An Unobtrusive Behavioral Model of "Gross National Happiness"

Adam D. I. Kramer
University of Oregon
Department of Psychology
Eugene, OR, USA
adik@uoregon.edu

ABSTRACT

I analyze the use of emotion words for approximately 100 million Facebook users since September of 2007. "Gross national happiness" is operationalized as a standardized difference between the use of positive and negative words, aggregated across days, and present a graph of this metric. I begin to validate this metric by showing that positive and negative word use in status updates covaries with self-reported satisfaction with life (convergent validity), and also note that the graph shows peaks and valleys on days that are culturally and emotionally significant (face validity). I discuss the development and computation of this metric, argue that this metric and graph serves as a representation of the overall emotional health of the nation, and discuss the importance of tracking such metrics.

Author Keywords

Psychology, quantitative methods, emotion, statistics, Facebook

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Measurement, Theory, Verification.

INTRODUCTION

Interest in the basic happiness or well-being of a person, group of people, or nation has grown over the past several decades, receiving a great deal of attention in the psychological literature. The most notable drive to measure this was undertaken by Ed Diener and his colleagues, starting in the mid 1980s, formalizing the notion of "subjective well-being," specifically the "satisfaction with life" (SWL) component, which has since come to represent the extent to which a person feels that their life is worthwhile or, in essence, "good" [2]. Diener and others have since championed use of subjective well-being as a comparable metric to socio-economic status or credit score, which can be used to classify individuals into categories or to represent people relative to each other in a broad (e.g., national) context.

Current methods of measuring GNH employ a self-report methodology [6,9]. The proponents of these metrics argue consistently and convincingly that self reports are appropriate for this context: Because the very construct is subjective, self-reports effectively have no "bias" due to misperception (unlike personality measures which may have some "error," for example when one's self-perceptions do not correspond to one's behavior). In other words, if I claim to be happy, who can argue that I'm not?

In parallel with the psychological study of happiness, research in communication and HCI has been growingly interested in "sentiment analysis." This research program is largely dedicated to the systematic or algorithmic extraction of a user's emotional state from text they produce naturally, such as in a blog post, tweet, or Facebook status update.

Though there are many methods of extracting text from natural language posts, I use the process described in [14], referred to as a "word count" procedure. In this approach, a set of words (in this case, the positive and negative emotion word categories, described and defined in [14]) is defined as having some psychological meaning (in this case, positive or negative emotion), such that a user or group of users who use more words from a certain category are higher in the psychological construct that the category is designed to measure. This top-down approach is useful for the study of known topics, as it allows for validation studies to cross word-use contexts and corpora: A more bottom-up approach, such as LSA, may allow discovery of positive terms not present in the LIWC corpus [12], but would require separate validation of the resulting model. The word-count approach, conversely, has been used extensively in the fields of HCI and psychology. For example, [7] showed that in short blog posts, users known to be angrier show higher incidence of LIWC negative emotion words, while more joyful authors use more LIWC positive emotion words; [8] showed that these word categories could be used to differentiate happy romantic couples from unhappy couples' instant message communications. Further, since I model word use in terms of its variability, ensuring that the words I use are correctly coded (which a top-down model provides) is more important than ensuring that I have counted every emotion

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

word (several of which would require a bottom-up model to discover).

Though these (and several other) approaches have provided consistent evidence that it is reasonable to use a word-categorization approach to extract sentiment from online word use, I present the first few steps towards taking this approach "to scale," and provide preliminary evidence for the validity of a daily national-level happiness index.

DATA SET

This approach to modeling GNH is based on Facebook status updates, of which there are over 40 million posted per day [5]. I then narrowed the focus to users from United States who had selected English as the language in which they preferred to view the website; all such users' updates were anonymously analyzed beginning September 7, 2007. Status updates are short-format (m=9 words per update) notes that are broadcast to some or all of the user's friends. These updates start with the user's name, and contain text provided by the user (e.g., "Joe Smith ____," where Joe could fill in the blank). I chose status updates on Facebook for several reasons:

- These updates are largely "undirected:" there is no specific target for the status update. This makes status updates a better choice than wall posts (which are always directed), tweets (updates posted on Twitter [17], which may or may not be directed), or blog entries, as I am interested in emotions of the poster, rather than the poster's relationships (c.f. [8]).
- The very name "status update" indicates that individuals should in fact be updating their friends with their status. In other words, this is a self-descriptive text modality, optimized and designed to elicit updates about the self, many of which contain emotional or affective content. This makes status updates a better choice than tweets, which are not explicitly updates about the self, but contain a wide variety of information [17].
- Status updates are a very lightweight source of data, containing few words. This allows the presence or absence of positive or negative emotional words to "stand out" more than they would in a longer-format post such as a blog post.

I note that the above claims regard status updates in general; I do not argue that every update fulfills these qualities, but rather that they do as a corpus. I also do not argue that an analysis of other Internet data, such as tweets or blog posts, would be inappropriate or biased: Rather, I chose status updates because I believe them to provide the strongest signal of the emotional well-being of the posters.

I used the Hive data warehousing infrastructure atop the Hadoop framework [18], an open-source system that allows massively parallel processing of custom database queries (i.e., counting and categorizing words for the status updates of millions of users over the course of two years), using the Text Analysis and Word Count (TAWC) program [10], as the LIWC software itself is ill-suited to the scale of millions of updates per day. TAWC counted the positive and negative emotion words (as defined in the LIWC dictionary) used in each update, as well as the total number of words. To protect users' privacy, data was piped directly from Hive into TAWC, so that no update was ever seen by a researcher during analysis. This resulted in a count of the number of positive words, the number of negative words, and the total number of words for every status update made by every individual who logged in from the United States and used American English as the language in which they preferred to view Facebook.

AGGREGATION

To control for "wordiness," I computed a positivity (percent of words that were positive) and negativity (ibid) score for each status update. As such, a status update of "I am happy today" would get a positivity rating of .25 (the word "happy" is positive, no others are) and negativity of 0, while an update of "yes yes YES YES YES" would receive a positivity rating of 1.0 (because the word "yes" is positive) and a negativity of 0, and an update of "Today was kinda good, kinda bad" would receive a positivity score of .17 because of the word "good" and a negativity score of .17 because of the word "bad."

These percentage scores, however, are not directly comparable, because the potential for positive and negative word use is not equivalent: The LIWC dictionaries define 506 words as "negative" but only 407 as "positive," indicating that meaning of "percent positive" is due in part to the use of the English language and perhaps the LIWC dictionaries. To account for this and generate a metric that is interpretable independent of language and dictionary, I used this formula:

$$GNH_d = \frac{\mu_{pd} - \mu_{p\bullet}}{\sigma_{p\bullet}} - \frac{\mu_{nd} - \mu_{n\bullet}}{\sigma_{n\bullet}}$$

where GNH_d represents GNH for a specific day. μ_{id} represents the percent of words that were positive (p) or negative (n) for a given day (d), averaged across every status update for every Facebook user with a United States address who viewed the site in English. $\mu_{i\bullet}$ and $\sigma_{i\bullet}$ indicate a "meta-average" or the average and sd of these daily averages, across all days analyzed.[1] Readers will note that this process was conducted separately for positive words and negative words: This allows positivity and negativity to be weighted equally in the final analysis. This is important to note, as users may opt to report only some emotional events: By standardizing positivity and negativity separately, I focus on variation in each emotional valence separately. In other words, even if people dramatically

[1] In fact, I used the inner 90% of days to compute this metric, so that especially happy or unhappy days did not affect the units used for standardizing.

underreport negative events in status updates, each day's *relative* negativity should still be informative. Similarly, even if negativity were completely unreported (which was not the case), I would still argue (as most emotion researchers do) that those who are feeling more negative will express less positivity. By then subtracting the standardized negativity score for a day from the standardized positivity score for a day, I effectively weight negativity and positivity equally, as equal representations of "happiness." The result indicates the difference between how "remarkably positive" the day is and how "remarkably negative" the day is, as positivity and negativity are not precisely opposites [1]. If a day is far more positive than usual (for example, on holidays; see below), the GNH score for that day will be higher (unless that day is also far more negative than usual).

RESULTS

The results of this graph can be viewed online at [11], via a Flash-driven Facebook app, showing positivity and negativity scores, as well as the GNH aggregate. This graph updates automatically every day with a two-day delay and provides the GNH score for every day; readers are encouraged to view the online application.

VALIDATION

I present two methods of validating the use of this metric to represent GNH: Convergent validity measured by showing that Facebook users' life satisfaction scores predict the positivity of their own personal status updates (i.e., validating the use of word counts and the process of using standardized differences between positive and negative word percentages), and face validity by examining the high and low points of the GNH graph.

To demonstrate convergent validity, I show that for individual status updates, this aggregation (standardized difference between percent of words that are positive from those that are negative) produces a variable that is related to a validated measure of well-being at the level of the individual. To compute the aggregate for an individual status update, I used the following formula:

$$H_u = \frac{p_u - \mu_{pd}}{\sigma_{pd}} - \frac{n_u - \mu_{nd}}{\sigma_{nd}}$$

where in this case I normalize the percent of positive and negative words in each status update relative to the average and standard deviation across all updates for that day. This allows removal of "day" effects as well as "language" and "dictionary" effects from each individual's post. This is desirable because one would only call a status update "remarkably positive" if it were positive for the day in question: For example, if everybody uses the word "happy" on Thanksgiving (i.e., to say "(name) wishes everyone a Happy Thanksgiving"), then an update with 20% positive words would be unremarkable. I attempted to predict this metric using Diener and colleagues' [3] SWL scale, which I collected from n_{users}=1,341 Facebook users who had at least three status updates. These users all had English listed as their primary language, and opted into the SWL web-based survey at some point during August of 2009 by following an advertisement posted on Facebook; they filled out the SWL questionnaire [3], and all of their status updates were analyzed in the same anonymous manner described above. I found that 90% of status updates analyzed in this manner fell between -1.7 and 1.8 (m=-0.04, sd=1.31). I then conducted a simple hierarchical linear model (HLM [16]), using the *nlme* packages's *lme* function for the R project for statistical computing [13,15]. This allowed prediction of the positivity of status updates (of which there were between 3 and 3,141 per user, m=243.8, total $n_{updates}$=347000) from SWL scores (of which there was one score per user), effectively asking the question, "Does knowing how satisfied one is with one's life predict how positive one's status updates are?" If so, this would indicate that this method of coding status updates represents a true measure of SWL, a component of happiness [19].

SWL was a significant predictor, b=0.05, $t(1339)$=6.27, $p < .001$, corresponding to a correlation of about r=.17: Those more satisfied with their life do indeed score higher on the metric, relative to other users, for a given day.

To demonstrate face validity, I examined the peaks and dips of the graph itself (see [11]): The graph provides a face-valid measure of national happiness for a given day if the graph is high on days when the nation is expected to be happy and low on days that the nation is expected to be unhappy. The graph does indeed show this pattern, with peaks occurring on national and cultural holidays (e.g., Christmas, Thanksgiving, Halloween, New Year's Day, Independence day, and others); I also note that Mother's Day was happier in 2009 than 2008, representing the rapid growth of older demographics and mothers on Facebook which took place between May 2008 and May 2009), and two marked dips on days of national tragedy: January 22, 2008, on which both an actor popular in America (Heath Ledger) died and the Asian stock market crashed, and June 25, 2009, the day that American cultural icon Michael Jackson died unexpectedly. I also note a short (7-day) cycle throughout the graph, corresponding to the common knowledge that Fridays are the best day of the week, a full 9.7% happier than the worst day of the week (Monday), $t(205)$=273.5, $p < .001$.

LIMITATIONS AND FUTURE DIRECTIONS

There are several limitations to the conclusion as well as the method. The first limitation is the question of demographics: Facebook's early adopters were primarily college students, though several reports suggest that the demographic composition of Facebook is broadening [4,5]. I also note that even if the demographic of individuals using Facebook is notably nonrepresentative of the national population, the graph will disproportionately represent the happiness of some citizens over others. Consistent disparities of happiness among demographic groups within

the same country, however, presents a larger issue at the level of measuring happiness at the national level at all.

The second limitation is the possibility that the choice of sentiment analysis method (word counts) may throw off the model on certain days. For example, the word "happy," a positive emotional word, is also used as a salutation for holidays (e.g., "Happy Thanksgiving"). I note first that the separate modeling of positivity and negativity provides a check on these data: Holidays also show a dip in negativity corresponding to the positive spikes on these holidays [11], which cannot be due to the fact that positive salutations may not indicate a positive state. I also note that wishing someone a happy holiday is itself a positive emotional act designed to make others feel good and to raise holiday cheer: I do not believe that the word in this context is in fact emotionally "blank," and so I did not actively seek to eliminate this word from the model.

Perhaps the most obvious future direction is to extend this work to model other countries' GNH in a manner that allows comparisons among countries. To this end, the process and system of computation have been developed in a manner that is independent of language and word corpus used; in this sense, the task of extending this project to other countries is the development of a corpus of positive and negative words for other languages or dialects.

The second future direction, currently underway, is to show the validity of the metric using other national-level metrics of happiness (e.g., the Gallup-Healthways Well-Being Index [6] and economic indicators of well-being).

The current utility of this graph, however, is to have a behavioral method with which to track the emotional health of the nation, both in terms of evaluating whether the population as a whole is in a positive or negative state. In brief, this work uses well-established HCI methods (word counts), taken in an unobtrusive manner (status updates were not provided for purposes of the study), creates an aggregate metric out of citizens' posts (happiness of a day), validates the metric (using the SWL scale), scales the metric to a national level (the US), and publishes it online (pushing it beyond "pure research" and making it a designed product itself).

ACKNOWLEDGMENTS

I thank Moira Burke, Danny Ferrante, Ravi Grover, Cameron Marlow, and the Facebook Data Team for their support on this project. This work was conducted during the author's internships at Facebook in the summers of 2008-9.

REFERENCES

1. Cacioppo, J. T., Gardner, W. L., & Bernston, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review, 1,* 3-25.
2. Diener, Diener, & Diener (1995). Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology, 69,* 851-864.
3. Diener, E., Emmons, R. A., Larson, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49,* 71-75.
4. Facebook Data Team. (2009). *How diverse is Facebook?* Retrieved from http://www.facebook.com/note.php?note_id=205925658858&id=8394258414&ref=mf
5. Facebook. (2009). *Facebook Press Statistics.* Retrieved Jan 5, 2009, from http://www.facebook.com/press/info.php?statistics
6. Gallup. (2008). *Well-being index*. http://www.well-beingindex.com.
7. Gill, A. J., French, R. M., Gergle, D., Oberlander, J. (2008). The language of emotion in short blog texts. *Proc. CSCW 2008,* 299-302.
8. Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in online communication. *Proc. CHI 2007,* 929-932.
9. Kahneman, D., Diener, E., & Schwarz, N. (Eds.). (2003). *Well-being: The foundations of hedonic psychology*. New York: Russell Sage Foundation.
10. Kramer A. D. I., Fussell, S. R., & Setlock, L. D. (2004). Text analysis as a tool for analyzing conversation in online support groups. *Proc. CHI 2004,* 1485-1488.
11. Kramer, A. D. I. & Grover, R. (2009). United States Gross National Happiness on Facebook. http://apps.facebook.com/usa_gnh/
12. Mishne, G. (2005). Experiments with mood classification in blog posts. *Proc. Style2005.*
13. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & The R Core Team (2009). *nlme: Linear and nonlinear mixed effects models.* R package version 3, 1-93.
14. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychological properties of LIWC2007*. Retrieved from http://www.liwc.net
15. R Development Core Team (2009). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
16. Raudenbusch, S. W. & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Newbury Park: Sage.
17. Twitter. http://www.twitter.com
18. Thusoo, A., and Facebook Data/Infrastructure (August 2009). Hive – A warehousing solution over a map-reduce framework. *Proc. 35th. Intl. Conf. on Very Large Data Bases*. Lyon, France.
19. Veenhoven, R. (2007). Measures of gross national happiness. *OECD: Statistics, Knowledge and Policy 2007: Measuring and fostering the progress of societies* (pp. 231-253). OECD Publishing: London.

The columns on the last page should be of approximately equal length.