On #agony and #ecstasy: Potential and pitfalls of linguistic sentiment analysis

Seth Flaxman (flaxman@stats.ox.ac.uk) and Karim Kassam (kassamk@steelers.nfl.com)

Abstract

With the ready availability of social media data, researchers are increasingly undertaking large-scale studies of online emotion. Many of these studies employ sentiment analysis—automatically inferring emotional information from text written on blogs, status updates, or tweets. We compare the momentary, daily, and day-of-week patterns of affect data extracted from Twitter to affect data generated by directly polling a demographically representative sample. We highlight striking inconsistencies, casting doubt on the direct application of sentiment analysis tools to measure population-level well-being. Whereas sentiment analysis tools appear to capture negative affect reasonably accurately, the same tools produce estimates of positive affect that are uncorrelated with direct measurement, because the frequency of positive words is not a reliable indicator of positive affect. As a proof of concept, we use a simple feature selection algorithm to propose a new lexicon of words to enable accurate inference about population-level positive emotion as well as negative emotion.

Introduction

The burgeoning field of computational social science promises new scientific insights from the combination of big data and computational data analysis. A case in point has been large-scale studies of emotion and well-being. Using sentiment analysis to automatically infer emotional information from text written on blogs, status updates, and tweets, researchers have produced maps of happiness by US state (Mitchell, Frank, and Harris 2013), proposed indices to capture well-being over time (Kramer 2010), and made novel scientific claims about the psychology of emotion (Golder and Macy 2011). This body of research is based on the premise that automated sentiment analysis can be used to reliably measure population-level emotion. This premise depends on an untested assumption that any bias inherent in automated sentiment analysis tools is random rather than systematic, and therefore not an issue when sample sizes are very large.

We explore this assumption and investigate the reliability of sentiment analysis for the large-scale study of emotion by comparing patterns of affect extracted from Twitter to patterns extracted from a separate, demographically representative sample of subjects who were directly polled about their feelings. In examining the inconsistencies between the two datasets, our contribution serves to underscore the importance of combining computational methods and social science. We treat negative and positive sentiment separately (Watson et al. 1999) and find that automated analyses are much more accurate at capturing negative than positive emotion experience. We confirm this finding through a novel application of the bootstrap in a big data natural language setting. We conclude by creating a new lexicon of words for more accurate inference about population-level emotion through the use of a simple feature selection algorithm.

Measuring Well-Being

Emotions are multifaceted processes involving numerous response channels (Larsen and Fredrickson 1999), but the gold standard for measuring emotional experiences is simply to ask people about them. Verbal self-reports represent the most common, and possibly the best way to measure how people feel (Barrett 1996; Robinson and Clore 2002). They are non-invasive, inexpensive and reliable (Courvoisier et al. 2010), and correlate with a variety of behaviors (Frey and Stutzer 2002). Moreover, self-report is the only "ground truth" available—it is not possible to measure emotional experience in any other way, unless that method has itself been validated against self-report (Gilbert 2006). To be sure, self-report is not perfect, and a variety of studies have demonstrated circumstances under which self-reports and other measures of emotion deviate (e.g. when participants do not want to disclose taboo attitudes as those surrounding race, Greenwald, McGhee, & Schwartz, 1998). These are exceptions, however. In general, verbal self-reports provide statistically valid data that is generalizable to larger populations (Watson, Clark, and Tellegen 1988).

The simplest measure of emotional experience one can

obtain is of valence, the net positivity or negativity experienced. Even when researchers assess emotion with other methods, factor analyses typically yield valence as the largest underlying dimension (Feldman 1995). When assessed directly, valence can be measured either with a single bipolar scale (ranging from negative to positive), or with two unipolar scales (one ranging from neutral to negative and one from neutral to positive). The latter approach allows for states in which people feel both positive and negative simultaneously (Kron et al. 2013), i.e. to differentiate between ambivalence and agnosticism.

When positive and negative states are assessed separately they tend to be negatively correlated, i.e. the more positive one feels, the less negative he/she feels (Tellegen, Watson, and Clark 1999). They are not completely symmetric, however, and a variety of effects or behaviors that correlate with one measure bear little relationship to the other. Positive and negative emotions are also unequal in their typical magnitudes – negative stimuli tend to have greater impact (Baumeister et al. 2001). Negative information is paid more attention (Eastwood, Smilek, and Merikle 2001; Olofsson et al. 2008), and is weighed more heavily in decision making (Kahneman and Tversky 1984). Negative events have a larger initial impact on moods (David et al. 1997), and that impact tends to last longer (Baumeister et al. 2001).

Bad events also wear off more slowly than good events. Whereas one bad day predicts another bad day will follow, a good day does not predict anything about the following day (Baumeister et al. 2001). In affect regulation, Baumeister et al (2001) summarize previous research that shows people try much harder to decrease negative affect than they do to increase positive affect. In the social support literature, negative aspects of social networks (conflicts, upsetting interactions) were found to be correlated with a variety of negative measures of well-being and mental health, while positive aspects and interactions were not correlated with well-being (Baumeister et al. 2001). Thus, negative emotion is more powerful by a variety of measures.

Despite the greater power of negative events to shape well-being, people are more likely to share positive content over the internet (Berger and Milkman 2012). Sharing positive messages reflects positively on the sender, and self-presentation and identity communication are important motivators (Wojnicki and Godes 2008). These presentational concerns, combined with asymmetric emotion magnitudes, may cause differences in the reliability with which positive and negative emotions can be inferred using automated methods.

Automatic Measurement of Well-Being at Scale

Automated measurement of well-being typically involves taking a large text corpus, with geographic or temporal variation, and applying sentiment analysis tools to code that text. Average sentiment scores can then be calculated over time or geographic area (e.g., Golder & Macy, 2011). Sentiment analysis holds the alluring possibility of scale – averaging across millions of individuals should produce reliable estimates and reveal hidden truths at the population level using freely available social media data.

The most widely used tool for automated textual analysis in psychology studies, Linguistic Inquiry and Word Count (LIWC; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), was developed to count word frequencies in psychologically-relevant categories. LIWC is an efficient alternative to relying on human raters, who are slow and do not always achieve high inter-rater reliability. In order to create LIWC, words were first compiled by hand by consulting dictionaries and other lexicons. Then, judges rated whether the words belonged in each category (Tausczik and Pennebaker 2010). Finally, categories were refined and shown to new judges for testing. The positive emotion category, for example, contains words like love, warm, and brave. The negative emotion category contains words like neglect, sad, and weapon. LIWC has been widely used to investigate a variety of psychological phenomena. To pick just one of hundreds of examples (a compendium can be found in Tausczik and Pennebaker 2010), Heberlein et al. (2003) used LIWC's positive and negative emotion categories to assess the effects of damage to right-hemisphere brain structures on emotional judgment. Experimental evaluations provide support for the use of LIWC as a sentiment analysis tool: Gonçalves et al. (2013) compared a variety of automated sentiment analysis tools on social media messages which had been hand-coded by human raters as positive or negative. LIWC had the highest average agreement with the other automated methods. To be sure, more sophisticated methods exist, e.g. VADER (Hutto and Gilbert 2014). We focus on LIWC due to its already widespread use, especially in psychology studies, and its high agreement with other methods.

For LIWC and other automated sentiment analysis tools to be valid measures of emotional experience, however, they need to do more than correspond with hand-coded analyses of the same data – they need to correspond with actual emotional experience. At a high level, the validity of these methods rests on their ability to produce unbiased estimates—if errors are random, ample data are available with which to extract signal from noise. How reliably can we infer positive and negative emotional states from the use of positive and negative words in a tweet, status, or blog post? The challenge is to establish whether errors will be systematic or random.

Assessing Sentiment Analysis

One method of examining whether sentiment analysis can accurately assess emotional experience is to compare the results of sentiment analysis to population-level ground

truth based on self-report data. Temporal patterns in emotion offer relatively stable and generalizable patterns for comparison between various measures of emotion and populations (e.g. Clark, Watson, & Leeka, 1989; Stone et al., 2006). Because temporal patterns in emotion have both stable external/environmental drivers and correlates (e.g. work-week vs. weekend; Hasler, Mehl, Bootzin, & Vazire, 2008) and stable internal/physiological drivers (Boivin et al. 1997; Murray, Allen, and Trinder 2002), we should expect to find these patterns with all instruments designed to assess emotion.

One pattern ubiquitous in the emotion literature is the negative correlation between positive affect and negative affect assessed simultaneously. Positive affect and negative affect are sometimes posited to be independent variables (e.g. Cacioppo & Berntson, 1994), but in practice, correlations are typically negative and significant. Russell and Carroll (1999) review 31 data sets and found correlations between positive and negative affect between -0.25 to -0.75. The more positive one feels in a given moment, the less negative he or she will feel, and vice versa. In contrast, analysis of positive affect and negative affect via text reveals almost no relationship with r = -0.10 (see results below) and r = -0.08 (Golder and Macy 2011).

A second well established pattern is the diurnal trend in positive affect[1]. Typically, positive affect starts low and rises over the course of the day before falling again in the late evening (Clark, Watson, and Leeka 1989; Murray 2007; Stone et al. 2006; Watson et al. 1999; Wood and Magnello 1992). The resulting inverted "U" shape mirrors the circadian rhythm of average body temperature (Murray, Allen, and Trinder 2002; Thayer 1978; Watson et al. 1999). In contrast, textual analysis of sentiment reveals a "U"-shaped pattern of positive affect, starting high, decreasing to a long trough from noon to 5pm, then increasing until midnight (Golder & Macy, 2011, see also below).

Third, day of week effects – better moods on weekends than on weekdays, and better moods on Fridays than on other weekdays – are likewise well established (Stone, Schneider, and Harter 2012). Sentiment analysis of textual data has found day of week effects for positive emotions with Saturday the most positive day (Dodds et al. 2011). We assess the presence of these three patterns in online sentiment analyses below.

Our Contribution

We attempt to resolve the question of whether sentiment analysis is an effective tool for inferring population-level emotion by analyzing two datasets. One is a large, experience sampling survey in which subjects self-reported their emotions every half hour (while awake) for a period of ten days. Our dataset includes a larger sample than typical experience-sampling studies (3867 participants and 1,126,116 observations), and took place beyond the confines of a laboratory: subjects were provided with smartphones on which they made their self-reports during the course of their normal day-to-day activities. The second is a large sample of Twitter messages, which we analyze using sentiment analysis coding strategies to calculate temporal trends in emotion.

We extensively compare these two datasets to each other and to previous studies. Our results highlight striking inconsistencies, casting doubt on the direct application of sentiment analysis tools to measure population-level well-being. Whereas sentiment analysis tools appear to capture negative affect reasonably accurately, the same tools produce estimates of positive affect that are uncorrelated with direct measurement. We argue that this is due to the fact that the frequency of positive words is not a reliable indicator of underlying positive affect. As a proof of concept, we turn to the question of whether and how sentiment analysis could be improved to generate reliable population-level estimates. Inspired by the fact that LIWC's negative affect dictionary correlates well with direct emotion measurement, we investigate whether any other dictionary-based methods for sentiment analysis yield diurnal trends in line with direct self-report. Using simple statistical techniques for high-dimensional data we derive new, more reliable dictionaries for coding both positive and negative affect.

Methods

Data Collection

We obtained experience sampling (ES) data from a demographically representative market research survey in which paid participants carried iPhones with them for 10 days. Each half-hour while awake, subjects answered a series of questions about what they were doing, who they were with, where they were, and about their emotional state. The emotion questions consisted of a series of emotion words paired with cartoon faces depicting the corresponding facial expressions, and subjects checked off as many of the pairs as they wished. The positive emotion words were: confident, excited, happy, hopeful, loving, contented, interested, and the negative emotion words were: angry, bored, frustrated, overwhelmed, sad, worried, exhausted, lonely. Subjects also rated their mood and how relaxed they were in the previous 30 minutes, each on a scale from 1 to 5. All observations are date and time-stamped. Demographic information, including zip codes, was also provided about survey respondents.

[1] Trends in negative affect have been less conclusive, with some evidence suggesting that negative affect may not exhibit a regular diurnal trend (e.g. Hasler et al., 2008; Murray et al., 2002).

Coding Positive and Negative Emotion from Self Report Data

Throughout, we adopt the following simple method of coding the positive affect of a given self-report: count the number of positive emotions selected by the participant. Similarly, define negative affect as the number of negative emotions selected by the participant. These scores range between zero and eight.

Coding Positive and Negative Emotion from Social Media Data

Replicating a sentiment coding strategy used previously (Golder and Macy 2011) we apply the positive and negative emotion dictionaries in Linguistic Inquiry and Word Count (Tausczik and Pennebaker 2010) to a set of messages sent over Twitter. The dataset was obtained from the Decahose/Gardenhose Twitter stream using the sampling strategy and preprocessing pipeline described in Eisenstein et al. (2012) which was restricted to geotagged messages only, meaning that most messages consisted of short, casual conversations, sent by Twitter users from their mobile devices. We further restricted our sample to users within the United States, and converted all time stamps to local time. This sample is very different from the one considered in Golder and Macy (2011), so it serves as a robustness check on their results.

Each dictionary in LIWC consists of a list of words. To code a message sent over Twitter according to a given dictionary, we calculate the fraction of words in the message which appear in the dictionary. For example, the phrase "good morning" would score 0.5 on the positive emotion scale because the word good is in the positive emotion dictionary, and it comprises half of the words in the message. The phrase "crap, I lost my keys" would score 0.4 on the negative emotion scale because the words "crap" and "lost" are in LIWC's negative emotion dictionary.

Building a New Dictionary for Sentiment Analysis

Treating the diurnal trends derived from our experience sampling dataset as a "gold standard," we used simple statistical methods (greedy feature selection with a held-out training set) and the LIWC dictionaries to create a new dictionary of terms. Unlike sentiment-analysis based dictionaries, which are meant to evaluate the latent sentiment in a piece of text, our new dictionary is meant to accurately capture population-level emotion. After splitting our data for training and validation, we recoded the full set of tweets using our new dictionary.

Results

We calculated the correlation between negative affect and positive affect taking a single tweet or self-report as the unit of measurement. In the Twitter data, we found a correlation of -0.10, similar to a previous analyses of Twitter data (r = -0.08; Golder & Macy, 2010). In the ES data, we found the correlation to be -0.28, within the range established by theoretical and empirical research on the relationship between measurements of negative and positive affect (-0.25 to -0.75), and very close to the mean correlation or -0.33 found in dichotomous (yes/no) measures of affect (Russell and Carroll 1999).

Next, we calculated diurnal patterns in negative and positive emotion as measured in our experience sampling dataset (Figure 1), and compared them to patterns derived from sentiment analysis applied to Twitter data (Figure 2) . Throughout, we adjust our estimates for individual-level effects. For negative affect as shown in Figure 1 (right) and Figure 2 (right), the diurnal trends mirror one another, with a Pearson correlation of r(45) = 0.77 (95% CI: 0.62, 0.86), $p < 0.0001$. The patterns for positive affect shown in Figure (1) left and Figure 2 (left), however, are noticeably different with an insignificant negative correlation: r(45) = -0.11 (-0.38, 0.19), $p = 0.48$. Furthermore, the results from the ES dataset are in line with previous research (Clark, Watson, and Leeka 1989; Stone et al. 2006), suggesting that sentiment analysis on Twitter data is failing to capture experienced positive affect.

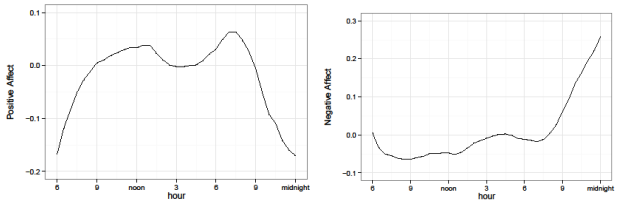

Figure 1. Diurnal trends in self-reported positive emotion (left) and negative emotion (right) from our Experience Sampling (ES) dataset.

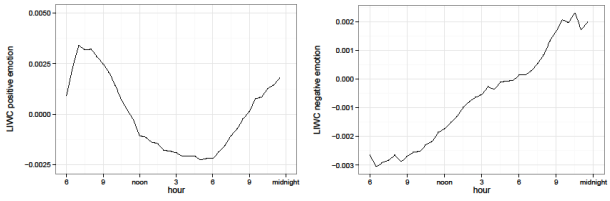

Figure 2. Diurnal trends derived from Twitter for positive emotion (left) and negative emotion (right) using Linguistic Inquiry Word Count, an automated sentiment analysis tool.

We turn to day of week (DOW) effects. In our ES dataset (Figure 3 (top)), there were significant DOW effects for both negative and positive affect (F's(6, 1126109) > 100; p's < 0.001). Effect sizes were modest; Fridays and weekends were less negative (d's > 0.01, p's < 0.002) and more positive (d's > 0.04, p's < 0.003) than weekdays, in line with previous research (Stone, Schneider, & Hart 2012).

In the Twitter dataset (Figure 3 (bottom)), DOW effects were vanishingly small for both negative and positive affect (d's < 0.002). The size of the Twitter dataset (104 million observations) allows for very small effects to reach conventional levels of statistical significance, but even with this massive sample the relationships that were found were inconsistent. Omnibus tests revealed significant DOW effects for negative affect ($F(6, 104704358) = 954$, $p < 0.0001$), and positive affect ($F(6,104704358) = 738$, $p < 0.0001$), but further tests revealed that while each day was significantly different for negative affect, the only day that was significantly different for positive affect was Sunday ($d = 0.01$; $p < 0.0001$), as shown in Figure 3 (bottom) where all effects are tiny but the significant effect of Sunday for positive affect is clear. (Note that the p-value for Saturday is 0.01 which we consider to be insignificant because of the huge sample sizes.) Fridays were less negative ($d = 0.01$, $p < 0.0001$) than weekdays but, in contrast to the ES dataset, not significantly different in terms of positive affect ($d = 0.0005$, $p = 0.06$). Puzzlingly, weekends were simultaneously *more* negative ($d = 0.01$; $p < 0.0001$) and *more* positive ($d = 0.01$; $p < 0.0001$) than weekdays. In sum, whereas day of week trends are consistent in experience sampling data, they are inconsistent or absent in our analysis of the Twitter dataset.

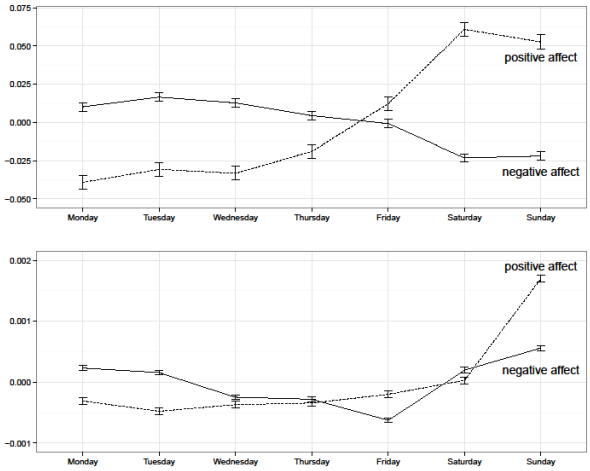

Figure 3. Day of week trends from experience sampling dataset (top) and Twitter dataset (bottom). Error bars show 95% confidence intervals around the mean.

To further understand differences in reliability of sentiment analysis for positive affect and negative affect, we calculated the correlation between the diurnal trend for each individual word from LIWC's negative emotion and positive emotion word lists with the diurnal patterns from the experience sampling dataset. For the negative emotion wordlist, 58% of words had a positive correlation with the negative affect trend from the ES data, with mean correlation 0.03 while for the positive emotion wordlist, only 43% of words had a positive correlation with the positive affect trend from the ES data, with mean correlation -0.04. There were individual words with reasonably high correlations, e.g. safe, happy, and loves had correlations above 0.4 for positive affect and savage, fatal, and skeptic had correlations above 0.4 for negative affect.

Ultimately, it is not these individual words that matter, but their aggregation into a dictionary. We used a bootstrapping approach to investigate random subsets of LIWC's positive and negative emotion dictionaries. We repeatedly subsampled 20% of the terms from the positive emotion dictionary, and for each subsample we applied this new dictionary of terms to the Twitter data, using the exact same methods as before, thus deriving a diurnal pattern. Then, we calculated the correlation between this diurnal pattern and the experience sampling trend in positive affect. We repeated this process 2000 times. We show the histogram for the correlations with the positive dictionary in Figure 4 (left). We repeated the same bootstrapping approach with LIWC's negative emotion dictionary, as shown in Figure 4 (right), calculating correlations with the ES trend in negative affect. The difference between the two plots is stark: while the smallest correlation we find after 2000 repetitions with the negative dictionary is 0.59, and the distribution is sharply peaked at a mean of 0.80, subsampling the positive dictionary results in a bimodal distribution, with a mode around -0.37 and another mode around 0.22. The largest correlation is 0.74, but correlations above 0.5 occur less than 4% of the time.

Figure 4 strongly suggests that whereas the frequency of negative words is a robust indicator for population-level negative affect, the frequency of positive words is, simply put, not an indicator of positive affect. While there are subsets of positive words which are positively correlated with positive affect, a larger fraction of subsets are *negatively* correlated with positive affect. This means that while positive words in written text may be a reliable indicator of the general positive sentiment of the text, it is a wholly unreliable indicator of population-level positive affect.

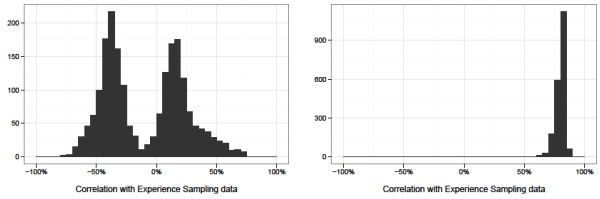

Figure 4. Random subsets of terms were drawn from the Linguistic Inquiry Word Count (LIWC) dictionaries for positive emotion (left) and negative emotion (right) and used as dictionaries to code diurnal positive and negative affect. The correlation was calculated between the daily trends and our Experience Sampling data for each subset. At left, the histogram of these correlations is plotted for the positive emotion terms, and a marked bimodal

pattern is evident, with some subsets of terms showing a weak positive correlation and other subsets showing a stronger negative correlation. By contrast, the figure at right shows correlations with the negative emotion terms. The correlations are much more robust and always positive.

Building a New Dictionary for Sentiment Analysis

We split our data into a training (80%), testing (10%), and validation (10%) set. We used the 64 separate psychologically-relevant dictionaries in LIWC to code each tweet. We used the experience sampling positive affect and negative affect diurnal trends as benchmarks in order to simultaneously create new positive and negative dictionaries, that met three criteria: 1) analysis using the positive affect dictionary should have a high correlation with the positive affect benchmark, 2) analysis using the negative affect dictionary should have a high correlation with the negative affect benchmark, 3) and the individual message-level correlation between the new dictionaries should be large in magnitude and negative. After adjusting for user-level effects, we used a greedy forward selection strategy to find a subset of dictionaries: on each step, alternating between the positive and negative dictionaries, we selected the LIWC category which when added to the previously selected categories maximized the correlation with the target dependent trend (criteria 1 and 2 above) and minimized the moment-to-moment correlation between the positive and negative dictionaries (criterion 3), where we weight these criteria equally by summing the correlations. This selection was performed on the training dataset, with the resulting correlations replicated on the validation dataset.

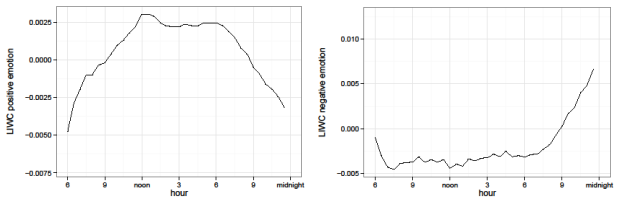

Figure 5. Diurnal trends derived from Twitter using a new dictionary of terms subselected with forward stepwise regression from Linguistic Inquiry and Word Count's positive emotion (left) and negative emotion (right) dictionaries.

For our final set of dictionaries, we chose the sets with the largest combined objective on the validation dataset. For the positive emotion words, the dictionaries were: Articles, Money, Family, Health, Numbers, Fillers, First Person Plural, Motion, Impersonal pronouns, Ingestion, Feel, Future tense, Third person plural, Religion, and Humans for a total of 1192 words. For the negative emotion words, the dictionaries were: Sexual, Negations, Sadness, Certainty, Second person, Nonfluencies, Affective processes, First person singular, Conjunctions, See, Inhibition, Body, Assent, Achievement, and Friends, for a total of 1658 words.

The result of recoding the entire dataset of Twitter messages using these dictionaries is shown in Figure 5. On the held-out test set, the correlation with the self-report data is (Figure 1) is r = 0.87 for positive affect and r = 0.87 for negative affect (compare to r = -0.11 and r = 0.77 for standard application of LIWC sentiment analysis on the same data). The moment-to-moment correlation between our positive and negative dictionaries is r = -0.28, in line with the result from the ES dataset and previous research. We also calculated DOW effects, obtaining results that were consistent with the ES dataset and previous research: a significant DOW effect for both negative (F(6, 104704356) = 1035, p < 0.0001) and positive affect (F(6, 104704356) = 1133, p < 0.0001), Fridays and weekends were less negative (p's < 2e-16) and more positive (p's < 0.0001) than weekdays with effect sizes between 0.005 and 0.02.

General Discussion

Our investigation into the reliability of sentiment analysis for measuring well-being shows both the promise and pitfalls of computational social science. By comparing the results of popular methods for measuring well-being using automated linguistic sentiment analysis to the results of a large, experience-sampling study based on self-report, we demonstrated striking inconsistencies in the temporal (momentary, diurnal, and weekly) patterns of positive, but not negative, emotion. Analyses of random subsets of the positive and negative sentiment analysis dictionaries showed both the robustness of the negative dictionary, and the inconsistency of the positive dictionary. Automated coding of positive emotion yielded temporal patterns that were negatively correlated with self-reported emotion.

The use of negative words appears more diagnostic of a negative state than the use of positive words is diagnostic of a positive state. A number of factors might result in this divergence: users may present themselves in certain ways on social media rather than share their experienced emotion (Wojnicki and Godes 2008); self-selection may interact with emotion and time, either promoting or inhibiting the sharing of certain content during the day; emotional experiences may be different during social media interactions, just as they are different during social interactions (Clark & Watson, 1988); social media may be used by an unrepresentative population that exhibits atypical diurnal patterns; and the strength of negative emotions may compel more accurate reporting (Baumeister et al. 2001).

Our findings are in line with Wang et al. (2012) who examined the validity of Facebook's "Gross National Happiness" (FGNH) by comparing temporal trends from

LIWC-inspired sentiment analysis of Facebook status updates to survey responses from a (non-random) sample of Facebook users who took Diener's Satisfaction with Life Scale (SWLS) survey on the same days. Depending on how they aggregated scores to calculate the correlations, they found non-significant correlations and correlations with incorrect signs. Similarly, Beasley and Mason (2015) compare the average sentiment of a year of a user's postings on Facebook and Twitter to a self-report of general emotional state and found very low, and at times insignificant correlations. These results all suggest strongly that sentiment analysis, as it is currently applied, does not yield valid population-level results.

Impact of sentiment analyses

Recent papers relying on automated sentiment analysis have generated much popular attention. But these findings deserve deep scrutiny in light of the lack of validation studies. Recently, much controversy surrounded the publication of an experimental study on emotional contagion in which the Facebook news feeds of users were manipulated to show fewer LIWC-coded positive or negative messages (Kramer, Guillory, and Hancock 2014). Our findings point at the untested assumptions underlying these methods, due to the unreliability of LIWC's positive emotion dictionary. The fact that removing LIWC-coded positive messages from a user's stream led users to post fewer LIWC-coded positive messages suggests that the content of messages posted by a user is influenced by the content of messages that user sees. However, we question the degree to which this relationship revolves around positive or negative emotion.

Self-presentation

That people attempt to present favorable impressions is well established in the psychological literature (Baumeister and Jones 1978; Goffman 1959). Though users of online social media strive to project accurate representations (Back et al. 2010), there is little doubt that some users curate content to promote positive evaluations at least some of the time (DiMicco and Millen 2007; Manago et al. 2008). If this is the case, it is little surprise that analysis of shared sentiment does not accurately reflect experienced emotion. Resharing positive stories or messages requires a minimum of effort on social networks (one or two clicks). Facebook in particular may promote this presentational bias by including a "like" (positive affirmation) button, and no corresponding "dislike" button. Writing messages with positive emotional content might be less cognitively demanding than writing messages with negative emotional content, perhaps because negative emotions result in more cognitive processing than positive emotions, as well as receiving more attention (Baumeister et al. 2001). Recent online evidence shows that positive news stories are more likely to be shared than negative stories (Berger and Milkman 2012), which could serve to further muddy the relationship between positive words and positive felt emotion.

Conclusion

Measuring population-level well-being is a challenging task, not one for which sentiment analysis tools like LIWC were designed, and not one for which the validity of LIWC has been formally tested. As with a variety of other sentiment analysis methods, LIWC focuses on the task of measuring the emotional content of a piece of text, and not on the emotional state of the person who wrote that text. Given the large psychological literature on presentational concerns and self-enhancement this subtle distinction becomes crucial to valid measurement. Our results strongly question the use of positive word frequencies as an indicator of positive affect, and thus we expect more sophisticated sentiment analysis methods to run into the same issues as LIWC.

Nevertheless, the challenge of measuring population-level well-being may be achievable, especially if the focus is placed more on negative than positive emotion. Negative emotion is ultimately a more robust, durable, and reliable indicator than positive emotion. As a proof of concept, we demonstrated using simple statistical methods that LIWC dictionaries can be used to obtain temporal trends which match those derived from our experience-sampling dataset. This demonstration suggests that it is possible to apply automated sentiment analysis methods for measuring well-being.

Resolving the issues we have raised will require further research, and the collection of new rich data sets that include both self-report data and concurrently written text. Our analysis represents the first step along this path.

References

Back, Mitja D, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. 2010. "Facebook profiles reflect actual personality, not self-idealization." *Psychological science* 21(3): 372–4. http://www.ncbi.nlm.nih.gov/pubmed/20424071 (July 18, 2014).

Barrett, Lisa Feldman. 1996. "Hedonic Tone, Perceived Arousal, and Item Desirability: Three Components of Self-reported Mood." *Cognition & Emotion* 10(1): 47–68. http://dx.doi.org/10.1080/026999396380385 (July 20, 2014).

Baumeister, Roy F., Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. "Bad is stronger than good." *Review of General Psychology* 5(4): 323–370. http://doi.apa.org/getdoi.cfm?doi=10.1037/1089-2680.5.4.323 (May 27, 2014).

Baumeister, Roy F., and Edward E. Jones. 1978. "When self-presentation is constrained by the target's knowledge: Consistency and compensation." *Journal of Personality and Social Psychology* 36(6): 608–618. http://psycnet.apa.org/journals/psp/36/6/608.html (July 20, 2014).

Beasley, Asaf, and Winter Mason. 2015. “Emotional States vs. Emotional Words in Social Media.” In *Proceedings of the 2015 ACM conference on Web science - WebSci ’15*,.

Berger, Jonah, and Katherine L Milkman. 2012. “What Makes Online Content Viral?” *Journal of Marketing Research* 49(2): 192–205. http://journals.ama.org/doi/abs/10.1509/jmr.10.0353 (July 12, 2014).

Boivin, Diane B., Charles A Czeisler, Derk-jan Dijk, Jeanne F Duffy, Simon Folkard, David S Minors, Peter Totterdell, and James M Waterhouse. 1997. “Complex Interaction of the Sleep-Wake Cycle and Circadian Phase Modulates Mood in Healthy Subjects.” *Archives of General Psychiatry* 54(2): 145. http://archpsyc.jamanetwork.com.ezp-prod1.hul.harvard.edu/article.aspx?articleid=497749 (July 22, 2014).

Cacioppo, John T., and Gary G. Berntson. 1994. “Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates.” *Psychological Bulletin* 115(3): 401–423. http://www.mendeley.com/catalog/relationship-between-attitudes-evaluative-space-critical-review-emphasis-separability-positive-negat/ (September 12, 2014).

Clark, Lee A., and David Watson. 1988. “Mood and the mundane: Relations between daily life events and self-reported mood.” *Journal of Personality and Social Psychology* 54(2): 296–308. http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.54.2.296 (September 27, 2014).

Clark, Lee Anna, David Watson, and Jay Leeka. 1989. “Diurnal variation in the Positive Affects.” *Motivation and Emotion* 13(3): 205–234. http://www.springerlink.com/index/10.1007/BF00995536.

Courvoisier, Delphine S, Michael Eid, Tanja Lischetzke, and Walter H Schreiber. 2010. “Psychometric properties of a computerized mobile phone method for assessing mood in daily life.” *Emotion (Washington, D.C.)* 10(1): 115–24. http://psycnet.apa.org/journals/emo/10/1/115.html (September 27, 2014).

David, James P, Peter J Green, René Martin, and Jerry Suls. 1997. “Differential roles of neuroticism, extraversion, and event desirability for mood in daily life: An integrative model of top-down and bottom-up influences.” *Journal of Personality and Social Psychology* 73(1): 149 – 159.

DiMicco, Joan Morris, and David R. Millen. 2007. “Identity management: multiple presentations of self in facebook.” In *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP ’07*, New York, New York, USA: ACM Press, p. 383. http://dl.acm.org/citation.cfm?id=1316624.1316682 (July 20, 2014).

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M Kloumann, Catherine a Bliss, and Christopher M Danforth. 2011. “Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter.” *PloS one* 6(12): e26752. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3233600&tool=pmcentrez&rendertype=abstract (November 3, 2012).

Eastwood, John D., Daniel Smilek, and Philip M. Merikle. 2001. “Differential attentional guidance by unattended faces expressing positive and negative emotion.” *Perception & Psychophysics* 63(6): 1004–1013. http://www.springerlink.com/index/10.3758/BF03194519 (September 27, 2014).

Eisenstein, Jacob, Brendan O’Connor, Noah A. Smith, and Eric P. Xing. 2012. *Mapping the geographical diffusion of new words*. . Computation and Language; Physics and Society. http://arxiv.org/abs/1210.5268 (July 17, 2014).

Feldman, LA A. 1995. “Variations in the Circumplex Structure of Mood.” *Personality and Social Psychology Bulletin* 21(8): 806–817. http://psp.sagepub.com/content/21/8/806.short (September 13, 2014).

Frey, Bruno S, and Alois Stutzer. 2002. “What can economists learn from happiness research?” *Journal of economic literature* 40(2): 402 – 435.

Gilbert, Daniel. 2006. *Stumbling on happiness*. New York: A.A. Knopf.

Goffman, Erving. 1959. *The presentation of self in everyday life*. Garden City N.Y.: Doubleday.

Golder, Scott A, and Michael W Macy. 2011. “Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures.” *Science (New York, N.Y.)* 333(6051): 1878–81. http://www.ncbi.nlm.nih.gov/pubmed/21960633 (November 11, 2012).

Gonçalves, Pollyanna, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013. “Comparing and combining sentiment analysis methods.” *Proceedings of the first ACM conference on Online social networks - COSN ’13*: 27–38. http://dl.acm.org/citation.cfm?doid=2512938.2512951.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. “Measuring individual differences in implicit cognition: The implicit association test.” *Journal of Personality and Social Psychology* 74(6): 1464–1480. http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.74.6.1464 (September 11, 2014).

Hasler, Brant P., Matthias R. Mehl, Richard R. Bootzin, and Simine Vazire. 2008. “Preliminary evidence of diurnal rhythms in everyday behaviors associated with positive affect.” *Journal of Research in Personality* 42(6): 1537–1546. http://linkinghub.elsevier.com/retrieve/pii/S0092656608001013 (November 23, 2012).

Heberlein, Andrea S., Ralph Adolphs, James W. Pennebaker, and Daniel Tranel. 2003. “Effects of Damage to Right-Hemisphere Brain Structures on Spontaneous Emotional and Social Judgments.” *Political Psychology* 24(4): 705–726. http://doi.wiley.com/10.1046/j.1467-9221.2003.00348.x.

Hutto, C J, and Eric Gilbert. 2014. “VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.” *Eighth International AAAI Conference on Weblogs and Social Media*: 216–225.

Kahneman, Daniel, and Amos Tversky. 1984. “Choices, values, and frames.” *American Psychologist* 39(4): 341 – 350.

Kramer, Adam D I. 2010. “An unobtrusive behavioral model of ‘gross national happiness.’” *Proceedings of the 28th international conference on Human factors in computing systems - CHI ’10*: 287. http://portal.acm.org/citation.cfm?doid=1753326.1753369.

Kramer, Adam D I, Jamie E Guillory, and Jeffrey T Hancock. 2014. “Experimental evidence of massive-scale emotional contagion through social networks.” *Proceedings of the National Academy of Sciences of the United States of America* 111(24): 8788–90.

http://www.pnas.org/content/111/24/8788.full?tab=author-info (July 9, 2014).

Kron, Assaf, Ariel Goldstein, Daniel Hyuk-Joon Lee, Katherine Gardhouse, and Adam Keith Anderson. 2013. "How are you feeling? Revisiting the quantification of emotional qualia." *Psychological science* 24(8): 1503–11. http://www.ncbi.nlm.nih.gov/pubmed/23824581 (June 3, 2014).

Larsen, Randy J., and Barbara L. Fredrickson. 1999. "Measurement issues in emotion research." In *Well-Being: Foundations of Hedonic Psychology*, eds. Daniel Kahneman, Edward Diener, and Norbert Schwarz. New York: Russell Sage Foundation, p. 40–60.

Manago, Adriana M., Michael B. Graham, Patricia M. Greenfield, and Goldie Salimkhan. 2008. "Self-presentation and gender on MySpace." *Journal of Applied Developmental Psychology* 29(6): 446–458. http://www.sciencedirect.com/science/article/pii/S0193397308000749 (July 20, 2014).

Mitchell, Lewis, MR Frank, and KD Harris. 2013. "The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place." *PloS one* 8(5): 1–20. http://dx.plos.org/10.1371/journal.pone.0064417 (July 4, 2014).

Murray, Greg. 2007. "Diurnal mood variation in depression: a signal of disturbed circadian function?" *Journal of affective disorders* 102(1-3): 47–53. http://www.sciencedirect.com/science/article/pii/S0165032706005301 (July 22, 2014).

Murray, Greg, Nicholas B Allen, and John Trinder. 2002. "Mood and the Circadian System: Investigation of a Circadian Component in Positive Affect." 19(6): 1151–1169.

Olofsson, Jonas K, Steven Nordin, Henrique Sequeira, and John Polich. 2008. "Affective picture processing: an integrative review of ERP findings." *Biological psychology* 77(3): 247–65. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2443061&tool=pmcentrez&rendertype=abstract (July 10, 2014).

Pennebaker, James W, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. *The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin*. Austin, TX.

Robinson, Michael D., and Gerald L. Clore. 2002. "Belief and feeling: Evidence for an accessibility model of emotional self-report." *Psychological Bulletin* 128(6): 934–960. http://psycnet.apa.org/journals/bul/128/6/934.html (September 27, 2014).

Russell, J a, and J M Carroll. 1999. "On the bipolarity of positive and negative affect." *Psychological bulletin* 125(1): 3–30. http://www.ncbi.nlm.nih.gov/pubmed/9990843.

Stone, Arthur a, Joseph E Schwartz, David Schkade, Norbert Schwarz, Alan Krueger, and Daniel Kahneman. 2006. "A population approach to the study of emotion: diurnal rhythms of a working day examined with the Day Reconstruction Method." *Emotion (Washington, D.C.)* 6(1): 139–49. http://www.ncbi.nlm.nih.gov/pubmed/16637757 (November 23, 2012).

Stone, Arthur a., Stefan Schneider, and James K. Harter. 2012. "Day-of-week mood patterns in the United States: On the existence of 'Blue Monday', 'Thank God it's Friday' and weekend effects." *The Journal of Positive Psychology* 7(4): 306–314. http://www.tandfonline.com/doi/abs/10.1080/17439760.2012.691980 (July 4, 2014).

Tausczik, Y. R., and J. W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1): 24–54. http://jls.sagepub.com/cgi/doi/10.1177/0261927X09351676 (May 28, 2014).

Tellegen, Auke, David Watson, and Lee Anna Clark. 1999. "On the Dimensional and Hierarchical Structure of Affect." *Psychological Science* 10(4): 297–303. http://www.mendeley.com/catalog/dimensional-hierarchical-structure-affect/ (July 4, 2014).

Thayer, Robert E. 1978. "Toward a psychological theory of multidimensional activation (arousal)." *Motivation and Emotion* 2(1): 1–34. http://link.springer.com/10.1007/BF00992729 (July 24, 2014).

Wang, N., M. Kosinski, D. J. Stillwell, and J. Rust. 2012. "Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National Happiness Index." *Social Indicators Research* 115(1): 483–491. http://link.springer.com/10.1007/s11205-012-9996-9 (July 4, 2014).

Watson, David, Lee A. Clark, and Auke Tellegen. 1988. "Development and validation of brief measures of positive and negative affect: The PANAS scales." *Journal of Personality and Social Psychology* 54(6): 1063–1070. http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.54.6.1063 (September 1, 2014).

Watson, David, David Wiese, Jatin Vaidya, and Auke Tellegen. 1999. "The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence." *Journal of Personality and Social Psychology1* 76(5): 820–838.

Wojnicki, Andrea C., and David Godes. 2008. "Word-of-Mouth as Self-Enhancement." *SSRN Electronic Journal*. http://papers.ssrn.com/abstract=908999 (July 20, 2014).

Wood, C, and M E Magnello. 1992. "Diurnal changes in perceptions of energy and mood." *Journal of the Royal Society of Medicine* 85(4): 191–4. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1294720&tool=pmcentrez&rendertype=abstract (July 22, 2014).