APPENDIX A

Understanding the Thomson Reuters MarketPsych Indices

Since 2004, MarketPsych has honed its unique methodology for extracting detailed, relevant concepts from a variety of business and investment text. The MarketPsych lexicon is an extensive, expert-curated repository of simple and complex English-language words and phrases of potential interest for traders, investors, and economists. Used in conjunction with the MarketPsych lexicon, MarketPsych's natural-language processing software employs grammatical templates customized to extract meanings from financial news, social media, earnings conference call transcripts, and executive interviews.

SOURCE TYPE CUSTOMIZATION

There is a vast difference in communication styles between social and news media. Compared to news, social media contains significant levels of sarcasm and irony, incomplete thoughts, misplaced or excessive punctuation, misspellings, nonstandard grammar, case insensitivity, and crude language. Additionally, in social media many common words are used with colloquial meanings. A statement such as "That trade was the bomb!" with reference to a successful trade is far different from a reference to warfare, as would be interpreted by a historically trained linguistic analysis engine.

Because new colloquial language enters social use periodically, including expressions such as "You killed it!" (as a compliment), MarketPsych's text analytics dictionaries and grammatical technology are updated every two to three years (we're on commercial version 2.2 currently). When new proper nouns or companies enter the lexicon, including countries such as "South Sudan" or companies such as "China Life," they are included during monthly entity updates.

Trading on Sentiment: The Power of Minds Over Markets. Richard L. Peterson

New text sources are added to the data feed over time as they become active, such as Twitter content in 2009. Over the years, the media and its audience migrate; most notably Yahoo! Finance message board volume has dropped by 80 percent while social media–consuming investors migrated to alternative social media sites such as Twitter and SeekingAlpha. Eventually, these sources will fade in significance as well. Given the changing nature of communication over the past 17 years of social Internet data, MarketPsych's analysts look for universal themes in text topics and in source audiences, and the focus is domain-specific. For example, only business, investing, and political articles are accepted for text analytics. Sometimes entertainment articles are included, as when two movie studios are undergoing a corporate merger, but these are excluded if they are not related to corporate activity.

A significant difference between social and news media lies in how viewpoints are conveyed. In social media, there is typically less editorial oversight and more leeway for a passionate author to unreservedly express his or her opinion or emotional state. In contrast, journalists are trained to offer multiple perspectives on the underlying story. Rather than conveying their own emotions, journalists see their role as describing the emotional states of those they are reporting on. As a result, information obtained from social media is typically less inclusive of contrary viewpoints and more emotionally expressive from the first-person perspective than news information.

Direct expressions of emotion in news and social media also vary. In social media, authors may utilize a complex array of text or graphic emoticons (e.g., ">:-(") and acronyms (e.g., "LOL") that developed organically, with regional, industrial, and national differences. Furthermore, word context is much more important in social media than in news media for interpreting intended meaning.

As a result of all these differences between news and social media, sentiment scoring accuracy is improved by text analytic models calibrated to source type. MarketPsych currently uses differentiated models for news, social media forums, tweets, SEC filings, and earnings conference call transcripts.

LEXICAL ANALYSIS

There are a variety of approaches used in sentiment analysis. The most common technique is called lexical analysis, and this approach is used in many historical academic studies of sentiment and stock returns.[1] Lexical analysis identifies explicit words and phrases in a body of text. Relevant content is organized and scored according to a hard-coded ontology. The simplest example of a lexical approach is called "bag of words." In the "bag of words" technique, all words are counted according to their frequency, and no additional grammatical or relational post-processing is performed.

There are several known limitations to a purely lexical approach. The most significant one, for the purposes of producing TRMI, is that most lexical approaches are focused only on extracting one-dimensional sentiment. In cases where a variety of sentiment dimensions may be scored using lexical analysis, such as when using the *Harvard General Inquirer* dictionary, the word tokens representing specific sentiments are occasionally incongruent with meanings in contemporary business English.

Another weakness of using uncurated dictionaries is lexical ambiguity across domains. For example, financial terms such as *investor* and *financier* are classified as negative sentiment terms in some open-source sentiment dictionaries. MarketPsych has overcome lexical ambiguity with extensive business-specific customization and curation of lexicons.

Insensitivity to grammatical structures is perhaps the most significant weakness of the lexical approach. In order to address this weakness, MarketPsych engineers embedded a complex grammatical framework with traits specific to different text sources such as social media, earnings conference call transcripts, financial news, and regulatory filings. The result is that customized lexicons, superior disambiguation, and optimized grammatical structures stand behind MarketPsych's textual analytics. For space reasons, we will not describe the grammatical nuances of the natural language processing underlying the TRMI.

ENTITY IDENTIFICATION AND CORRELATE FILTERING

Consider that entities such as IBM may be referred to as "IBM," "Big Blue," and "International Business Machines" in the press. Additionally, international press may or may not use accent marks in common location names such as Düsseldorf. In order to identify entities such as IBM and Düsseldorf that have multiple spellings or reference names, MarketPsych prepared a list of over 60,000 entity names with aliases. This list has been improved by human review, and it is updated monthly with new and changed (acquired, merged, etc.) entities.

To improve entity name disambiguation, MarketPsych used supervised machine learning to identify correlate and anti-correlate words in proximity of ambiguous entity references. For example, gold and silver are commonly spoken of as both commodities and constituents of jewelry, but every two years they are frequently mentioned as Olympic medals. To prevent entity identification errors, anti-correlate filters are utilized to eliminate Olympic references such as "gold medal" and "won a silver." Another example is the South Korean won, which could be confused with a successful competition by a South Korean athlete who "won" an event. Anti-correlate filtering and case-sensitivity both improve precision of the scoring process and entity identification.

In addition to an anti-correlate filter to exclude irrelevant entities, for some entities MarketPsych software uses a correlate filter to ensure that only entities with the correct co-references are included in the entity identification. For example, when a Twitter user tweets that "I am enjoying my instant oats," MarketPsych's software will not count that reference as applicable to the commodity oats. References to oats are counted only if they also contain key identification correlates such as "prices" and "futures."

LINGUISTIC ANALYSIS FLOW

When applied to text, the confluence of the various text processing described above generates over 4,000 variables (Vars), each with the potential to be applied to a different entity. Alphabetically, a few Vars include:

AccountingBad

AccountingGood

Ambiguity

Anger

Each Var is then qualified by tense, such as the following:

AccountingBad_n: present-tense negative accounting news

AccountingGood_p: past-tense positive accounting news

Ambiguity_c: conditional-tense uncertainty

Anger_f: anger about anticipated events

SENTENCE-LEVEL EXAMPLE

Using the principles outlined above, let's now take a closer look at the MarketPsych software in action and see how it analyzes the following sentence:

"Analysts expect Mattel to report much higher earnings next quarter."

The language analyzer performs the following sequence:

1. Associates ticker symbol MAT with entity reference "Mattel."
2. Identifies "earnings" as an Earnings word in the lexicon.

3. Identifies “expect” as a future-oriented word and assigns future tense to the phrase.
4. Identifies “higher” as an Up-Word.
5. Multiplies “higher” by 2 due to presence of the modifier word “much.”
6. Associates “higher” (Up-Word) with “earnings” (Earnings) due to proximity.

The analysis algorithm will report:

Date	Time	Ticker	Var	Score
20110804	15:00.123	MAT	*EarningsUp_f*	2

In the example above, 2 is the raw score produced for EarningsUp_f.

CREATING AN INDEX

The TRMI themselves derived from two groups of sources—news and social media—and the data feed itself consists of three feeds: a social media feed, a news media feed, and an aggregated feed of combined social and news media content. The TRMI are updated minutely. Over 2 million articles are processed daily and contribute to the TRMI feed within minutes of their publication. The following sections further describe the construction of the TRMI, from raw content to Vars to published TRMI.

SOURCE TEXT

The TRMI are derived from an unparalleled collection of premium news, global Internet news coverage, and a broad and credible range of social media. The TRMI social media feed consists of both MarketPsych and Moreover social media content. Moreover Technologies’ aggregated social media feed is derived from tens of thousands of social media sites and is incorporated into the TRMI from 2009 to the present. MarketPsych social media content was downloaded from public social media sites from 1998 to the present.

The TRMI News indices are derived from live content delivered via Thomson Reuters News Feed Direct and two Thomson Reuters news archives: a Reuters-only one from 1998 to 2002 and one with Reuters and select third-party wires from 2003 to the present. In addition, we incorporate Moreover Technologies aggregated newsfeed, which is derived from 40,000 Internet news sites and spans 2005 to present. MarketPsych crawler content from hundreds of financial news sites is also included. MarketPsych-specific sources of text include *The New York Times, The Wall Street Journal, Financial Times, Seeking Alpha*, and dozens more sources widely read by professional investors.

Figure A.1 shows a graphic displaying the time course of each text feed within the TRMI. The TRMI thus cover the period 1998 through the present. Currently, all source text for the MarketPsych sentiment products is English-language.

INDEX CONSTRUCTION

Each TRMI is composed of a combination of variables (Vars). First, the absolute values of all TRMI-contributing Vars, for all asset constituents, over the past 24 hours are determined. These absolute values are then summed for all constituents. This sum is called the "Buzz," and it is published in conjunction with each asset's TRMIs. More specifically, where V is the set of all Vars underlying *any* TRMI of the asset class, where *a* denotes an asset, and where C(*a*) is the set of all constituents* of *a*, we can define the Buzz of *a* as the following:

$$Buzz(a) = \sum_{c \in C(a),\ v \in V} |Var_{c,v}|$$

Each TRMI is then computed as a ratio of the sum of all relevant Vars to the Buzz. We define $V(t)$ as the set of all Vars relevant to a particular TRMI t. Next we define a function to determine whether a Var $v \in V(t)$ is additive or subtractive to a TRMI as the following:

$$I(t, v) = \begin{cases} +1 \ \textit{if additive} \\ -1 \ \textit{if subtractive} \end{cases}$$

Thus the TRMI t of asset a can be computed as the following:

$$TRMI_t(a) = \frac{\Sigma_{c \in C(A),\ v \in V(t)} (I(t, v) \times PsychVar_v(c))}{Buzz(Asset)}$$

* For example, Mattel is a constituent of MarketPsych's Nasdaq 100 index proxy asset (MPQQQ).

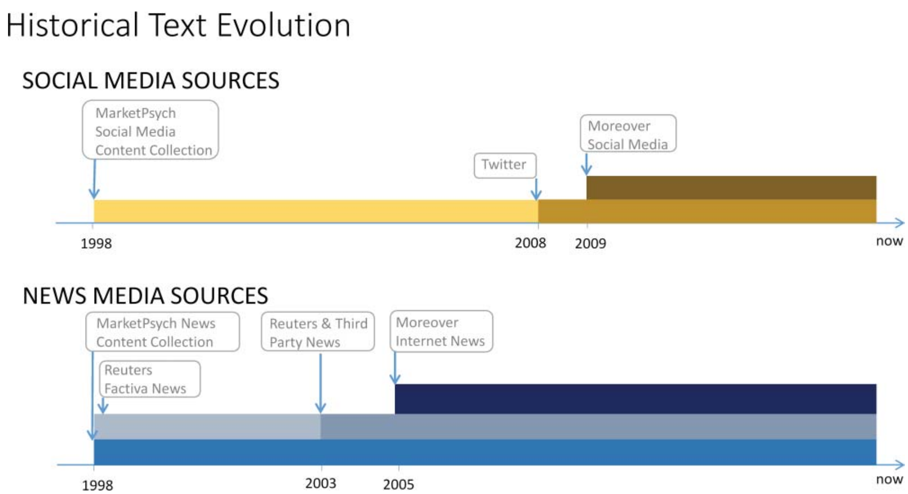

FIGURE A.1 Timeline of textual content analyzed for the social and news media TRMI.

FIGURE A.2 Asset classes covered by the Thomson Reuters MarketPsych Indices.

It's worth noting that, particularly for Equities where the assets all correspond to indices and sectors, an individual constituent may contribute to multiple assets. For example, Mattel is a constituent of both the Consumer Goods sector and the Nasdaq 100 index proxies. As a result, Mattel's Var scores will be incorporated into the TRMI for both.

Similarly, a single Var can contribute to multiple TRMI. For example, the earningsUp_f Var noted in the "Sentence-level Example" section above is not only a constituent of earningsForecast but also of the Sentiment, Optimism, and fundamentalStrength TRMI.

ASSET CLASSES COVERED

The Thomson Reuters MarketPsych Indices cover tradable assets in five different asset classes. Please see an abbreviated list of coverage in Figure A.2.

TRMI DEFINITIONS

The Thomson Reuters MarketPsych Indices consist of several different sentiments, 14 of which are common to all five scored asset classes. Macroeconomic and topic TRMI vary by asset class. More documentation about the individual assets and indices covered is available in the online Thomson Reuters MarketPsych Indices User Guide.[2]

Company and Equity Index TRMI Indices

There are 31 TRMI indices for the companies and equity index asset classes. Each TRMI carries six significant digits past the decimal point. Negative numbers have a leading minus (−) sign. The table below summarizes these fields.

Index	**Description:** *Score of references in news and social media to …*	Range
sentiment	overall positive references, net of negative references	−1 to 1
optimism	optimism, net of references to pessimism	−1 to 1
fear	fear and anxiety	0 to 1
joy	happiness and affection	0 to 1
trust	trustworthiness, net of references connoting corruption	−1 to 1

(*continued*)

Index	**Description:** *Score of references in news and social media to …*	Range
violence	violence and war	0 to 1
conflict	disagreement and swearing net of agreement and conciliation	–1 to 1
gloom	gloom and negative future outlook	0 to 1
stress	distress and danger	0 to 1
timeUrgency	urgency and timeliness, net of references to tardiness and delays	–1 to 1
uncertainty	uncertainty and confusion	0 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	–1 to 1
longShort	buying, net of references to shorting or selling	–1 to 1
longShortForecast	forecasts of buying, net of references to forecasts of shorting or selling	–1 to 1
priceDirection	price increases, net of references to price decreases	–1 to 1
priceForecast	forecasts of asset price rises, net of references to forecasts of asset price drops	–1 to 1
volatility	volatility in market prices or business conditions	0 to 1
loveHate	love, net of references to hate	–1 to 1
anger	anger and disgust	0 to 1
debtDefault	debt defaults and bankruptcies	0 to 1
innovation	innovativeness	0 to 1
marketRisk	positive emotionality and positive expectations net of negative emotionality and negative expectations. Includes factors from social media found characteristic of speculative bubbles—higher values indicate greater bubble risk. Also known as the "bubbleometer."	–1 to 1
analystRating	upgrade activity, net of references to downgrade activity	–1 to 1
dividends	dividends rising, net of references to dividends falling	0 to 1

Index	**Description:** *Score of references in news and social media to …*	**Range**
earningsForecast	expectations about improving earnings, less those of worsening earnings	−1 to 1
fundamentalStrength	positivity about accounting fundamentals, net of references to negativity about accounting fundamentals	−1 to 1
layoffs	staff reductions and layoffs	0 to 1
litigation	litigation and legal activity	0 to 1
managementChange	changes in a company's management team, net of references to stability in the management team	−1 to 1
managementTrust	trust expressed in a company's management team, net of references to reports of unethical behavior among the management team	−1 to 1
mergers	merger or acquisition activity	0 to 1

Currency TRMI Indices

There are 21 TRMI indices for the currency asset class.

Index	**Description:** *Score of references in news and social media to …*	**Range**
sentiment	overall positive references, net of negative references	−1 to 1
optimism	optimism, net of references to pessimism	−1 to 1
fear	fear and anxiety	0 to 1
joy	happiness and affection	0 to 1
trust	trustworthiness, net of references connoting corruption	−1 to 1
violence	violence and war	0 to 1
conflict	disagreement and swearing net of agreement and conciliation	−1 to 1
gloom	gloom and negative future outlook	0 to 1
stress	distress and danger	0 to 1

(*continued*)

Index	**Description:** *Score of references in news and social media to …*	Range
timeUrgency	urgency and timeliness, net of references to tardiness and delays	–1 to 1
uncertainty	uncertainty and confusion	0 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	–1 to 1
longShort	buying, net of references to shorting or selling	–1 to 1
longShortForecast	forecasts of buying, net of references to forecasts of shorting or selling	–1 to 1
priceDirection	price increases, net of references to price decreases	–1 to 1
priceForecast	forecasts of asset price rises, net of references to forecasts of asset price drops	–1 to 1
volatility	volatility in market prices or business conditions	0 to 1
loveHate	love, net of references to hate	–1 to 1
carryTrade	carry trade	0 to 1
currencyPegInstability	the instability of a currency peg, net of references to the stability of a currency peg	–1 to 1
priceMomentum	currency price trend strength, net of references to trend weakness	–1 to 1

Agricultural Commodity TRMI Indices

There are 27 TRMI indices for the agricultural commodity asset class.

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
sentiment	overall positive references, net of negative references	–1 to 1
optimism	optimism, net of references to pessimism	–1 to 1
fear	fear and anxiety	0 to 1
joy	happiness and affection	0 to 1

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
trust	trustworthiness, net of references connoting corruption	−1 to 1
violence	violence and war	0 to 1
conflict	disagreement and swearing, net of agreement and conciliation	−1 to 1
gloom	gloom and negative future outlook	0 to 1
stress	distress and danger	0 to 1
timeUrgency	urgency and timeliness, net of references to tardiness and delays	−1 to 1
uncertainty	uncertainty and confusion	0 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	−1 to 1
longShort	buying, net of references to shorting or selling	−1 to 1
longShortForecast	forecasts of buying, net of references to forecasts of shorting or selling	−1 to 1
priceDirection	price increases, net of references to price decreases	−1 to 1
priceForecast	forecasts of asset price rises, net of references to forecasts of asset price drops	−1 to 1
volatility	volatility in market prices or business conditions	0 to 1
consumptionVolume	factors leading to increased consumption, net of references to factors leading to decreased consumption	−1 to 1
productionVolume	increased production, net of references to factors leading to decreased production	−1 to 1
regulatoryIssues	regulatory issues	0 to 1
supplyVsDemand	surplus supply and lack of demand, net of references to supply shortage and high demand	−1 to 1
supplyVsDemand Forecast	expectations of supply outstripping demand, net of references to expectations of demand outstripping supply	−1 to 1

(*continued*)

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
acreageCultivated	increases in acreage and crop cultivation, net or references to decreases in acreage and crop cultivation	–1 to 1
agDisease	commodity disease	0 to 1
subsidies	subsidies affecting commodity prices	0 to 1
subsidiesSentiment	increases in subsidies, net of references to decreases in subsidies	–1 to 1
weatherDamage	commodity weather damage	0 to 1

Energy and Material Commodity TRMI Indices

The 24 TRMI indices for the energy and material commodity asset class.

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
sentiment	overall positive references, net of negative references	–1 to 1
optimism	optimism, net of references to pessimism	–1 to 1
fear	fear and anxiety	0 to 1
joy	happiness and affection	0 to 1
trust	trustworthiness, net of references connoting corruption	–1 to 1
violence	violence and war	0 to 1
conflict	disagreement and swearing net of agreement and conciliation	–1 to 1
gloom	gloom and negative future outlook	0 to 1
stress	distress and danger	0 to 1
timeUrgency	urgency and timeliness, net of references to tardiness and delays	–1 to 1
uncertainty	uncertainty and confusion	0 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	–1 to 1
longShort	buying, net of references to shorting or selling	–1 to 1
longShortForecast	forecasts of buying, net of references to forecasts of shorting or selling	–1 to 1

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
priceDirection	price increases, net of references to price decreases	−1 to 1
priceForecast	forecasts of asset price rises, net of references to forecasts of asset price drops	−1 to 1
volatility	volatility in market prices or business conditions	0 to 1
consumptionVolume	factors leading to increased consumption, net of references to factors leading to decreased consumption	−1 to 1
productionVolume	increased production, net of references to factors leading to decreased production	−1 to 1
regulatoryIssues	regulatory issues	0 to 1
supplyVsDemand	surplus supply and lack of demand, net of references to supply shortage and high demand	−1 to 1
supplyVsDemand Forecast	expectations of supply outstripping demand, net of references to expectations of demand outstripping supply	−1 to 1
newExploration	new ventures/exploration	0 to 1
safetyAccident	safety accidents	0 to 1

Country TRMI Indices

The 48 TRMI indices for the country asset class.

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
sentiment	overall positive references, net of negative references	−1 to 1
optimism	optimism, net of references to pessimism	−1 to 1
fear	fear and anxiety	0 to 1

(*continued*)

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	**Range**
joy	happiness and affection	0 to 1
trust	trustworthiness, net of references connoting corruption	–1 to 1
violence	violence and war	0 to 1
conflict	disagreement and swearing net of agreement and conciliation	–1 to 1
gloom	gloom and negative future outlook	0 to 1
stress	distress and danger	0 to 1
timeUrgency	urgency and timeliness, net of references to tardiness and delays	–1 to 1
uncertainty	uncertainty and confusion	0 to 1
emotionVsFact	all emotional sentiments, net of all factual and topical references	–1 to 1
loveHate	love, net of references to hate	–1 to 1
anger	anger and disgust	0 to 1
debtDefault	debt defaults and bankruptcies	0 to 1
innovation	innovativeness	0 to 1
marketRisk	positive emotionality and positive expectations net of negative emotionality and negative expectations. Includes factors from social media found characteristic of speculative bubbles—higher values indicate greater bubble risk. Also known as the “bubbleometer.”	–1 to 1
budgetDeficit	a budget deficit, net of references to a surplus	–1 to 1
businessExpansion	businesses expanding, net of references to contraction	–1 to 1
centralBank	the central bank of a country	0 to 1
commercialRealEstateSentiment	positive references to commercial real estate, net of negative references	–1 to 1
consumerSentiment	positive consumer sentiment, net of references to negative consumer sentiment	–1 to 1

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
creditEasyVsTight	credit conditions being easy, net of references to credit conditions being tight	–1 to 1
economicGrowth	increased business activity, net of references to decreased business activity	–1 to 1
economicUncertainty	uncertainty about business climate, net of confidence and certainty	–1 to 1
economicVolatility	increasing economic volatility, net of economic stability	–1 to 1
financialSystem Instability	financial system instability, net of references to financial system stability	–1 to 1
fiscalPolicyLooseVs Tight	fiscal policy being loose, net of references to fiscal policy being tight	–1 to 1
governmentAnger	anger and disgust about government officials and departments	0 to 1
government Corruption	fraud and corruption in government, net of references to trust in government	–1 to 1
governmentInstability	governmental instability, net of references to governmental stability	–1 to 1
inflation	consumer price increases, net of references to consumer price decreases	–1 to 1
inflationForecast	forecasts of consumer price increases, net of forecasts of consumer price decreases (deflation)	–1 to 1
interestRates	interest rates rising, net of references to rates falling	–1 to 1
interestRatesForecast	forecasts of interest rates rising, net of forecasts of rates falling	–1 to 1
investmentFlows	investment inflows, net of references to investment outflows	–1 to 1
monetaryPolicyLoose VsTight	monetary policy being loose, net of references to monetary policy being tight	–1 to 1
naturalDisasters	natural disasters	0 to 1

(*continued*)

Index	**Description:** *24-hour rolling average score of references in news and social media to …*	Range
regimeChange	regime change	0 to 1
residentialRealEstateGrowth	residential real estate expansion, net of references to contraction	–1 to 1
residentialRealEstateSales	residential real estate sales rising, net of references to sales decreasing	–1 to 1
residentialRealEstateSentiment	positive references to residential real estate, net of negative references	–1 to 1
residentialRealEstateValues	residential real estate values rising, net of references to declining values	–1 to 1
sanctions	sanctions or embargoes emanating from or against a country	0 to 1
socialInequality	social inequality	0 to 1
socialUnrest	social unrest and calls for political change	0 to 1
tradeBalance	exports, net of references to imports	–1 to 1
Unemployment	unemployment rising, net of references to unemployment falling	–1 to 1

VISUAL VALIDATION

One simple technique for validating that the TRMI data reflect their intended output is to visualize actual events. Social unrest is one event with high psychological impact that has been in the news following the Arab Spring and other revolutions against totalitarianism. The SocialUnrest TRMI can be seen in Figure A.3, which demonstrates the general accuracy of the TRMI in tracking important global events where darker shading indicates higher levels of socialUnrest. TRMI for many Sub-Saharan African nations are not published in version 2.2, and their shading is light gray in the figure.

NOTES

1. P. Tetlock, "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance* 62(3) (2007).
2. Available to Thomson Reuters customers at: https://customers.reuters.com/a/support/paz/Default.aspx?pId=2381.

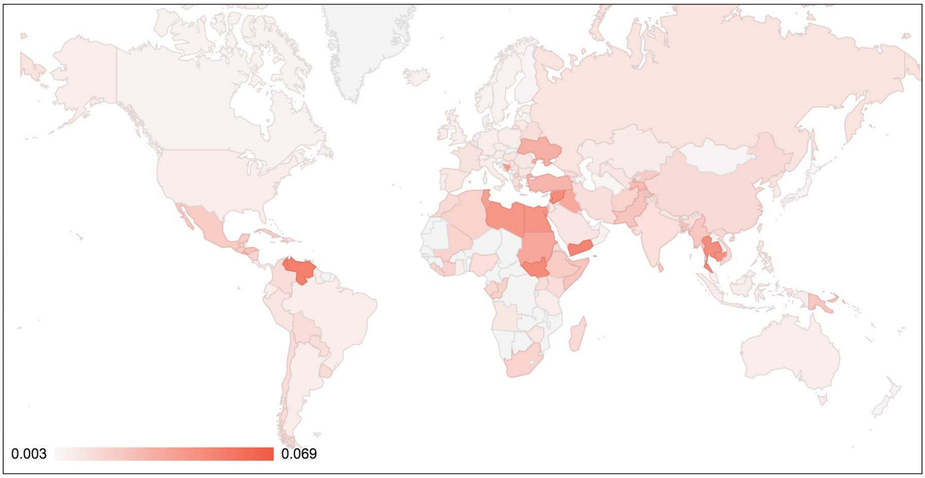

FIGURE A.3 An image of average SocialUnrest TRMI values for countries in the year 2014.