A Mathematical Approach to Identifying and Forecasting Shifts in the Mood of Social Media Users[1]

Dr. Les Servi and Dr. Sara Beth Elson
The MITRE Corporation
202 Burlington Road
Bedford, MA 01773
USA

lservi@mitre.org, selson@mitre.org

ABSTRACT

Social media offers a promising opportunity to identify and understand the direction of the mood of people using these platforms, just as conventional radars help one identify and understand physical motion. To date, many of the methods used to analyze social media in this way are qualitative, relying on the inputs of human subject matter experts. Those quantitative approaches which have been validated are in their infancy. This paper presents a new quantitative approach to characterizing the mood of social media users that can complement existing qualitative methods. This novel method combines a validated computer program (LIWC) with a mathematical algorithm to follow trends in past and present moods and detect breakpoints where those trends changed abruptly. First steps have also been taken to further develop this method so that it can also predict future trends in mood and possibly forecast related events. Validation is an important aspect of this part of the overall study. Finally, preliminary guidance for putting the output of the breakpoint analysis and forecasting into context is provided. The paper concludes with an overview of directions for continued research.

1.0 INTRODUCTION

Social media offer an important window into the emotions of those who use the platform. Media like Facebook and Twitter can reveal what people are thinking and feeling, often about very current events. The rapidly growing number of social media users around the world creates a wealth of data that complements traditional public-opinion polls and other types of surveys. Much work has been done in recent years to develop ways to analyze this rich new dataset. As articulated in [11], social media provides an opportunity to create a "Social Radar … to sense perceptions, attitudes, beliefs and behaviors." One area of focus has been on using social media to detect significant changes in public mood. A closely related goal has been to forecast future shifts in the mood of a country's citizens. Being able to identify and forecast these mood shifts with a high degree of accuracy is the essential first step toward interpreting what they mean politically and socially, and potentially anticipating future events to which they might lead.

To date, a number of the methods developed to tap into social media in these ways depend heavily on subject matter experts (SMEs) and are highly qualitative. While these approaches have value, there has been concern that some lack scientific rigor and objectivity. At MITRE, we developed a novel approach that answers this concern. Our method builds on a recognized tool designed to analyze the content of texts:

[1] The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

a computer program called "Linguistic Inquiry and Word Count 2007" (LIWC) [14][15]. In recent years, this program has been used successfully to probe the content of social media. Starting with the results that LIWC can generate from sizeable datasets of social media posts, the our approach employs a mathematical algorithm to detect changes in public mood. Points in time at which public mood shifts substantially are referred to as "breakpoints." [2] Our method computes and mathematically updates estimates of when these breakpoints occur, as well as the trends in public mood in between breakpoints. It can do this either in real time or historically. The selection of breakpoints is unbiased by global events or any person's subjective views of where the world is going. This freedom from any potential bias offers a decided advantage in comparison with SME-oriented approaches.

We began this work with a large existing dataset of posts to Twitter—called "tweets." As it developed its new method, we used this existing dataset as the basis for a test case. This was because the dynamics of public mood during the period examined were extraordinary, and the dataset was readily available. The test case both demonstrates the method and provides concrete examples of the type of outputs it produces.

While our method in its current form provides objective rigor in detecting breakpoints and monitoring trends in public mood in the past and present, it has also been designed to be able eventually to assist in forecasting the future direction of public mood. Work toward this end is fully underway. Forecasting, especially using social media, is very challenging and is still far from an exact science. We has made a good start at applying its mathematical approach to forecasting future LIWC indicators in a manner that brings desired rigor. In due course, the our method will be able to apply these types of forecasts to actual events.

Identifying, monitoring, and forecasting public mood are part of a larger process whose goal is to explain the meaning, in political and social terms, of the raw trends and shifts that one is observing. Without first accurately identifying when such changes occur, reliable interpretations are not possible. But the ultimate objective is well-informed explanations and, possibly, recommendations stemming from those explanations. We has already made a first foray in its work to date into developing a way to draw such political and social implications from the output of its algorithm. That work is ongoing.

1.1 Organization of the paper

Each section in the paper can be read independently of other sections to suit the interests of various readers. Section 2 provides a background to LIWC and its application to Twitter data. Section 3.1 describes a new mathematical approach to best characterize abrupt shifts in the emotions of Twitter users. We illustrate this approach first with a simple example and then with a Twitter data set. Section 3.2 then repeats the analysis of Section 3.1 in a mathematically precise manner. Expanding on the mathematical approach presented in Section 3, Section 4.1 presents a set of forecasting algorithms using the output of Section 3 and initial validation for one of those algorithms over the others. Section 4.2 then presents methodological considerations to address when using the techniques presented in this paper. Finally, Section 5 presents future directions for this research.

2.0 GENERATING THE DATA NEEDED TO APPLY THE MITRE METHOD: THE LINGUISTIC INQUIRY AND WORD COUNT (LIWC) PROGRAM

[2] The working premise here is that, in general, public mood evolves quite smoothly, but on occasion, it changes very suddenly and dramatically. When this happens, it's as if a "reset button" were being pushed. In our work, we treat points in time before the "reset' as irrelevant to points in time after it.

In order to examine the emotions expressed by Twitter users, we used the computerized text analysis program, LIWC. In essence, LIWC provided our means of converting the textual messages of tweets into quantitative data that we could analyze mathematically. When one uses LIWC to "process" texts, the program counts the total number of words contained in that text. It then counts the number of words falling into each of approximately 80 categories (such as emotion categories like "positive emotion," "sadness," social categories like "family," "friends," and other types of categories). For each of the roughly 80 categories, LIWC then computes a ratio of the number of words in the text falling into that category to the total number of words in the text. On both a daily and weekly basis, we combined all tweets posted in a day / week into a text file and then ran these files through LIWC. Doing so enabled us to generate a series of ratios across time – one for each day or week - for each category. It was these ratios that we used as the raw material on which to run the algorithms presented in this paper.

LIWC has a scientific pedigree, in that it has been validated across many studies. It processes texts by counting words in psychologically meaningful categories [17]. Researchers have used LIWC in a wide variety of experimental settings, such as to show attentional focus [16], emotionality [5], cognitive styles [12], individual differences [13], and social relationships [6].

In addition, the use of certain function words (such as "the," "with," or "they")[3] has been linked with personality and social processes, psychological states including depression, biological activity, reactions to individual life stressors, reactions to socially-shared stressors, deception, status, gender, age, and culture [2]. For example, people who are feeling physical or emotional pain tend to focus their attention on themselves and therefore use more first-person singular pronouns in their speech or writing while they are in that state [15].

In addition to function words, the use of emotion words (such as "happy," "terrified," or "jealous"), how people express emotion, and the valence of that emotion can tell us how people are experiencing the world [17]. People react in radically different ways to traumatic or salient events, and how they react may reveal much about how they cope with the event and the degree to which the event will play a role in the future. Previous research using LIWC suggests that it identifies emotion accurately in language use. For example, people use positive emotion words (e.g., "love," "nice," or "sweet") when writing about a positive event and negative emotion words (e.g., "hurt," "ugly," or "nasty") when writing about a negative event [7]. In addition, LIWC ratings of positive and negative emotions words in written texts correspond with human ratings of the emotional content of those same written texts [1]. As such, the function and emotion words people use provide psychological markers of their thought processes, emotional states, intentions, and motivations [17][4].

At the center of LIWC is a set of dictionaries – one each for approximately 80 categories of words [15]. A dictionary consists of the set of words that define a particular category. When processing text files (e.g., from blogs, novels, essays, poems, etc.), the program first processes each file word by word and compares each word with those contained in the dictionary files; it then counts the number that are in each dictionary file and calculates a ratio of the number of words in each dictionary to the total number of words in the text file, e.g., 2% of all words in a text file are sadness words.

To develop the emotion word categories, human judges had to evaluate which words were appropriate for each category [15]. For all of the subjective categories, Pennebaker et al. first selected word candidates,

[3] Function words include the following categories: pronouns, propositions, articles, conjunctions, and auxiliary verbs [2]. These words have minimal lexical meaning yet they tie a sentence together. They frequently reveal psychological states which people are experiencing

[4] Of course, like all other automated programs of its nature, LIWC does not perfectly capture the meanings of all words.

and then groups of three judges independently rated whether each word candidate fit within the overall word category [15]. Based on the responses, a rigorous protocol was used to remove or add words. Next, the entire process was repeated by a separate group of three judges. The final percentages of judges' agreement for the second rating phase ranged from 93% to 100%.

In the past, researchers have applied LIWC to a variety of different text genres, including college writing samples, science articles, blogs, novels, talking, and newspapers. [5], [15] Although previous research has laid a foundation for the use of Twitter and of LIWC to forecast the direction of Twitter users' emotions and events on the ground, any such forecasting would rely on manual inspection of trends in LIWC indicators. For this reason, MITRE developed a set of mathematical algorithms to determine, objectively, when trends in emotion have shifted and to forecast where these trends are headed in the near future.

Figure 1 depicts raw tweet data. Each dot signifies a LIWC ratio for the swear indicator and the sadness indicator, computed based on an aggregate of an entire week's worth of tweets. In looking at the pattern of these ratios, it is all too easy to make a judgment call as to when Twitter users' emotions shifted up or down and where to draw the line between various phases of their emotions. The danger in doing so lies in the fact that different observers may disagree as to where, exactly, to draw such lines. Moreover, if observers take into account world events, then doing so may lead to them to draw the lines differently depending on their views of the impact these events had on public mood.

Because of this potential problem, we developed a mathematical approach that determines, objectively, when Twitter users' emotions have shifted (or are shifting, as events unfold in real time). We introduce this algorithm in Section 3.

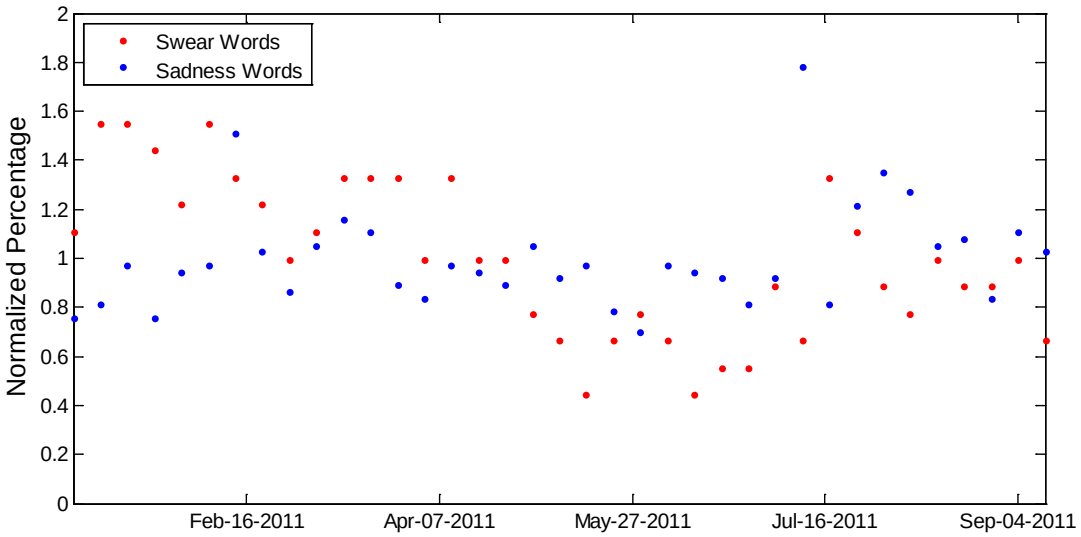

Figure 1: Normalized Percentages of LIWC Swears and Sadness Words in Weekly Aggregates of Tweets

3.0 THE NEW APPROACH: AN AUTOMATIC BREAKPOINT ANALYSIS

3.1 A Qualitative Description of the Approach

Using a simple example with a single small hypothetical dataset, we will illustrate this mathematically unbiased approach to determining when shifts occurred in a dataset. Consider the set of five data points in Fig. 2a, on the left:

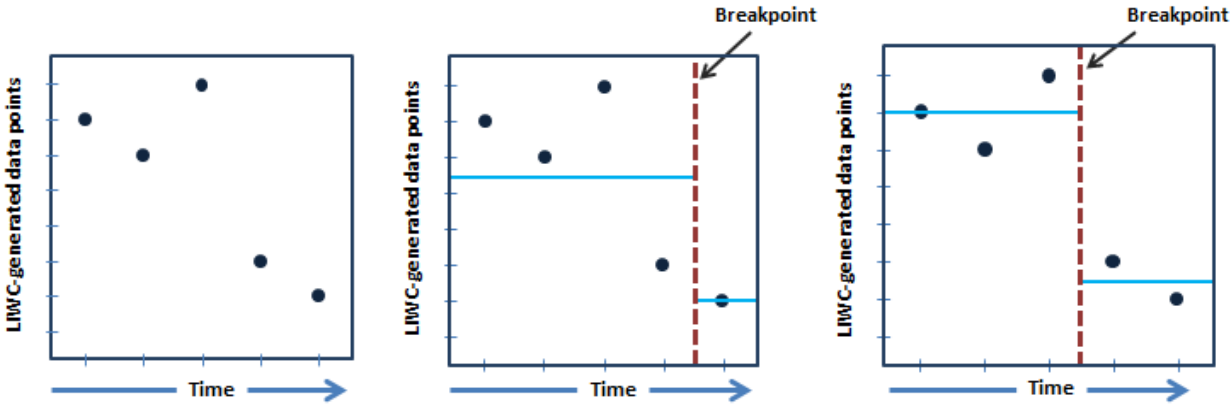

Figure 2a. Five Raw Data Points in a Hypothetical Set

Figure 2b. Alternative Breakpoint #1

Figure 2c. Alternative Breakpoint #2

Figure 2: Five Raw Data Points in a Hypothetical Set

Suppose the underlying dynamics of a system are such that it is reasonable to assume that these five data points represent noisy measurements of a system in one state which, some time while these data are being collected, discontinuously moves to another state (perhaps due to a world event). For example, if the discontinuity tasks place after the fourth point in time, the horizontal blue lines (which corresponds to the average value of the data during each time region) would correspond to the Fig 2b, in the middle. Alternatively, if the break point occurred between the third and fourth data point, the blue lines would similarly correspond to Fig. 2c, on the right:

Which breakpoint is more reasonable and why?

An unbiased approach to determining which selection of breakpoints is better would be to drop a perpendicular orange arrow from each data point to the blue line (corresponding to the average in each region), (Fig. 3a and Fig. 3b), and summing the sizes, the distances, of these orange arrows.

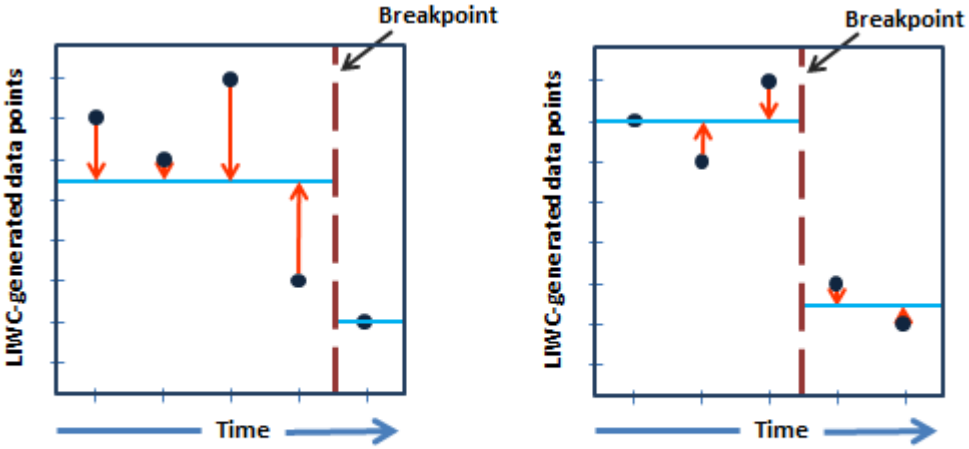

Figure 3a. Mathematically Suboptimal Breakpoint

Figure 3b. The Mathematically Best Breakpoint

Figure 3: Selecting the Mathematically Best Breakpoint

Clearly, the division on the right has a smaller summation of the orange arrows. Hence, one arrives at the intuitive conclusion that the breakpoint on the right is the better choice.

For the case of finding a single division in the small hypothetical raw data in Fig. 2a, the best division could have been found intuitively. However, for the raw Twitter data (Fig. 1), intuition would not be sufficient. That is, for large data sets, we need the following natural generalization. While the generalization clearly requires more computation to complete, it is conceptually no harder to find, and each step can be made mathematically quite rigorous.

(i) Rather than examine only two possible sets of breakpoints, the left and the right choices, examine *all* possibilities.
(ii) Rather than assume that the system has two regions, assume it may have more.
(iii) Rather than consider a single LIWC indicator consider multiple LIWC indicators.

Applying these first two assumptions to the raw weekly Twitter data of LIWC swears and sadness indicators[5], one can algorithmically determine (in a manner described in section 3.2) the **best** set of breakpoints for the data, These breakpoints are shown in Fig. 4.

[5] Combining the use of both the swear and sadness indicators yields more informative breakpoints than either indicator alone, so the rest of this paper uses both. Regardless of how many or which indicators one decides to use, one can algorithmically determine that the *best* set of breakpoints, as illustrated in Fig. 4. In this figure, the vertical axis is a normalization of the percentages (making each have a mean of 1) for clarity. Note that adding multiple data sets will typically change the dates of the best breakpoints.

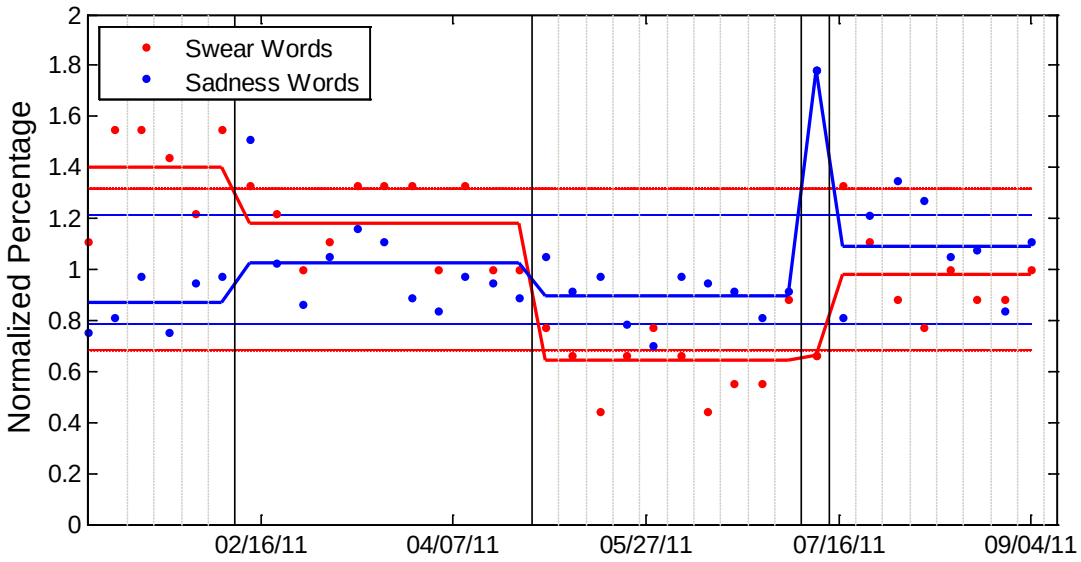

Figure 4: Breakpoint Analysis of Normalized Percentages of LIWC Swear and Sadness Words in Weekly Aggregates of Tweets, using a Constant Model

The above analysis is called a *constant model* because of the implicit assumption that the level of the LIWC emotions are a constant within a breakpoint region. This assumption can be relaxed to assume that the level of the LIWC emotions are a line within a breakpoint region. This will be referred to as the *linear model*. As indicated in the next section, the linear model can be analyzed in a manner analogous to the constant model. The result, applied to the same data is shown in Fig. 5.

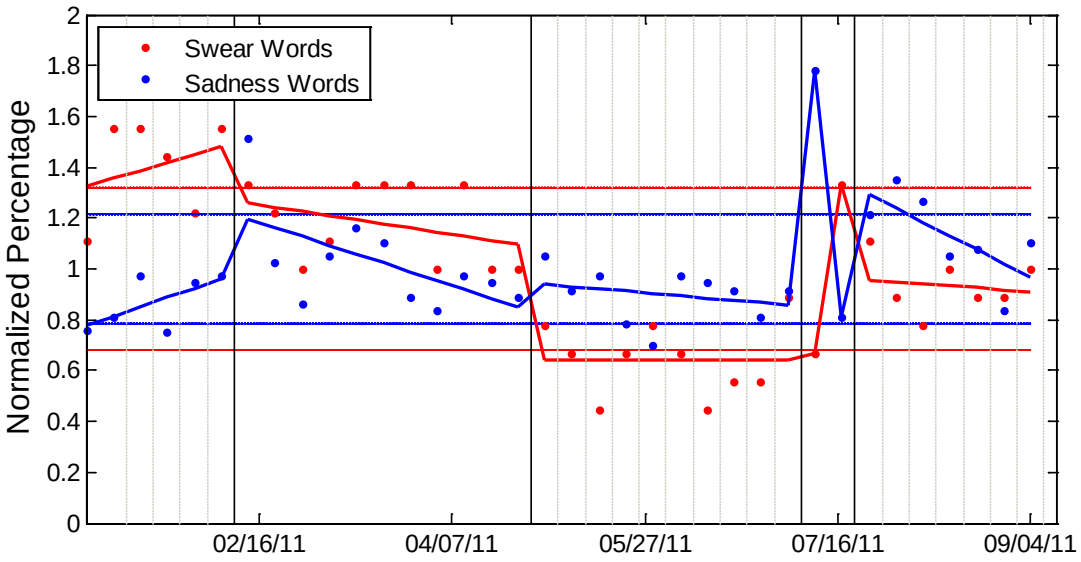

Figure 5: Breakpoint Analysis of Normalized Percentages of LIWC Swears and Sadness Words in Weekly Aggregates of Tweets, using a Linear Model

A similar analysis can be conducted using daily tweet data; this analysis is illustrated in Fig. 6.

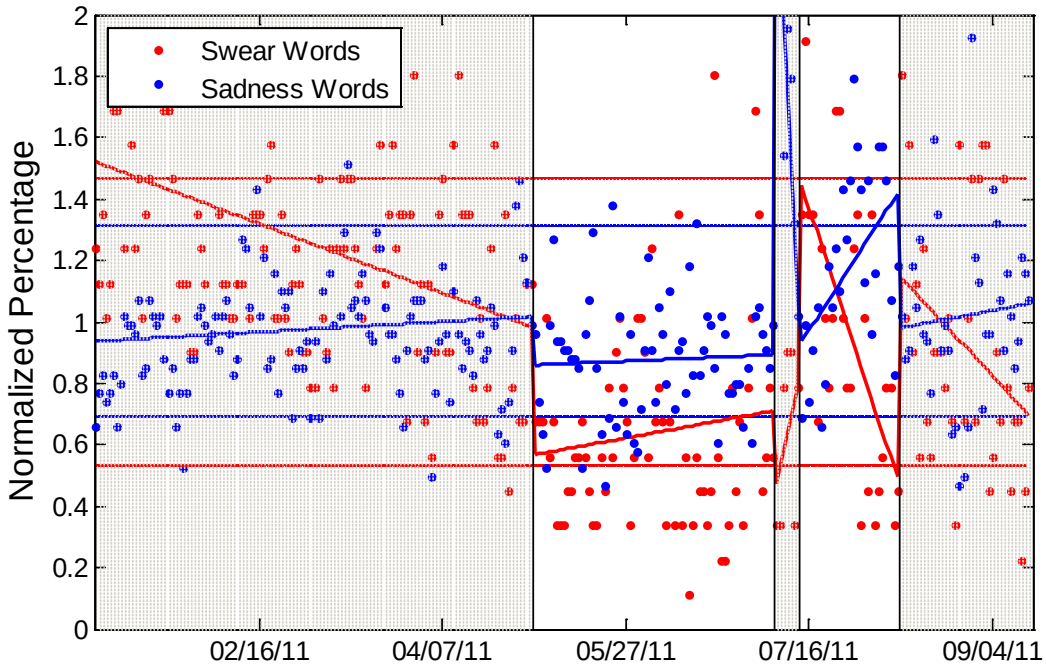

Figure 6: Breakpoint Analysis of Normalized Percentages of LIWC Swears and Sadness Words in Daily Aggregates of Tweets, using a Linear Model

While not used in this analysis, three additional generalizations that can be easily implemented would be

- Rather than assume that each data point has the same importance, assume that the importance may be different for each (perhaps reflecting differing confidences in each point). This differing confidence might reflect the number of tweets which were sampled to generate the data points.

- Rather than assume that each data set (derived, for example, from LIWC indicators such as sadness) has the same importance, assume that the importance may be different for each (perhaps reflecting different experiences with each set for different applications).

- Rather than processing the data after the fact, it might be useful to process the data in real time as they are being created and measured. Doing so enhances situational awareness.

As a final, important note, we wish to emphasize that as new data emerge over time, the algorithm re-calculates the breakpoints to determine whether they still provide the best set of phases to the ever-changing data. Therefore, over time, breakpoints can appear, disappear, and re-appear if the picture of trends changes with newly-acquired data. One should not, therefore, consider the breakpoints to be permanent markers of where a phase-shift occurred in public mood. Thinking about this idea at a more conceptual/ historical level, it often becomes clearer, with the passage of time, the points at which public opinion/ mood entered a new phase regarding some historical event. But it may not be as clear at the very point when public mood is shifting that it is, in fact, shifting.

3.2 A Mathematical Description of the Approach

While Section 3.1 describes the key ideas behind the breakpoint analysis algorithm, it is useful to revisit it in a mathematically more precise manner for the interested reader.

Suppose we have a set of K time series x_n^k for $n = 1,...,N$ and $k = 1,...,K$ with time, *n*, separated into M partitions with breakpoints at $m_0 (= 0), m_1, ..., m_{M-1}, m_M = N$. Suppose, within a breakpoint region, it is reasonable to assume the data arises either from a collection of piecewise constants, i.e., $x_n^k \approx y_j^k$, for $n = m_{j-1}+1,...,m_j$, a collection of piecewise linear functions, i.e., $x_n^k \approx y_j^k(1) + n y_j^k(2)$, for $n = m_{j-1}+1,...,m_j$, or more generally

$$x_n^k \approx A_j(\vec{m})\vec{y}_j^k \qquad \text{for j=1,…,M} \qquad (1)$$

where

$$\vec{x}_j^k(\vec{m}) = (x_{m_{j-1}+1}^k, x_{m_j+2}^k, ..., x_{m_j}^k) \qquad (2)$$

$$A_j(\vec{m}) = \begin{pmatrix} 1 & m_j+1 & (m_j+1)^2 & ... & (m_j+1)^{D-1} \\ ... & ... & ... & ... & ... \\ 1 & m_{j+1} & (m_{j+1})^2 & ... & (m_{j+1})^{D-1} \end{pmatrix} \qquad (3)$$

(where D = 1 and D=2 are the cases of most interest) and $\vec{y}_j^k = (y_j^k(1), y_j^k(2), ..., y_j^k(D-1))^T$ is an unknown vector to be determined.

The *best* breakpoints are found from solving

$$m^*(\vec{m}) = \arg\min_{\vec{m}} \left[\sum_{j=1}^{M} \sum_{k} w_k \min_{\vec{y}_j^k} \| \vec{x}_j^k(\vec{m}) - A_j(\vec{m})\vec{y}_j^k \|_P^2 \right] \qquad (4)$$

Equation (4) can be interpretated as searching the space of all possible breakpoints, $\vec{m}$, for the one which minimizes the distance between each datapoint of each data set and the fitted estimate from equation (1) (which might be a constant, straightline, or another function) while the distances are weighted by the importance of the data point, (indicated by P), as well as the importance of the data set (indicated by $\vec{w}$). P might be set based on the number of tweets which generated the data point (as well as being diminished as one looks back in time. $\vec{w}$ might be set based on the seeming relevance of each data set on the issues of most interest.

The inner minimization is a simple linear regression, the best $\vec{y}_j^k$ in equation (4) is

$$\vec{y}_j^{k*} = \left(A_j^T(\vec{m}) P A_j(\vec{m}) \right)^{-1} A_j^T(\vec{m}) P \vec{x}_j^k \qquad (5)$$

4.0 INTERPRETING THE BREAKPOINT ANALYSIS AND FORECASTING

In Section 4, we introduce a mathematical algorithm to forecast, in the short term, the moods of Twitter users. We designed this algorithm for use by anyone interested in generating forecasts of moods. However, Section 4 also presents a more technical comparison (in Section 4.1) of the algorithm's predictive accuracy against that of two other algorithms – a baseline algorithm and an algorithm presented for comparison purposes. In the process, Section 4.1 presents initial validation work for the most desired forecast algorithm for the data available. Readers wishing only to see the forecast algorithm itself may wish to skim Section 4.1 and read Section 4.2 in more depth, as Section 4.2 presents an initial methodology for using the algorithm.

4.1 Forecasting: A Mathematical Forecasting Rule and Discussion of its Accuracy Compared with Other Rules

In Section 3, we introduced the idea of mathematically calculating the points at which public mood shifted into a new phase. One can use this technique to understand how Twitter users' emotions changed in the present and recent past as events unfolded. This technique also opens the door to forecasting where those emotions are headed in the near future. Hence, this section introduces a mathematical extension of the breakpoint technique (in the form of a "forecasting rule") and presents a discussion of how that forecasting rule was selected over other potential forecasting rules. Specifically, this section presents a comparison of three potential forecasting rules and the accuracy with which they actually forecasted short-term trends in Twitter users' emotions.

An ultimate goal of those using this approach might be to use the output of the breakpoint analysis to forecast actions that will occur on the ground, such as riots or protests. However, forecasting such actions using LIWC indicators would require, as a first step, making accurate forecasts regarding the future trends of the indicators themselves. Further research is needed to determine the extent to which LIWC indicators can be used to forecast actions – as opposed to just the emotions of Twitter users.

Forecasting, of course, is fraught with obstacles. Perhaps the most prominent is the possibility of self-negating prophesies where the actions of the public or those stimulating the mood (e.g., those proposing a new tax policy) negate the prophesy[6][7]. Therefore, rigorous testing is crucial when forecasting.

Below, we illustrate and present the three mathematical approaches ("Rules") to calculate a forecasted value (i.e., the ratio) of a LIWC indicator based on previous values of that indicator:

[6] This phenomenon is closely related to the 'paradox of warning' introduced and discussed in [9].

[7] In another domain, research on stock market prices has shown that the self-negating prophesy phenomenon hinders making forecasts of those prices according to the efficient market theory hypothesis [4]. Specifically, an anticipated future price change may spur traders to take immediate action, causing the change to take place in the present and, therefore, making the future unpredictable.

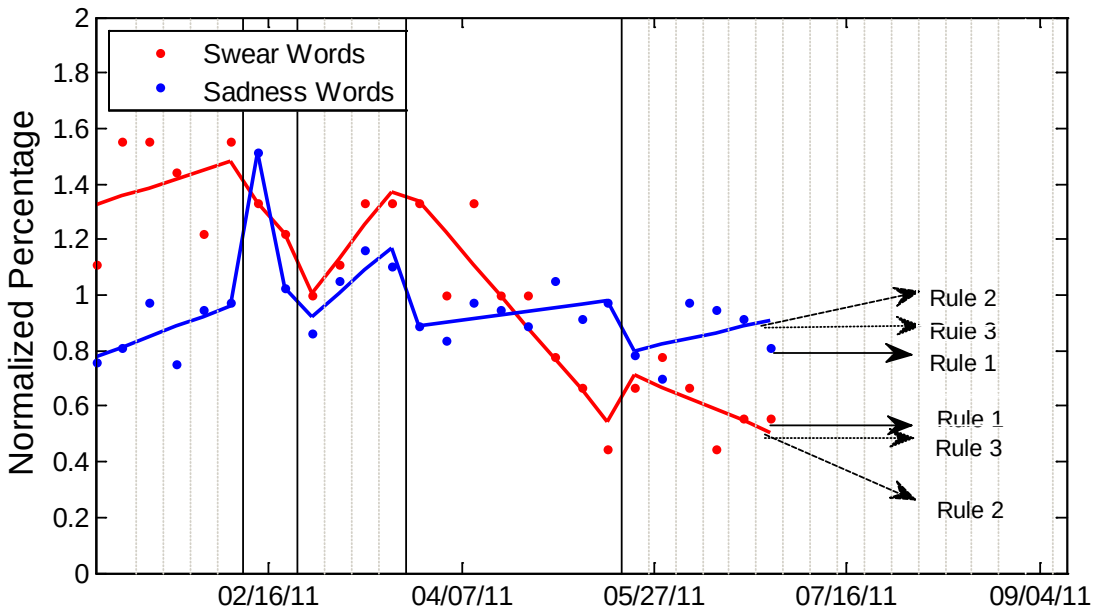

Figure 7: An Illustration of the Three Rules

Rule 1: *Use the last dot*: Forecast the value of a LIWC indicator *without using the breakpoint analysis* by assuming that the last data point is the best estimate for the future. If this rule outperforms the others, it will mean that the breakpoint analysis is not necessary. As such, this might be considered the baseline forecast rule.

Rule 2: *Project ahead:* Perform a breakpoint analysis with the linear model and use this analysis to forecast by projecting ahead in time, starting from the most recent breakpoint. That is, extend the most recently derived trend line out longer (continuing along its current slope) and use that line to determine the forecasted value. This is the most natural approach. However, as indicated below, we expect it will demonstrate a compartive advantage only if there is an extensive amount of data in each individual breakpoint region and if the trend lines have large slopes.

Rule 3: Use last estimate: Perform a breakpoint analysis with the linear model and use this analysis to forecast that future values will be the same as the last estimated data point, i.e., extend, in time, the most recently derived trend line out from its end but with a *horizontal* slope – rather than continuing along the current slope of the line. While this may be less accurate and less intuitive than Rule 2 if the data are plentiful, this is a more conservative and robust rule which will perform well for a broader range of data availability and slopes of the trend lines.

Below, we illustrate the accuracy of the three rules aggregated over *each* instant in time[8] using two LIWC indicators (swears and sadness) and examining the forecasted values generated at one, two, and three weeks into the future. We compared each set of predicted values[9] with the actual values that emerged and computed the amount by which our forecasts were off (i.e., the sum of squared errors).

[8] More precisely, the test was performed on each instant in time after at least 100 daily data points were processed. This warm-up period is needed to ensure the breakpoint analysis has reasonable accuracy.

[9] To facillitate comparisons between the swear and sadness predictions, all values were normalized so that the average value of both in the time region of interest was set to 1. In general, the more granular the data, the more accurate the forecasts will be. Our

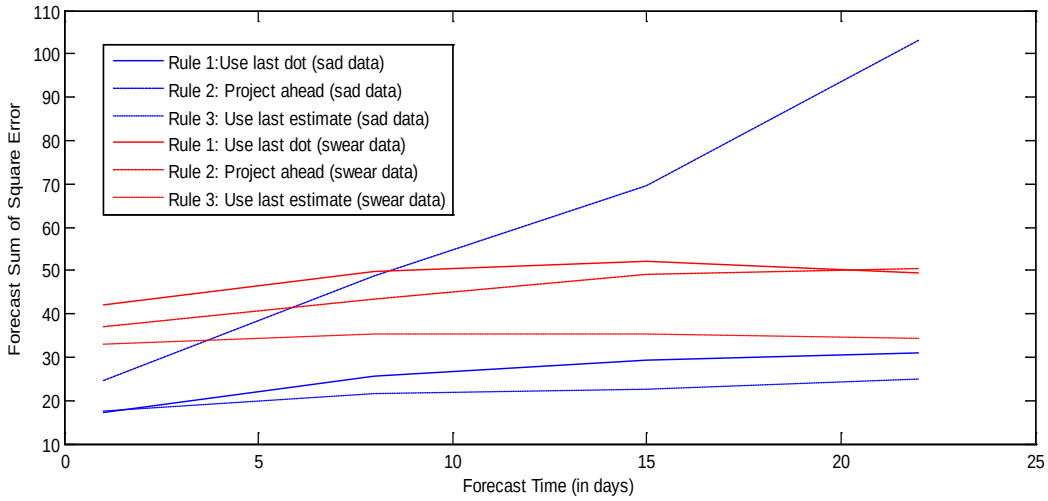

Figure 8: Sum of Square Error in Predicting Normalized Sadness and Swear Daily Data

The key finding depicted on this figure is that for this data, Rule 3 consistently outperformed the baseline Rule 1 (which was generated without the breakpoint analysis) as well as Rule 2[10]. In addition, Rule 2 outperformed the baseline Rule 1 when applied against the normalized swear data but fell short when applied against the normalized sadness data.

We performed the above analysis with limited data. As such, to make more definitive predictions and explore more subtle prediction rules would require a more exhaustive study with orders of magnitude more data.

4.2 Putting the Output of the Breakpoint Algorithm in Context

It should be clear that, by examining trend lines and breakpoints, one can gain situational awareness of Twitter users' emotions that is more illuminating and subtle than examining raw data alone (such as in Fig. 1). Yet, the trend lines and breakpoints must be put into a broader context in order to fully understand emotional patterns within Twitter. Certain questions are essential to address throughout the process of using the trend lines and breakpoints, such as: 1) are we applying them to an "extraordinary" period of emotional trends or to an "ordinary" one?, 2) how do we determine when a trend is really going "up" or "down"?, 3) what does it mean if a breakpoint disappears shortly after it appears?, and 4) how do we know that a trend line is really signifying the emotion we think it is?

---

analysis was performed on a *daily* version of the data described in previous sections of this paper and not weekly data - as our preliminary studies indicate that daily data offer the minimal granularity needed for reasonably accurate forecasts.

[10] The estimated slope of the trend line generated by the linear model is an imperfect estimate. As such, in cases such as the one explored in this section, the precision of Rule 2 might be insufficient to outperform the more conservative Rule 3. However, one would expect Rule 2 to perform better as the number of data points in the breakpoint region increases and/or the slope of the trend line increases. Note, however, that there may be situations where Rule 3 outperforms Rule 2; yet Rule 2 (the linear model) can accurately predict the *sign* of the slope of the trend line (i.e., the direction which the public mood is moving).

When conducting an analysis using trend lines, breakpoints, and the forecast rule, we recommend, firstly, taking a step back to examine the larger picture of public mood. Examining this picture entails asking questions like "Has public mood truly entered an "extraordinary" situation compared with what it looked like before, during more typical times? Are the ups and downs I'm seeing in the LIWC indicators (e.g., sadness, swears) really that big, in the overall scheme of Twitter users' emotions, or are they relatively small?" One way to answer these questions consists of comparing the possible extraordinary trends in LIWC indicators with "baseline" trends, collected before the supposed special period began. In some cases, the "special" period may be as much as two orders of magnitude greater than the baseline period with regard to average LIWC ratios and /or the variation in these ratios (i.e., the standard deviation). Further empirical study is needed to make recommendations regarding when LIWC values are elevated enough to warrant detailed breakpoint analysis. However, the higher the level of elevation, the more relevant the resulting interpretations are.

Another important consideration when examining trendlines is the fact that LIWC, like other automated content analysis programs, ignores context, irony, sarcasm, and the use of metaphors [2]. In that sense, although studies are yielding evidence that function words indicate emotional and biological states, status, honesty, and several individual differences, the imprecise measurement of word meaning and psychological states implies that researchers will not be able to detect those states with 100% accuracy by examining words automatically.

Another consideration in applying the breakpoint analysis is that breakpoints may shift as new data emerge, as discussed in Section 3. This possibility highlights the need to gain confidence that a breakpoint (signifying a shift in Twitter users' emotions) is robust - meaning that their emotions have truly entered a new phase – before trying to forecast where their emotions are going. By the same logic, if a breakpoint disappears quickly, there may be flux in Twitter users' emotions, making it hard to forecast where those emotions are headed. One could gain confidence that a breakpoint is robust by monitoring the breakpoint over time, and as the days continue, one's confidence increases. At this time, we advise initial caution when a breakpoint first appears before concluding that Twitter users' emotions have shifted into a new phase.

At this point, it is worth noting how valuable it can be to observe flux in the breakpoints, which signifies flux in the underlying Twitter users' emotions. The value of observing this flux is that one knows that the emotions have not settled yet. One also knows to temper one's certainty about the future.

While monitoring breakpoints to determine when Twitter users' emotions have shifted, one can also learn the directional change in emotions by examining whether trend lines are rising/falling and how sharply those rises/falls are (i.e., the slopes). Further research will need to determine, mathematically, how sharp a rise/fall must be for one to consider it significant. Naturally, as a trend line angles more sharply upward/downward, one gains confidence that mood is changing quickly.

5 CONCLUSION AND FUTURE DIRECTIONS

By combining the content analysis program, LIWC, with a new mathematical approach to detect shifts in Twitter users' emotions, the work presented in this paper offers an opportunity to use the data obtained from social media to gain objectively-derived insights into the dynamics of Twitter users' emotions. In sum, the MITRE method can yield situational awareness in two ways:

- Identify abrupt shifts in Twitter users' emotions that occurred in the recent past, and pinpoint when they took place[11]
- Draw trend lines that indicate the direction in which present Twitter users' emotions are changing in real time.

The work presented here has also laid the groundwork for forecasting mathematically what Twitter users' emotions will look like in the near future, and, ultimately, interpreting the meaning of changes with a high degree of reliability.

Our approach to date constitutes a first step; yet more research is needed in this area. In the future, it would be possible to pursue in a mathematical, evidence-based manner three trajectories of work that would secure even greater accuracy and enable the method to achieve broader impact:

Expand the diversity of the input used in an analysis. In this paper, we focused on two LIWC word categories—sadness and anger— to characterize certain dynamics of public mood. Yet, this was a test case, and a much greater breadth of word categories is available. Using a more diverse set of categories will produce a more complete representation of how Twitter users' emotions are evolving. Consequently, future work should explore, mathematically, the most informative combination of LIWC indicators for assessing each different social media dataset. Another way of gaining greater reliability and a more robust understanding of the dynamics of Twitter users' emotions is to utilize datasets drawn from a variety of social media platforms rather than just one—for example, blogs and Facebook in addition to Twitter. Finally, future studies that compare and then combine the input of subject matter experts with our mathematically derived analysis could yield fused insights that are very likely to surpass the results of either approach used alone.

Detect any biases in the dataset being used in an analysis, and correct for them. The work we present in this paper rests on the implicit assumption that the tweets contained in the dataset used in our test case represent an unbiased sample of Twitter users' emotions. In reality, statistical biases can arise from less-than-representative demographics within a sample of social-media users, varying levels of influence among different groups within a sample, and varying levels of passion of the people using social media. A next stage of work, therefore, should develop methods to identify and correct for these and other potential biases found in social media datasets.

Utilize larger data sets. Larger datasets than the one we used in our test case are certainly possible. Use of larger datasets would deepen the understanding that can be achieved of changes in public mood, because they would enable us to see subtleties that statistically, are not visible in smaller datasets. In more technical terms, larger datasets would permit us to explore more subtle rules for forecasting and the role played by self-negating prophesies in the accuracy of forecasting. They would also permit us to characterize the robustness of the dynamics of public mood mathematically.

With the work already completed and the next steps laid out here, MITRE is establishing a strong, scientifically-based approach that will serve as part of a social radar [11][3][10]. This new sensor can be utilized alongside other sensors now in place and complement the contributions of subject matter experts.

Acknowledgments: The authors thank Frank Chang, Paul Garvey, and Paul Lehner for their thoughtful comments regarding this paper. The authors also thank Scot Lunsford for his many long hours spent developing tools to facilliate this work.

[11] A special feature of the method is that it can accommodate new data that emerges in real time and re-compute its best guess of (a) the trend lines of past emotions and (b) when breakpoints occurred. In this way, the method generates maximally accurate computations of previous trends in Twitter users' emotions.

7.0 REFERENCES

[1] Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., et al. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior, 21,* 361-376.

[2] Chung, C.K. & Pennebaker, J.W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp 343-359). New York: Psychology Press.

[3] Costa, B., Boiney, J. A., and Schmorrow, D. (2012). Social media analysis with social radar. NATO Human Factors and Medicine Panel HFM-201-Specialists Meeting on "Social Media: Risks and Opportunities in Military Applications."

[4] Fama, Eugene (1970). Efficient Capital Markets: A review of theory and empirical work. *Journal of Finance* **25** (2): 383–417.

[5] Gortner, E. M., & Pennebaker, J. W. (2003). The archival anatomy of a disaster: Media coverage and community-wide health effects of the Texas A&M bonfire tragedy. *Journal of Social and Clinical Psychology, 22,* 580-603.

[6] Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., & Pennebaker, J.W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science*.

[7] Kahn, J.H., Tobin, R.M., Massey, A.E., & Anderson, J.A. (2007). Measuring emotional expression with Linguistic Inquiry and Word Count. *The American Journal of Psychology, 120*, 263-286.

[8] Kalata, PR, The Tracking Index: A Generalized Parameter for alpha-beta and alpha-beta-gamma Target Trackers, *IEEE Transactions on Aerospace electronic Systems*, Vol AES-20, 1984, pp. 174-182.

[9] Lehner, P and A. Michelson. (Dec, 2010). "Measuring the Forecast Accuracy of Intelligence Products", submitted to Studies in Intelligences.

[10] Mathieu, Jennifer, Fulk, Michael, Lorber, Martha, Klein, Gary, Costa, Barry and Schmorrow, Dylan (2012). Social radar workflows, dashboards, and environments. NATO Human Factors and Medicine Panel HFM-201-Specialists Meeting on Social Media: Risks and Opportunities in Military Applications.

[11] Maybury, M. (2010) "Social Radar for Smart Power." The MITRE Corporation. www.mitre.org/work/tech_papers/2010/20_0745/10 0745.pdf

[12] Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology, 84*, 857-870.

[13] Newman, M.L., Groom, C.J., Handelman, L.D., & Pennebaker, J.W. (in press). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*.

[14] Pennebaker, J.W., Booth, R.E., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC2007 – Operator's manual. Austin, TX: LIWC.net.

[15] Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net

[16] Rude, S.S., Gortner, E.M., & Pennebaker, J.W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotio*n, 18, 1121-1133.

[17] Tausczik, Y., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24-54.