Aug 09" would correspond with the locally normalized CALM time series value on August 6, 2008. The vertical lines in Figure II.3.3 collectively make it clear that the bottom panel's CALM graphic has been superimposed over the DJIA in the top panel of the figure. The orange line corresponds to the point labeled "Oct 08" on the horizontal axis and the pink line corresponds to a point on the DJIA curve that BMZ label "bank bail-out." In Section 2.4 of TMP, BMZ allude to a Federal Reserve bank relief action on October 13, 2008 which they call "a major bank bailout." If we are willing to identify the "bank bail-out" labeled in the curve with the "major bank bailout" referred to in the text of TMP, we could conclude that the pink vertical line corresponds with the date October 13, 2008 (Bernanke, 2008). There are four trading days in the period ranging from October 8, 2008 to October 13, 2008 and four kinks shown in the DJIA and CALM time series plots. Thus, the pink and orange lines confirm that the date labels on the horizontal axis labels correspond with each date's DJIA value. The point on the horizontal axis of Figure II.3.3 labeled "Aug 09" should not have a kink in the DJIA plot point aligning with this label because August 9, 2008 occurs on a Saturday in 2008. If we simply assume, as BMZ claim, that that point on the CALM time series plot corresponds with August 6, 2008, then there should be a kink on the CALM plot that corresponds to this day. In Figure II.3.3, using the yellow vertical line we see that, as predicted, there is no kink in DJIA curve on this day but there is also no kink in the CALM time series. This suggests that either BMZ are removing weekday CALM values that, when lagged three days, would correspond with DJIA weekends (i.e. all Wednesday and Thursday CALM values) and holidays or that the difference in slopes between the two days on either side of August 6, 2008 is so small that no kink is visible. For the remainder of this essay we assume the former.[35] Together these three suppositions allow us to align the labeled date horizontal axis in TMP's Figure 3 with the CALM time series in the bottom panel. In this case, for example, we could infer that the CALM point labeled "Oct 28" on the bottom panel (with brown vertical line running through it) of Figure 3.3 actually corresponds with the locally normalized CALM value from October 25, 2008. We can see that locally normalized CALM is increasing in the two days before October 25, 2008, i.e. locally normalized CALM increases from October 24 to October 25 as well as from October 21 to October 24 (since October 25 and October 26, 2008 fall on a weekend).

To confirm our supposition, we can use the fact that the points just identified also appear in Figure 2 of TMP. Using Figure 2, we can see that, indeed, the CALM time series rises from October 24, 2008 to October 25, 2008 and that, consistent with our supposition, the value of the locally normalized CALM time series is greater on October 24, 2008 than on October 21, 2008, so that if these points were plotted the slope would appear to increase as it does in Figure 3. If BMZ did not drop locally normalized CALM values on Wednesdays and Thursdays from the plot in TMP's Figure 3 then the plots would be contradictory. For instance, Figure 2 shows a decrease in the locally normalized CALM time series from October 23, 2008 to October 24, 2008 but Figure 3 would show an *increase* in the aforementioned time series that we naively assume span these two days.[36] We conclude that the plot shown in Figure 3 of TMP contains locally CALM time series points ranging from July 28, 2008 to October 27, 2008.

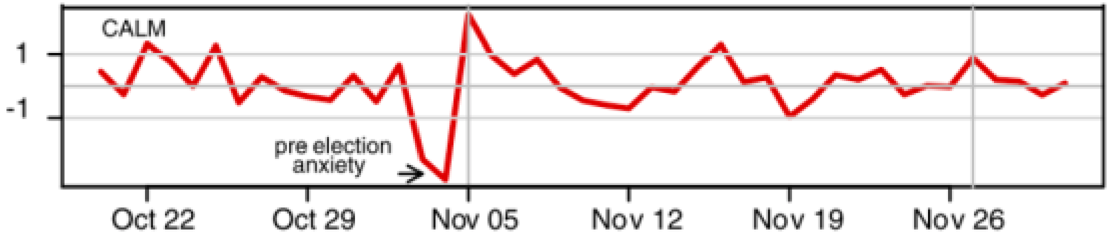

Figure II.3.1: This figure corresponds to Figure 2 in TMP. It appears to show the locally normalized CALM time series from October 20, 2008 to November 30, 2008. We infer that it includes weekends because there are 7 kinks in the curve shown in each period that would correspond to one week in calendar time, e.g. October 22, 2008 to the day before October 29, 2008.

[35] Assuming the former does not change any of the conclusions above while assuming the latter would throw many of our conclusions, which rely on a count of the number of kinks in the time series curves drawn between labeled horizontal axis values, into doubt.

[36] Another possibility for this discrepancy could be that BMZ choose different local normalization parameter k values for Figure 2 and Figure 3 in TMP. We considered this possibility however it would still not explain the lack of a kink in the CALM time series point aligned with the "Aug 09" label in Figure 3; therefore, we opt for the more parsimonious interpretation of these Figures presented in this paragraph.

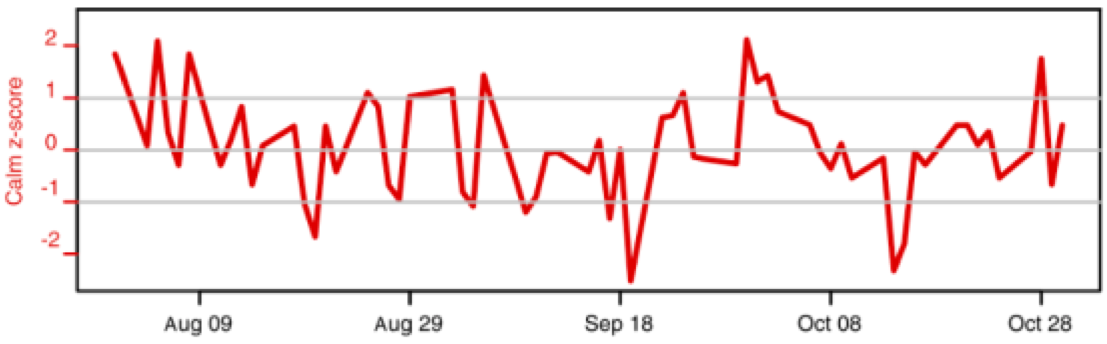

Figure II.3.2: This figure corresponds to the bottom panel presented in Figure 3 of TMP. It shows the 3-day lag of BMZ's CALM time series. It appears to show the locally normalized CALM time series from August 1, 2008 to October 30, 2008. In actuality, we have concluded that it contains the locally normalized CALM time series, less points that when lagged would correspond with weekends, from July 28, 2008 to October 27, 2008. We infer that CALM points which, lagged three days, would correspond with weekends and holidays have been removed because there are 15 kinks shown between the dates labeled August 9, 2008 and August 29, 2008, but only 13 kinks shown from the day after the date labeled August 29, 2008 to September 18, 2008 (corresponding with three weeks less one week day as well as Labor Day, a Monday, on September 1, 2008) and the point labeled "Aug 09", which is a Saturday in 2008, does not correspond with a kink in either the DJIA or CALM time series.

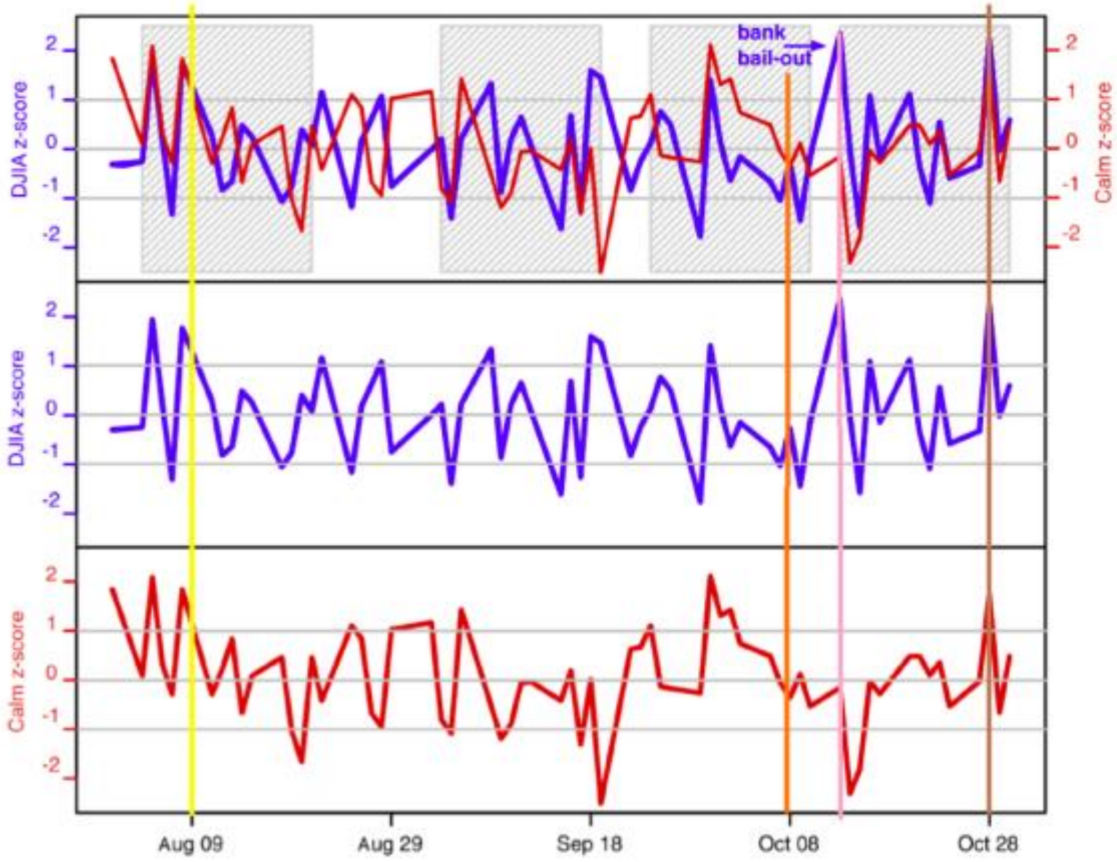

Figure II.3.3: This figure corresponds with Figure 3 in TMP. We have added the yellow, orange, pink and brown vertical lines to the original Figure.

Using the dating scheme just discussed, we identify four features from locally normalized mood in Figure II.3.2 and one feature in Figure II.3.1 that we would like our specification of (ℓ, K) to capture; we document our visual variable selection criterion in Table II.3.1. From Figure II.3.2, we choose dates corresponding to the two highest and two lowest locally normalized CALM readings and reject all (ℓ, K) that do not give local maxima and local minima, respectively, readings on these days.[37] From Figure 3.1, we observe that BMZ have identified the pre-election anxiety as significant. We reject any (ℓ, K) that do not identify November 4, 2008 as a local minima for composed-anxious readings. We also reject any (ℓ, K) that do not show the largest "local" increase in composed-anxious readings from November 4 to November 5, 2008. By limiting our selection to these six criterion, we parsimoniously eliminate all time series construction methodologies that are unable to detect large changes in BMZ's collective mood index while minimizing the subjectivity of our analysis.

	Date	Our "Calm"-ness Measure Must Have a	Criterion Justified by:
Criterion 1	11/04/2008	Local minima	Figure II.3.1
Criterion 2	11/05/2008	Large increase over previous day	Figure II.3.1
Criterion 3	09/16/2008	Local minima	Figure II.3.2
Criterion 4	10/11/2008	Local minima	Figure II.3.2
Criterion 5	08/01/2008	Local maxima	Figure II.3.2
Criterion 6	09/27/2008	Local maxima	Figure II.3.2

Table II.3.1: This table contains our visual selection criterion.

3.2.1 Individual and Collective Accuracy

BMZ validate their OF and GPOMS tools in two ways. First they graph each day's mood value and, if they are able to come up with an explanation for why the OF or GPOMS tool gives a particularly high or low value on a particular day, then the tool is declared a success. BMZ write:[38]

> Fig. 2 shows that the OF successfully identifies the public's emotional response to the Presidential election on November 4th and Thanksgiving on November 27th. In both

[37] Unfortunately, local minima and maxima are not invariant under the local normalization. For k = 2 and 3 the cases we have constructed under which local minima and maxima are not preserved appear to be pathological. By choosing the largest deviations from the local mean, we maximize the probability that local extreme values on the Figures BMZ present correspond to local extreme values in our time series.

[38] While the explanations BMZ give for the changes in the GPOMS dimensions over time make sense, one wonders: if BMZ had found a high Alert score on Thanksgiving, would they have attributed it to the GPOMS tool correctly detecting the stress of cooking and interacting with rarely-seen relatives? In general, it is difficult to see how this storytelling exercise could reject the GPOMS as a useful tool. For any large deviation, BMZ could attribute it to a news event from that day.

> cases OF marks a significant, but short-lived uptick in positive sentiment specific to these days.
> The GPOMS reveal November 3, 2008 is characterized by a significant drop in Calm indicating highly elevated levels of public anxiety. Election day itself is characterized by a reversal of Calm scores indicating a significant reduction in public anxiety, in conjunction with a significant increases [sic] of Vital, Happy as well as Kind scores. The latter indicates a public that is energized, happy and friendly on election day. On November 5, these GPOMS dimensions continue to indicate positive mood levels, in particular high levels of Calm, Sure, Vital and Happy. After November 5, all mood dimensions gradually return to the baseline. The public mood response to Thanksgiving on November 27, 2008 provides a counterpart to the differentiated response to the Presidential Election. On Thanksgiving day we find a spike in Happy values, indicating high levels of public happiness. However, no other mood dimensions are elevated on November 27. Furthermore, the spike in Happy values is limited to 1 day, i.e. we find no significant mood response the day before or after Thanksgiving.

The second method of validating their GPOMS tool is to regress the mood time series it generates on the OF time series. Presumably, we should assess the positive correlations detected between GPOMS Sure and Happy moods and the OF tool as evidence that all three time series are detecting similar latent variables. This is a standard methodology in political science but for the four moods that have no linear relationship with the OF tool, BMZ write:

> ...certain GPOMS mood dimension [sic] partially overlap with the mood values provided by OpinionFinder, but not necessarily all the mood dimensions that may be important in describing the various components of public mood e.g. the varied mood response to the Presidential Election. The GPOMS thus provides a unique perspective on public mood states not captured by uni-dimensional tools such as OpinionFinder.

In other words, regardless of the outcome of the multiple regression analysis in TMP, BMZ would have assessed their proprietary GPOMS tool as providing insight into collective mood states. As BMZ admit in Section 3 of TMP, they have no access to the ground truth of actual collective mood state readings and no evidence that their tool gives correct collective mood estimates. They write:

> ...although we have cross-validated the results of 2 different tools to assess public mood states, we have no knowledge of the "ground truth" for public mood states nor in fact for the particular subsample of the population represented by the community of Twitter.com users. This problem can only be addressed by research into direct assessments of public mood states vs. those derived from online communities such as Twitter.

Unfortunately, BMZ never made their GPOMS tool public so no such follow-up research has taken place. BMZ identify the problem of assessing the internal validity of mood analytics on social media datasets as a critical gap in the literature. Our survey of the mood analytics literature suggests that there exists no research correlating mood analyses obtained "in the wild" from social media with collective mood states offline. BMZ present no validation of their tool's ability to score individual Tweets, as is the industry and academic standard for mood analytics tools (Alm et al., 2005; Aman and Szpakowicz, 2008; Wang et al., 2012; Liu, 2012). Because we are chiefly concerned with replicating their time series over deriving a better mood analytics tool, we rely only on inspection between the graphs we generate and the apparent minima and maxima in TMP's normalized graphs.

In addition to the time series we construct by word count methods, we use the *sentimentr* R package's *sentiment_by* function to score Tweets using the extended composed-anxious lexicon we construct below. This is a commonly used tool in sentiment and mood analysis and a casual inspection of the outputs from it suggests that *sentimentr*'s scoring algorithm strictly dominates all word count methods in correctly classifying individual Tweets.

3.2.2 Character Matching

In our dataset, we count terms by simply matching strings of characters in Tweets to those in our lexicon. This is an unorthodox choice in computational linguistics but less unorthodox, though still rare, in Twitter applications.[39] The major risk of matching by characters is incorrect classification of covariates. For instance, "poise" is stemmed to "pois" which will match with "poison."[40] For Tweets, however, the 140 character limit incentivizes the use of the shorter words so that these kind of misclassifications are much less likely to occur than in a generic document. Second, hand inspection of 3000 Tweets found no examples of these kind of perverse matches occurring within our lexicons. However, we found many examples of incorrect negation matches. For instance, the word "normal" is unaffected by stemming and so it matches with both "nor" and "no" because both of these words are subsets of the word "normal". We correct this by only counting a negation term if there is also whitespace after it.[41] Character matching is computationally efficient and overcomes difficulties arising from the fact that many Twitter users appear to omit spaces between their words. For instance, in our data dataset we have the Tweet "im tensetensetense!" which will

[39] The standard procedure is to tokenize the Tweets. A token is a "part" of a sentence usually corresponding to a string or word. IBM Research defines a token as "more...than simply...strings delimited on both sides by spaces or punctuation. Different notions depend on different objectives, and often different language backgrounds. A token is linguistically significant and methodologically useful." Once a text has been tokenized it is easy to construct a bag of words representation of the text. Our word count uses the bag of words representation in the sense that we rely exclusively on word counts. The secondary time series generated by *sentimentr* takes word order into account and so it is not a bag of words method. https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en, accessed 2015/01/01.

[40] "Poison" cannot be stemmed further.

[41] The most perverse example we found was the word "nor" matching with "no" and "nor" thus giving it no effect.

not be detected by any method looking for the word “tense” surrounded by whitespace. On the other hand, our method correctly detects three instances of stemmed term “tens”.

3.2.3 Kernel-Lexicon Selection

Finally, we are tasked with constructing our daily mood time series. Let X_t be the collective mood value for the day. Day t has I_t Tweets $\in \{1, \ldots, I\}$. Each Tweet i consists of a BOW T_i . N_i is a count of the words in Tweet i that are negation terms. Then, with (ℓ, K) as our kernel-lexicon pair in which $\ell^C \cup \ell^A = \ell$, we have day t's collective mood as given in:

$$X_t := \frac{1}{I_t}\left[\sum_{i_t=1}^{I_t}(-1)^{N_{i_t}} K(\#\{T_{i_t} \cap \ell^C\}, \#\, \{T_{i_t} \cap \ell^A\}, \#\{T_{i_t}\})\right] \tag{23}$$

where $\#\{T_{i_t} \cap \ell\}$ is the number of matches between our lexicon and the words in Tweet *i* from day *t*.[42] Because it is possible that including negation as implemented (23) simply adds noise, we construct all our time series with and without negation terms. The negated subjectivity kernel has no natural interpretation and there is no reason to believe that the subjectivity kernel paired with any of our lexicons will replicate BMZ's GPOMS time series. We include it in our analysis anyway. Excluding negation terms is equivalent to setting $N_i = 2$ for all i in (23).

Monopolar Kernels	**Functional Form**
McDonald	C
Loughran	$-A$
Scaled McDonald	C/W
Scaled Loughran	$-A/W$
Subjectivity	$C + A$
Scaled Subjectivity	$(C + A)/W$
Bipolar Kernels	**Functional Form**
Breen	$(C - A)$
Scaled Breen	$(C - A)/W$

[42] In fact, all aggregation methods covered, including the Breen and Tetlock approaches, were tried. They give similar results to (23).

Scaled Tone	$\frac{(C-A)}{C+A}$

Table II.3.2: This table contains all of the bipolar kernels we will test. Let C be the number of matches from the composed lexicon and A be the number of matches from the anxious lexicon. Let W be the number of words in the Tweet.

We place all kernels we are going to test into Table II.3.2. Overall, we consider a total of 18 mood time series since we have nine kernels and a choice of whether or not to consider negation.

3.2.4 Visualizing Our Kernel-Lexicon Time Series

We visualize selected time series in Figures II.3.4 through II.3.12. Our results are striking. Many of our monopolar and bipolar kernels can identify all but one local minima and maxima, typically August 1, 2008. Nearly all kernels have statistically significant positive pairwise correlation with each other. As a sanity check, our subjectivity kernels, which we would not expect to proxy for the CALM time series in TMP a priori, are unable to identify a single minima or maxima in Table II.3.1's variable selection criterion. Nonetheless, every single kernel is rejected by the visual variable selection criterion we set out in above with one exception: $(\ell_I, Tone)$, using the aggregation method that ignored negation. We note that, relative to sentiment analysis techniques we have tested on the dataset for related projects, our inspection lexicon and tone kernel lead to a higher variance time series. It also leads to more local extreme values than all of the bipolar kernels we test in this project. Thus, we must note the possibility that it fits the selection criterion we place in Table 3.1 by chance.[43] Nonetheless, this will be the time series we use throughout the main text.

3.2.5 Logistical Details: Weekends and Holidays

Finally, before we can relate the DJIA and our mood time series, we must find a systematic way to deal with holidays and weekends. Specifically, the DJIA does not trade over these time periods, but we still have information about collective mood from these periods that we may not wish to discard. BMZ provide no details about how they adapt their mood time series to weekends and holidays, but one of the papers cited in TMP, Gilbert and Karahalios (2010) suggests that, for collective mood time series with a small number of data points, one should take the maximum value for the mood time series across the most recent

[43] Because we have no access to the ground truth, we have no way of verifying whether or not this is the result of our inspection lexicon being more a more powerful estimator or the result of it fitting noise in our data.

trading day, weekends and holidays. Since their time series purported to measure collective anxiety while ours is supposed to measure collective “calm”-ness, we follow a variation on their procedure, creating three time series. In the first time series, we take the minimum across weekends, holidays and the most recent trading day (before said weekends or holidays) and assign this minimum to the most the recent trading day. In the second time series, we follow the same procedure, but take the maximum instead. Intuitively, we expect that taking the minimum across weekends should yield better in-sample performance and this is what we find in our in-sample tests. Finally, in the third time series we simply drop weekend values. We work with the time series that drops weekends and the time series that takes minimums in the main text.

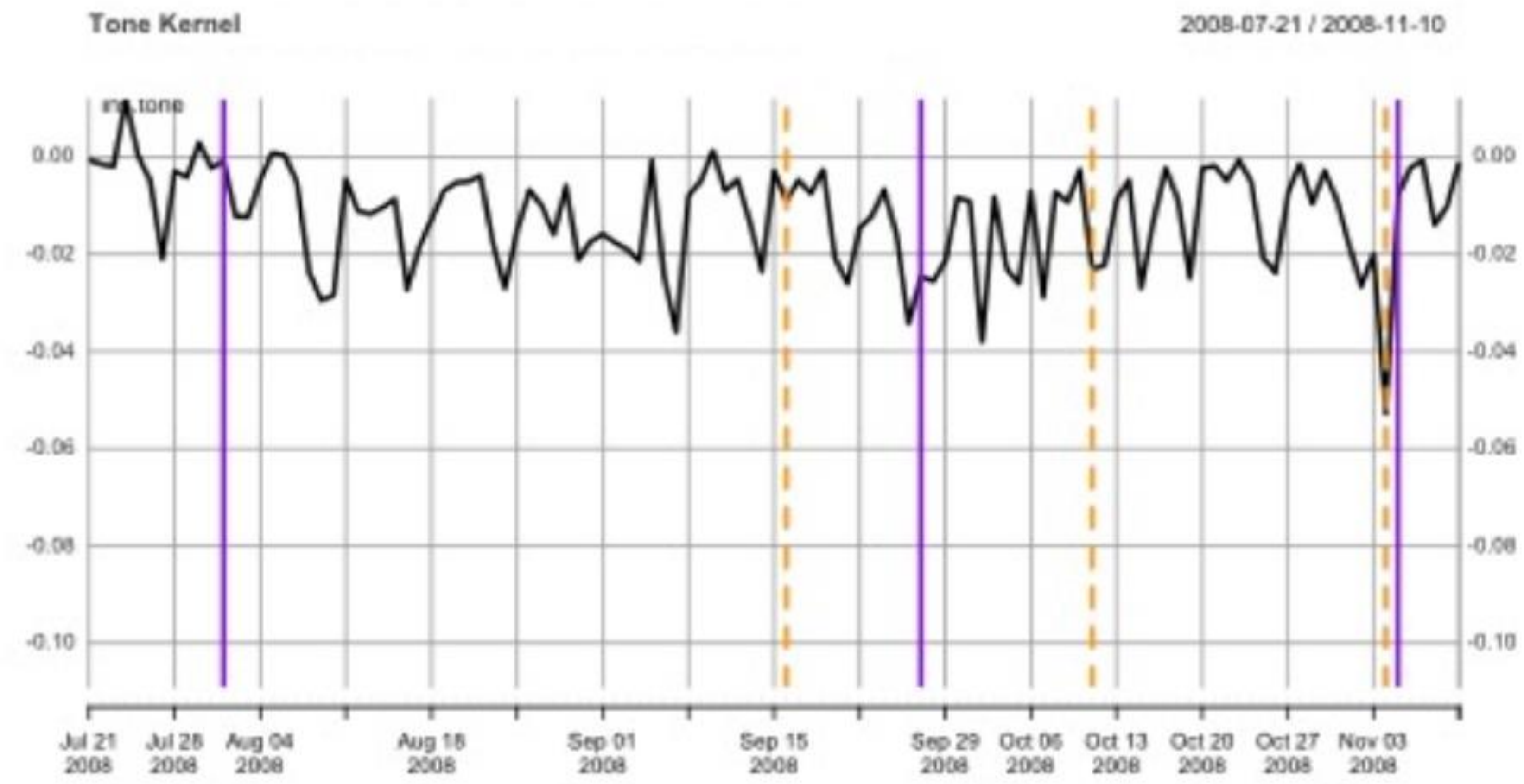

Figure II.3.4: This figure shows time series generated by the tone kernel. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). This kernel, paired with the lexicon generated by inspection, successfully identifies all local maxima, minima and relative increases identified in Table 3.1.

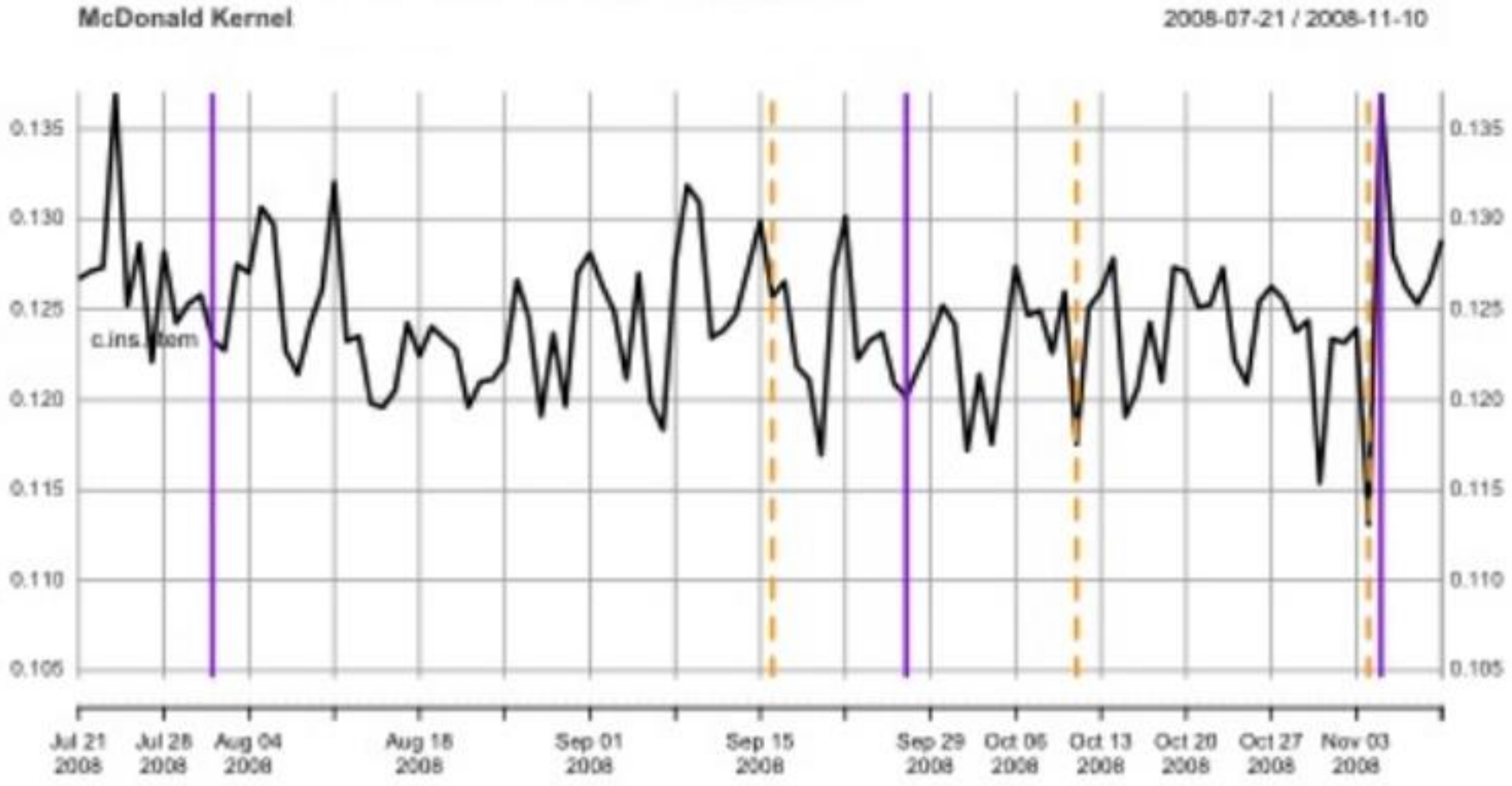

Figure II.3.5: This figure shows time series generated by the McDonald kernel. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel on the basis of its inability to identify August 1, 2008 as a local maximum.

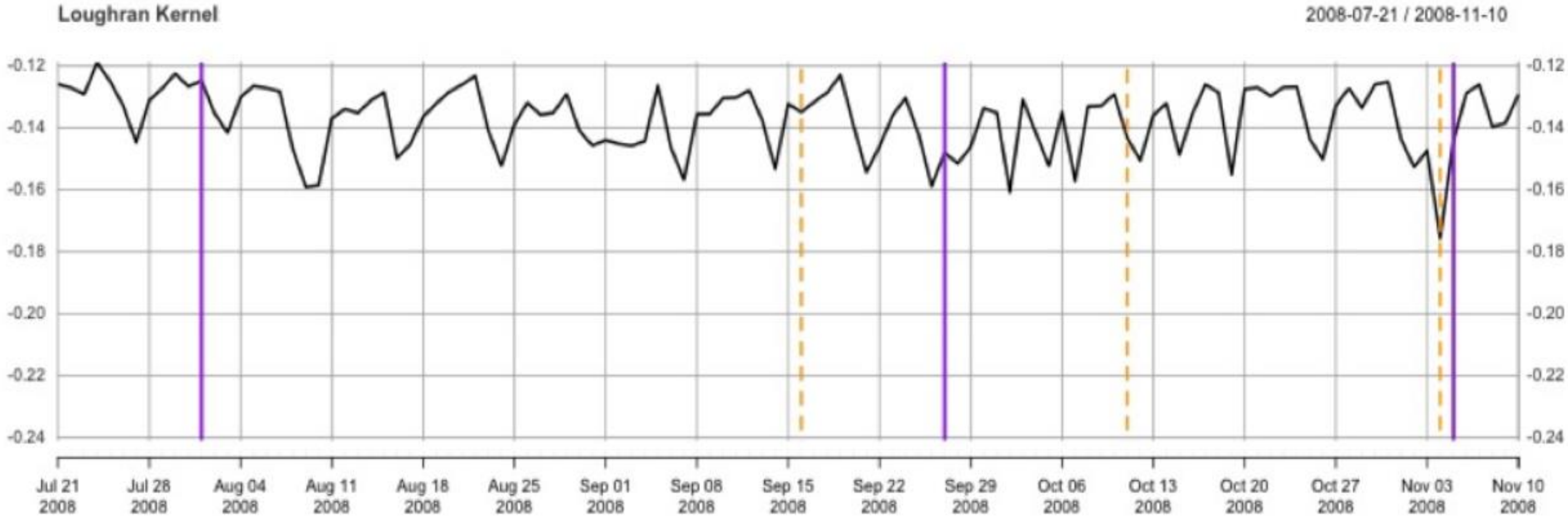

Figure II.3.6: This figure shows time series generated by the Loughran kernel. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel on the basis of its inability to identify October 11, 2008 as a local minimum.

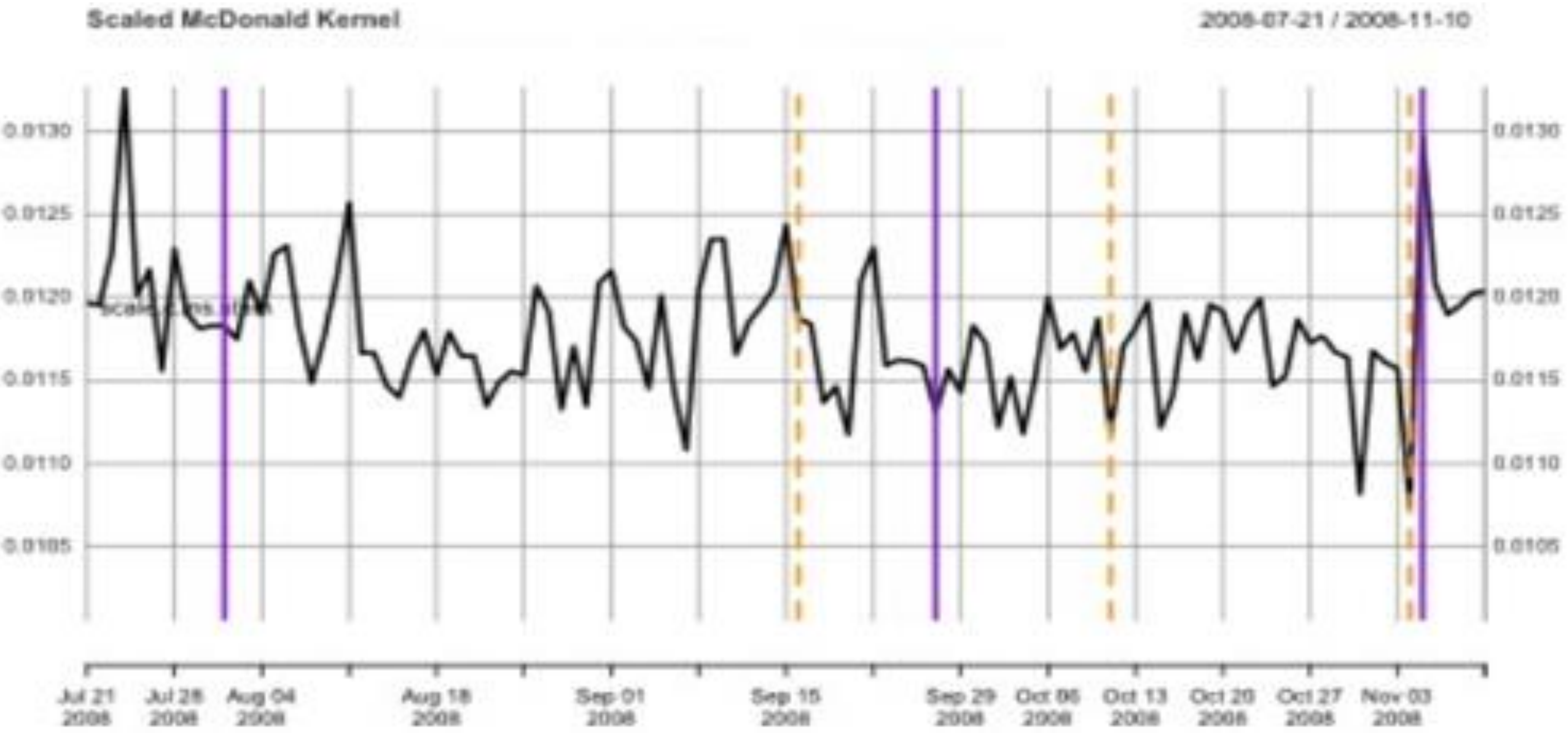

Figure II.3.7: This figure shows time series generated by the scaled McDonald kernel on both lexicons. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel on the basis of its inability to identify September 16, 2008 as a local maximum

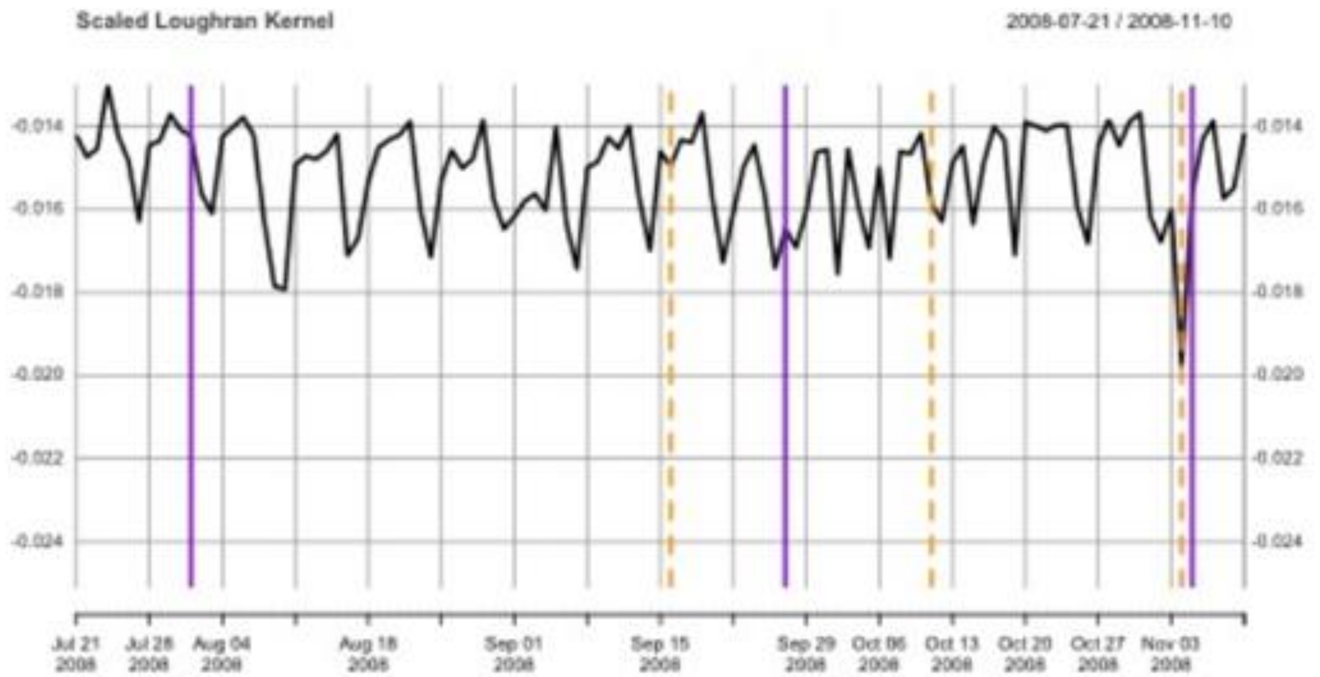

Figure II.3.8: This figure shows time series generated by the scaled Loughran kernel on both lexicons. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel on the basis of its inability to identify August 1, 2008 as a local maximum.

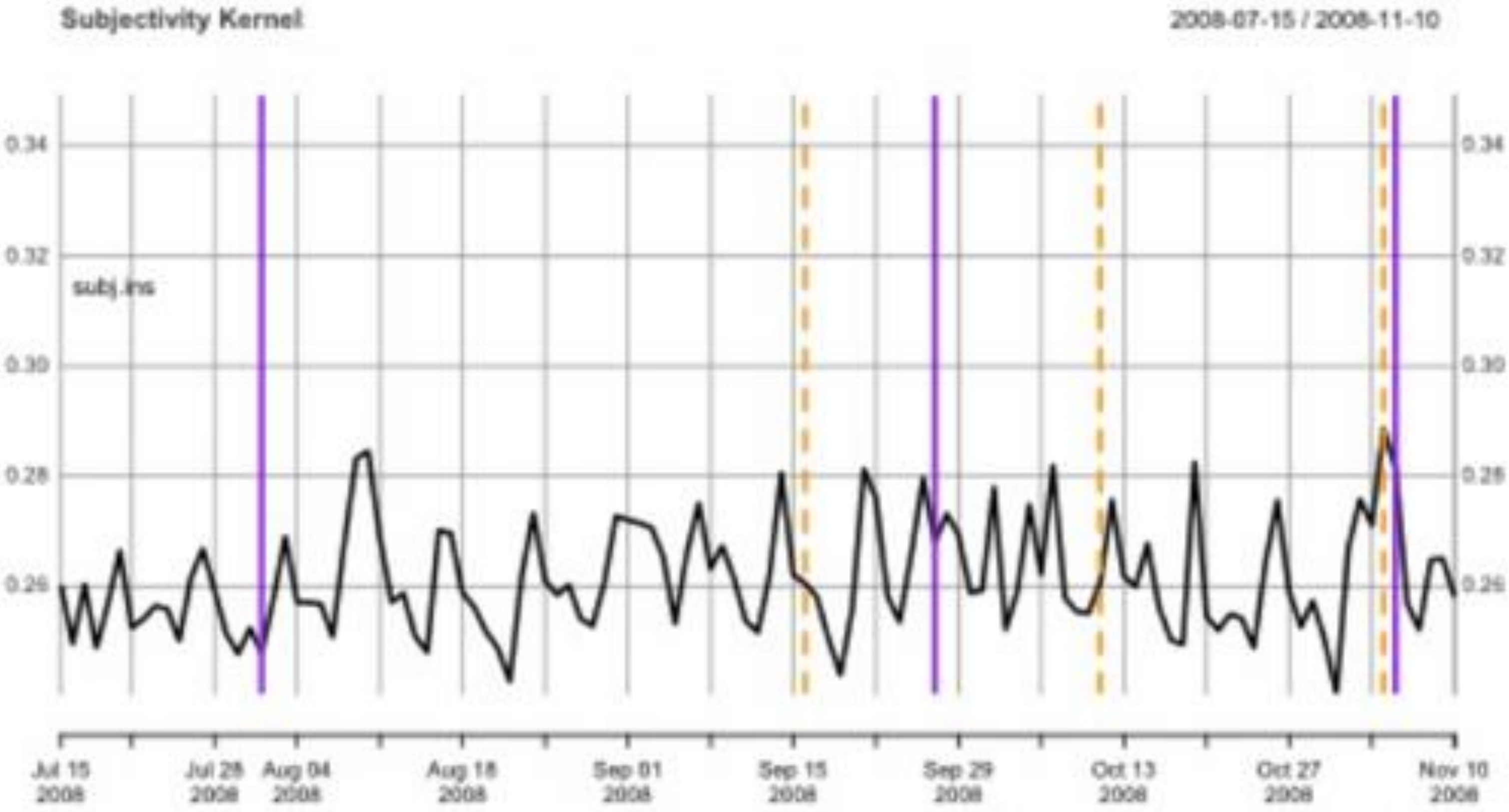

Figure II.3.9: This figure shows time series generated by the subjectivity kernel. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel because it is unable to identify any of our selection criteria's local minima or maxima.

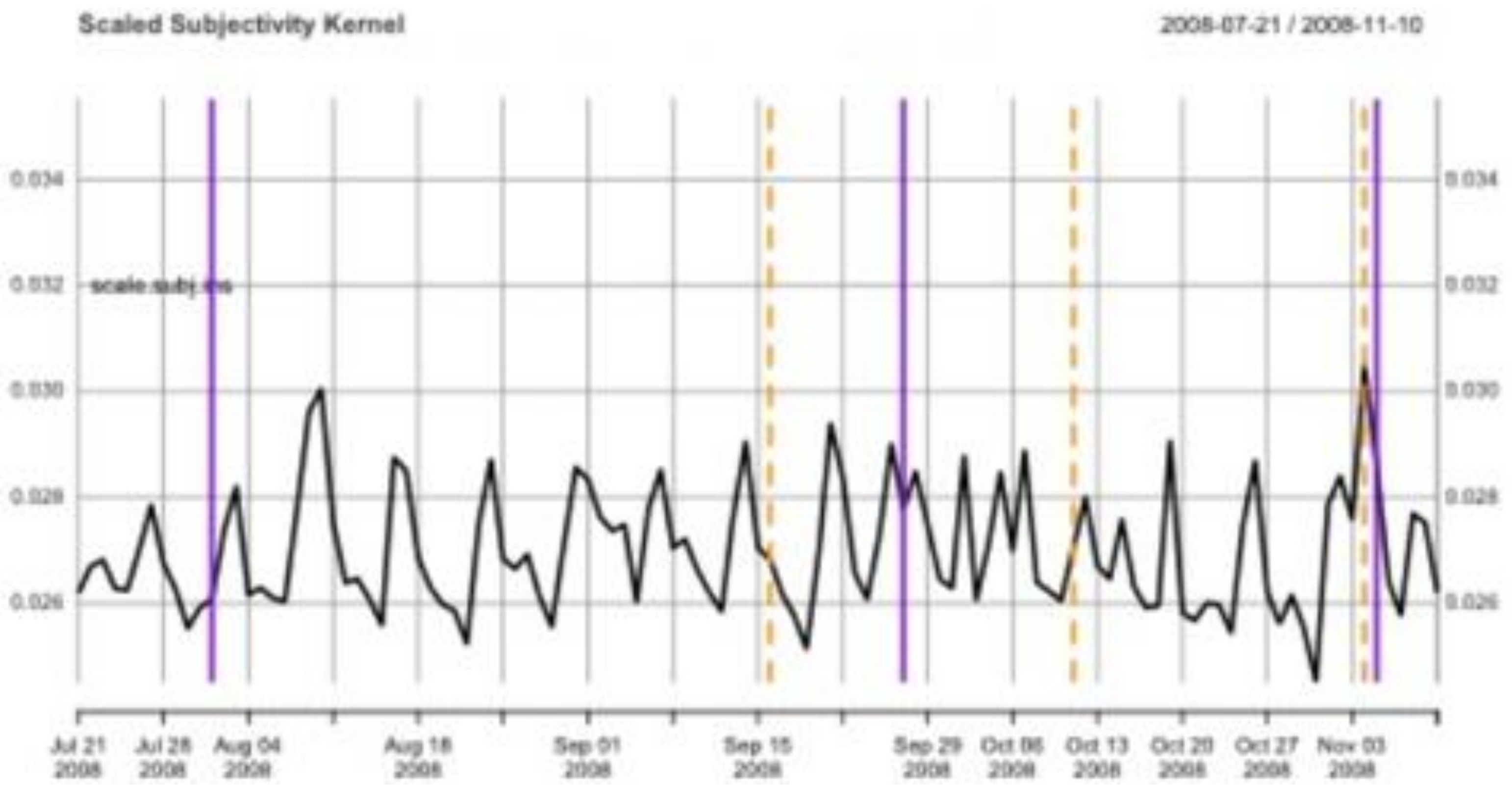

Figure II.3.10: This figure shows time series generated by the scaled subjectivity kernel on both lexicons. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel because it is unable to identify any of our selection criteria's local minima or maxima in our selection criteria.

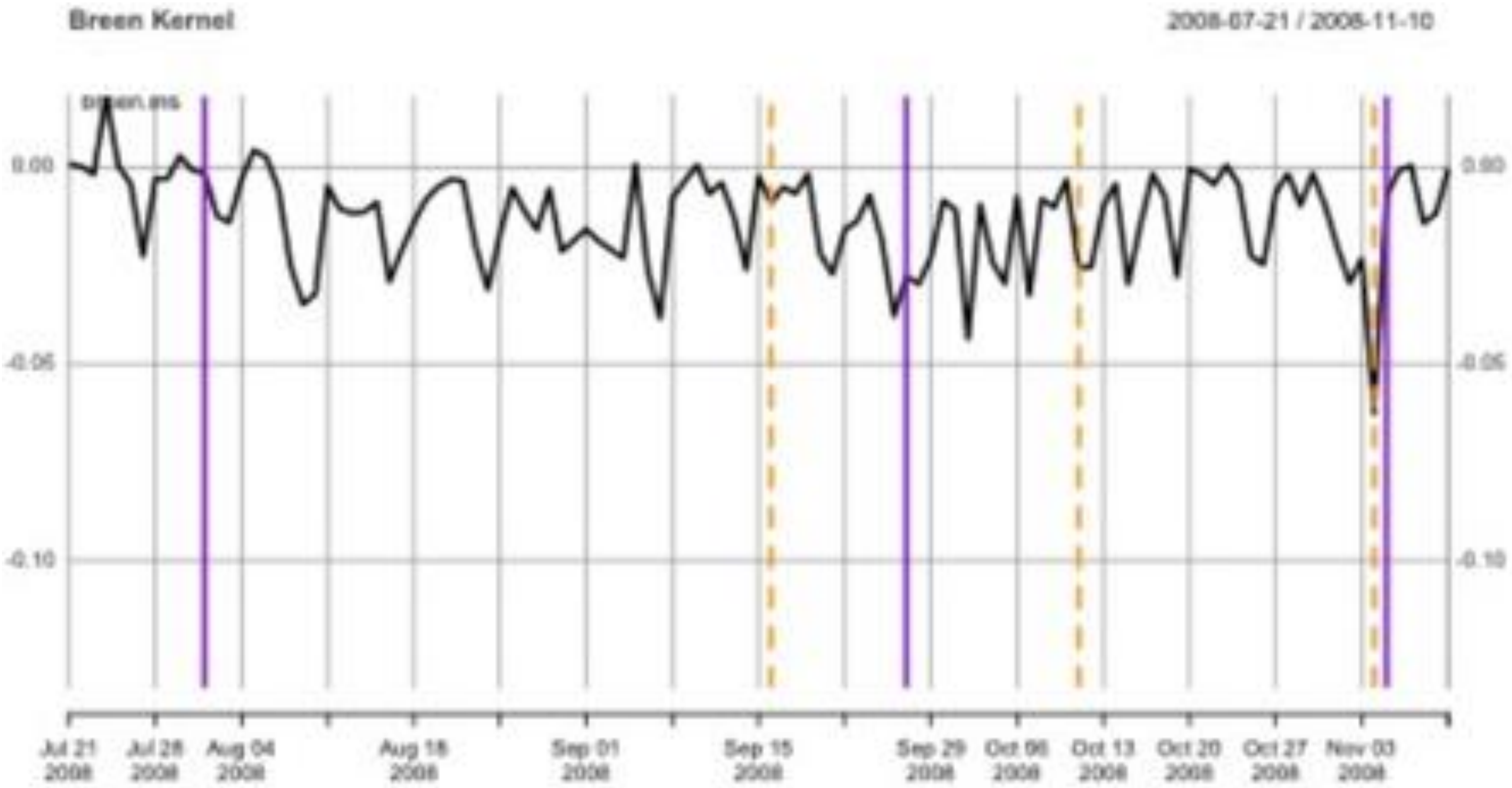

Figure II.3.11: This figure shows time series generated by the Breen kernel on both lexicons. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel on the basis of its inability to identify August 1, 2008 as a local maximum.

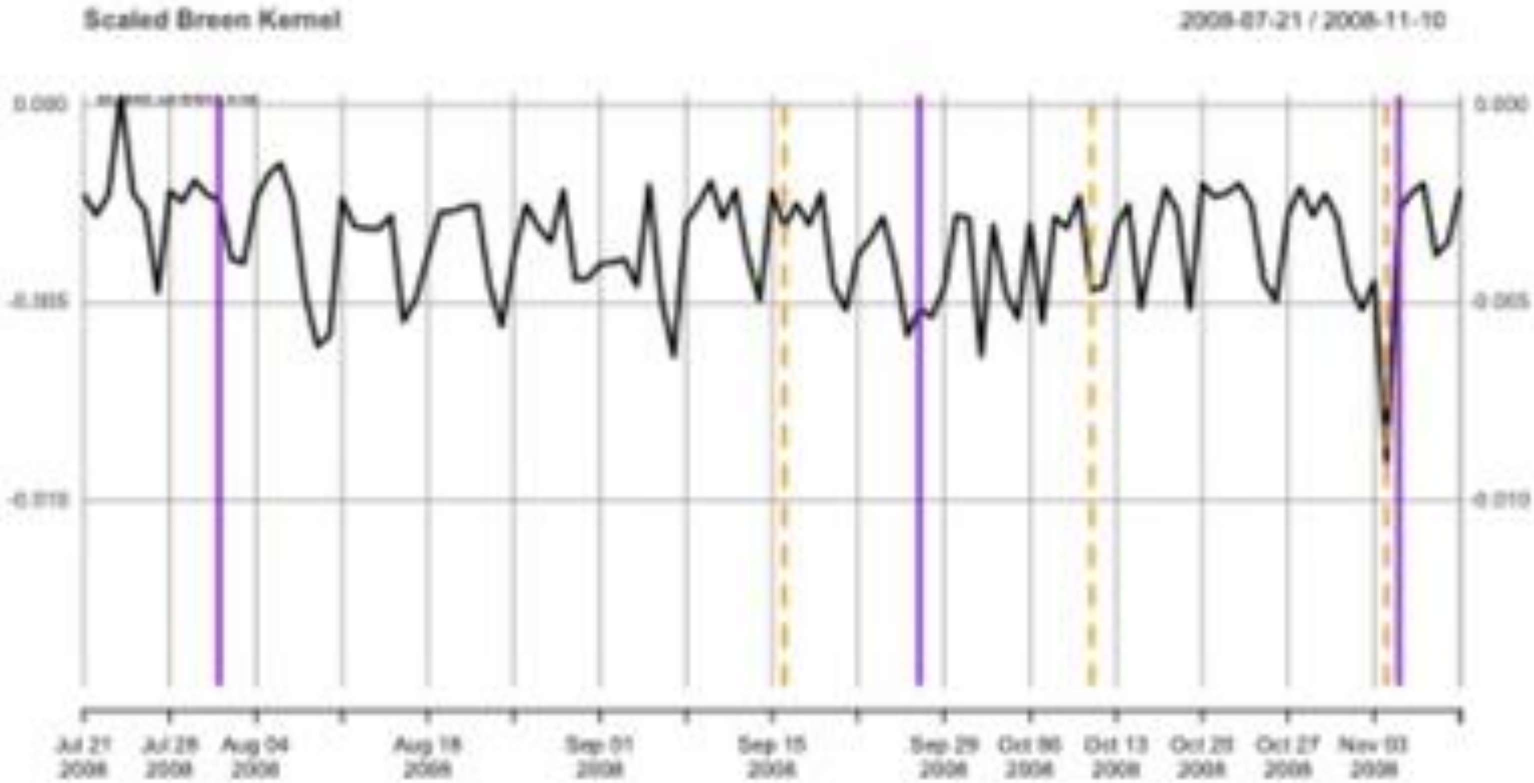

Figure II.3.12: This figure shows time series generated by the scaled Breen kernel on both lexicons. The orange, dotted lines correspond with where we should find local minima. From right to left, the first two purple, solid vertical lines correspond with where we should find local maxima while the third solid purple, line should correspond with the largest local increase (the day after the 2008 Presidential election). We reject this kernel on the basis of its inability to identify August 1, 2008 as a local max