Tracking "Gross Community Happiness" from Tweets

Daniele Quercia[§] **Jonathan Ellis**[‡] **Licia Capra**[‡] **Jon Crowcroft**[§]
[§]The Computer Laboratory, University of Cambridge, UK
[‡]Department of Computer Science, University College London, UK
dq209@cl.cam.ac.uk, j.ellis@cs.ucl.ac.uk, l.capra@cs.ucl.ac.uk, jac22@cl.cam.ac.uk

ABSTRACT
Policy makers are calling for new socio-economic measures that reflect subjective well-being, to complement traditional measures of material welfare as the Gross Domestic Product (GDP). Self-reporting has been found to be reasonably accurate in measuring one's well-being and conveniently tallies with sentiment expressed on social media (e.g., those satisfied with life use more positive than negative words in their Facebook status updates). Social media content can thus be used to track well-being of individuals. A question left unexplored is whether such content can be used to track well-being of entire physical communities as well. To this end, we consider Twitter users based in a variety of London census communities, and study the relationship between sentiment expressed in tweets and community socio-economic well-being. We find that the two are highly correlated: the higher the normalized sentiment score of a community's tweets, the higher the community's socio-economic well-being. This suggests that monitoring tweets is an effective way of tracking community well-being too.

Author Keywords
Psychology, emotion, community, Twitter

ACM Classification Keywords
H.5.m Information Interfaces and Presentation: Miscellaneous

INTRODUCTION

Policy makers have recently suggested that measuring community well-being will help governments do a better job at directing public policy towards promoting quality of life (happiness) rather than material welfare (GDP). The French president Nicolas Sarkozy recently announced he intended to include well-being in France's measurement of economic progress [13]. The UK prime minister David Cameron is initiating a series of policies, under the rubric "Big Society", that seek to make society stronger by getting more people running their own affairs locally all together. To assess how different "Big Society" policies will impact the well-being of communities, the UK Office of National Statistics will shortly be asked to produce measures of "general well-being" [13].

Measuring the well-being of single individuals can be easily accomplished by administering questionnaires such as the Satisfaction With Life (SWL) test, whose score effectively reflects the extent to which a person feels that his/her life is worthwhile [3]. Self-reporting has been shown to be reasonably accurate, and recent studies have further highlighted that it tallies with, for example, sentiment expressed in Facebook status updates [9].

To go beyond single individuals and measure the well-being of communities, one could administer SWL tests to community residents. But that would be costly and is thus done on limited population samples and at a frequency of years. Recent research findings suggest that the costing problem may be ameliorated by monitoring implicit data generated by community members. For example, in a 2010 Science article, Eagle *et al.* monitored the diversity of the communication networks formed by community residents, and showed that socio-economic well-being of communities strongly correlates with network diversity (with a striking correlation of $r = .74$). In particular, they built communication networks from phone records across the entire United Kingdom, cross-referenced it with socio-economic census data, and showed that members of well-off communities have diverse networks, while members of economically and socially disadvantaged communities have insular social relations [5].

Another way of tracking community well-being is to extract community residents' emotional state from text they produce on social media websites such as blogs, Twitter, or Facebook. As a result, "sentiment analysis", that is, extracting emotional state from text, is receiving a lot of attention from the research community [8]. Since self-reported content on social media is ready available, researchers can monitor sentiment without having to go through the time-consuming process of asking people explicit questions. Kramer, for example, has built a sentiment metric out of Facebook status updates (he used a standardized difference between percent of words that are positive and those that are negative), found that the metric correlates with self-reported satisfaction with life (it correlates with SWL scores with a statistically significant coefficient of about $r = .17$), and aggregated the metric at national US level. In so doing, he observed that the temporal graph of the aggregate US metric

To appear in the Proceedings of ACM CSCW 2012.

(named "Gross National Happiness") showed peaks occurring on national and cultural holidays [9].

Although Kramer's study suggests that one might be able to gauge a whole nation's well-being and overall emotional health from the sentiment expressed on social media, it is not clear whether the correspondence between sentiment of self-reported text and well-being would hold at *community level*, that is, whether sentiment expressed by community residents on social media reflects community socio-economic well-being. We test the hypothesized correspondence between sentiment and community well-being by making the following contributions:

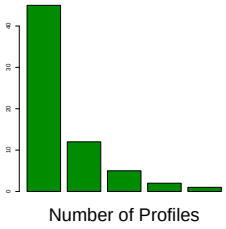
- For a variety of London communities, we crawl tweets produced by their residents and obtain census data of their socio-economic well-being (Section "Dataset"). More precisely, we crawl Twitter accounts whose user-specified locations report London neighborhoods. We geo-reference those accounts by converting their locations into longitude-latitude pairs. We also obtain the 2007 UK government's Index of Multiple Deprivation (IMD), which is a composite measure of relative prosperities of 32,482 communities, 78 of which are in London [11].

- We analyze the sentiment of tweets using two algorithms (Section "Sentiment Analysis"). The first is the word count technique proposed by Kramer and is based on computing the standardized difference between percent of words that are positive and those that are negative. The second is the Maximum Entropy classifier, which is a state-of-the-art method for classifying sentiment of single tweets [6]. We implement both classifiers and show that their results correlate and are accurate.

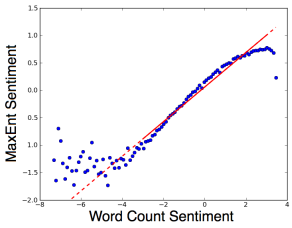
- We study the relationship between sentiment and socio-economic well-being (Section "Results"). We find that the higher the normalized sentiment score of a community's tweets, the higher the community's socio-economic well-being. The correlation coefficients are statistically significant and are as high as $r = .350$ for sentiment based on word count, and $r = .365$ for sentiment computed with Maximum Entropy. These results suggest that monitoring the sentiment of tweets may well be an accurate and cost-effective way of tracking the well-being of communities.

DATASET

Twitter Profiles. To control for any variability in the use of language across geographic areas, we have preferentially chosen Twitter profiles from London and did so as follows. We chose three popular London-based seed profiles of news outlets: the free subway newspaper Metro, the center-left newspaper The Independent, and the tabloid The Sun. These news outlets cover the entire UK political spectrum and have high penetration rates in the city. Each Twitter user who follows these seed profiles was crawled. This resulted in 250K profiles, 157K of which specified geographic locations (mostly city names) and 1,323 specified London neighborhoods (e.g., Hackney, Mayfair). For those 1,323 profiles, we converted user-specified home locations into longitude-latitude pairs using the Yahoo! PlaceMaker API[1]. To filter out profiles that are likely to be spam accounts or are not of real people, we crawled the PeerIndex realness score for each profile. This score is generated upon information such as whether the profile has been self-certified on the PeerIndex site and or has been linked to Facebook or LinkedIn. "PeerIndex realness score is a metric that indicates the likelihood that the profile is of a real person, rather than a spambot or twitter feed. A score above 50 means this account is of a real person, a score below 50 means it is less likely to be a real person"[2]. We filter out profiles that scored below 50 and are left with 573 profiles.



(a) Number of Profiles per Community



(b) Sentiment Scores: Word Count vs. Maximum Entropy

Figure 1. (a) Frequency distribution of the number of Twitter profiles in each community. (b) Maximum Entropy Sentiment vs. Word Count Sentiment: the two sentiment scores are highly correlated at aggregate level.

Socio-demographic Data. From the UK Office for National Statistics, we obtain the Index of Multiple Deprivation (IMD) score of each of the 78 census areas in London. We consider a census area to be a community. We choose such a definition of community because it has been widely used in studies of social deprivation (including the related article by Eagle *et al.* [5]) and because using IMD scores with any other definition of community would lead to results that are not ecologically valid. IMD is a composite score based on income, employment, education, health, crime, housing, and the environmental quality of each community [11]. The higher a community's IMD score, the more socially deprived the community (e.g., Tottenham, Hackney); whilst the lower the score, the less deprived the community (e.g., Mayfair, Belgravia). Of these 78 census areas, 51 contained at least one of the Twitter profiles we crawled; Figure 1(a) shows the frequency distribution of the number of profiles per community.

SENTIMENT ANALYSIS

After collecting tweets between the dates of 27 September and 10 December 2010, we measure the sentiment expressed by a profile's tweets and then compute, for each census region, an aggregate community-sentiment measure of all the profiles in the region. So, for starters, we need to measure the sentiment of a profile, and we do so using two classifications: Word Count and Maximum Entropy.

[1]http://developer.yahoo.com/geo/placemaker/
[2]http://www.peerindex.net/help/scores

Word Count. We use a dictionary called "Linguistic Inquiry Word Count". LIWC is a standard dictionary of 2,300 English words that capture 80% of the words used in everyday conversations and reflect people's emotional and cognitive perceptions. After removing stop-words from tweets, we count, for each profile, the number of words that are positive and those that are negative (words matching the two categories of 'positive emotions' and 'negative emotions' as defined in LIWC) and aggregate both counts to produce the "Word Count Sentiment" score, which is similar to the score proposed by Kramer [9]:

$$Sentiment_i^{WC} = \frac{p_i - \mu_p}{\sigma_p} - \frac{n_i - \mu_n}{\sigma_n} \quad (1)$$

where p_i (n_i) is the fraction of positive (negative) words for user i; μ_p (μ_n) is the fraction of positive (negative) words, averaged across all users; and σ_p (σ_n) is the corresponding standard deviation. The normalization using means and standard deviations accounts for the unbalanced distribution of positive and negative words of the English language [9].

Maximum Entropy. A limitation of the Word Count technique is that the vocabulary it uses does not contain all English words that are positive or negative. We thus need a classifier that can learn new positive/negative words. To this end, we resort to a machine learning technique called Maximum Entropy (MaxEnt for short). The technique has been proven to be effective in a number of natural language processing applications, including sentiment classification of tweets [1, 6]. To train MaxEnt, we used a dataset provided by a group of students from Stanford University that consists of 232,442 "smiley" tweets (containing the :-) emotion) and 151,955 "frowny" tweets (containing the :-(emotion). We consider "smiley" and "frowny" faces to reflect the ground-truth sentiment of the corresponding tweets, similarly to previous work by Go *et al.* [6]. After being trained on the ground-truth, MaxEnt is then able to classify future tweets. We do so for the tweets in our London dataset, and compute a profile's "Maximum Entropy Sentiment" using, again, formula (1).

Effectiveness of classifiers. Having the two classifiers at hand, we now need to measure how well they perform. Upon 10-fold cross validation, we find that the two classifiers show very similar *accuracy* upon tweets they are able to classify (precision is around 66%) but different *recall* in that Word Count leaves more tweets unclassified than what MaxEnt does (recall is 38% for Word Count and 68% for MaxEnt). However, these results are for single tweets. At profile level, the two classifiers perform similarly. We plot one classifier's profile sentiment scores versus the other's in Figure 1(b) and clearly observe that the two quantities are strongly correlated (Pearson correlation coefficient of $r = .73$): each profile, on average, is considered to be positive/negative to a very similar extent by both classifiers. Given these results at profile level, one concludes that the two classifiers perform similarly, all the more so at community level.

However, critics might rightly say that words not reflecting positive moods such as greetings during Christmas might well wrongly skew our sentiment classifications. This bias holds but is alleviated in three ways. First, we study profiles in the single area of London and, as a result, we control for any variability in the use of words across geographic areas. Second, our sentiment analysis is a comparative one - time-specific or area-specific words (e.g., greetings) do introduce biases but do so across all profiles and, as such, have little impact on a comparative analysis. Third, any bias left (e.g., one could argue that deprived areas always use less greetings) is of little concern because our classification is not of single tweets but of entire profiles in which the importance of individual greetings is largely diluted.

Gross Community Happiness. Upon the sentiment classifications of profiles, we finally compute "Gross Community Happiness" (*GCH*) as the mean sentiment score of the profiles in a community: $GCH_C = Mean\{Sentiment_i\}_{\forall i \in C}$, where $Sentiment_i$ is the sentiment score (computed with either word count or maximum entropy) of each profile i in community C. We choose the metric of *GNH* to ease the interpretation of our results: the mechanics behind *GNH* are easy to understand, as opposed to machine learning approaches that often act as black boxes, and *GNH* also makes our correlation results comparable with previous work. Furthermore, as for *GNH*'s input sentiment classifications, we have just shown that classifying (not tweets but) *profiles* does not require sophisticated techniques - Word Count does a reasonable job.

RESULTS

We now use the previously computed sentiment scores across different communities to make two assessments. First, we aim to gain an insight into the geographic distribution of sentiment across London neighborhoods. We do so by overlaying "Gross Community Happiness" (*GCH*) scores onto a city map using the Google Maps Javascript API (Figure 2). The map loosely indicates that the North West is the happiest quadrant of the city, whilst South East London is fairly awash with blue, indicating unhappy sentiment. Indeed, conventional wisdom among Londoners holds that "Generally London has a North West/South East gradient of posh to rough areas. It's not so much 'West end' vs. 'East end' as 'North West' vs. 'South East'. Many of the really dodgy parts of London are in the South East"[3]. These anecdotes are quantitatively supported by the gradient of IMD scores between North West London and South East London [11].

Second, we compute the Pearson correlation coefficient between IMD and "Gross Community Happiness" scores. In so doing, we learn that the two quantities are correlated to a significant extent: $r = .350$ if *GCH* is computed with word count ($p < 0.05$), and $r = .365$ if *GCH* is computed with MaxEnt ($p < 0.01$). These correlations are statistically significant. However, for areas with less than 9 profiles, correlation coefficients do vary but their variations are not easy to interpret as they are not statistically significant. For the 33 areas with more than 9 profiles, correlation coefficients become stable (around .35) and statistically significant, and the corresponding IMD distribution is still nor-

[3]What's the worst part of London? http://tinyurl.com/3ldhu8q

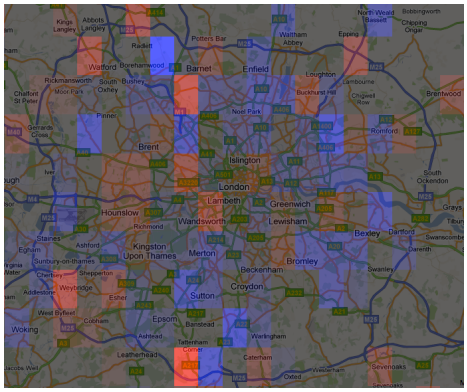


Figure 2. Mapping "Gross Community Happiness" in Greater London (best seen in color).

mally distributed. More generally, there is no correlation between one area's number of Twitter profiles and its IMD. Correlation coefficients are rather high considering that, at individual level, sentiment expressed on social media correlates with self-reported life satisfaction with a coefficient of about $r = .17$ [9]. These results ultimately suggest that users in more deprived London communities tweet, on average, more negatively than those in less-deprived communities.

DISCUSSION

Our analysis has demonstrated that the relation between sentiment and well-being does not only hold for individuals but scales to the level of communities too, so much so that we have created an aggregate metric out of community residents' tweets and have successfully validated it: the metric correlates with census socio-economic well-being of communities. This result suggests that it is possible to effectively track the emotional health of local communities from their residents' tweets in an unobtrusive way, as tweets are publicly available and easily crawled.

The significance of these results extends beyond merely tracking emotional health of local areas: they provide evidence that users' offline communities have a noticeable effect on their online interactions. To appreciate the importance of this insight, consider that past research has suggested two relations relevant to this study. The first is between where people live and their subjective well-being: income inequality, unemployment rates, urbanization, safety and deprivation of an area have all been shown to relate to people's subjective well-being across different countries and time periods [4]. The second relation is between subjective well-being and what people write on social media: sentiment expressed in user-generated content gets more positive as people are increasingly satisfied with their lives [9]. From these two relations, a third one might transitively follow but has never been tested: the relation between where people live and social media content they generate. We have now tested this third relation and found that, indeed, tweets from residents of socially deprived communities contain more negative emotions than those from residents of well-off communities. Interestingly, this link between offline and online worlds counsels cautions against claims that social-networking sites like Twitter "dehumanize" community life [7, 10]; on the contrary, they seem to reflect important aspects of physical communities. Our recent findings further support this view. We have studied the Twitter network and have recently found that one's sentiment strongly correlates with the average sentiment of one's Twitter friends and followers. This suggests, once again, that dynamics observed in the physical world such as homophily are reflected in digital interactions.

This study has three limitations that call for further investigation in the future. The first is demographic bias: 63% of Twitter users are less than 35 years old and 68% have a total household income of at least $60, 000 in the United States. The results we presented thus disproportionately represents the happiness of some citizens over others. This is one of reasons why we have chosen London: it had been the top Twitter-using city in the world until the beginning of 2010 [2], and as the service penetration rate increases, demographic bias is bound to decrease. The second limitation is that our results do not speak to causality. Though the causal direction is difficult to be determined from observational data, one could repeatedly crawl Twitter over multiple time intervals and use a cross-lag analysis to potentially observe causal relationships. The third limitation is that we have tracked sentiment but have not studied *what* actually makes communities happy. To that end, we are currently extracting the *subject matter* of tweets using topic models [12] and will then compare topics across communities. For example, given two communities, one talking about yoga and organic food, and the other talking about gangs and junk food, what can be said about their levels of social deprivation? The hope is that topical analysis will answer this kind of questions and, in so doing, assist policy makers in making informed choices regarding, for example, urban planning.

REFERENCES

1. L. Barbosa and J. Feng. Robust sentiment detection on Twitter from biased and noisy data. In *Proc. of the* 23^{rd} *COLING*, August 2010.
2. M. Butcher. London is the capital of Twitter, says founder. TechCrunch Europe, August 2009.
3. E. Diener, M. Diener, and C. Diener. Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology*, 69(5):851–864, 1995.
4. P. Dolan, T. Peasgood, and M. White. Do we really know what makes us happy? *Journal of Economic Psychology*, 29(1):94–122, 2008.
5. N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.
6. A. Go, R. Bhayani, and L. Huang. Twitter Sentiment Classification using Distant Supervision. In *Stanford Tech Report*, December 2009.
7. C. Irvine. Twitter CEO dismisses claims that social networking 'dehumanises'. The Telegraph, August 2009.
8. F. Kivran-Swaine and M. Naaman. Network properties and social sharing of emotions in social awareness streams. In *Proc. of the ACM CSCW*, March 2011.
9. A. Kramer. An unobtrusive behavioral model of "Gross National Happiness". In *Proc. of the* 28^{th} *ACM CHI*, April 2010.
10. J. Lanier. *You Are Not a Gadget: A Manifesto*. Knopf, January 2010.
11. M. Noble *et al.* The English Indices of Deprivation 2007. The Department of Communities and Local Government, March 2008.
12. D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In *Proc. of the AAAI ICWSM*, 2010.
13. A. Stratton. Happiness index to gauge Britain's national mood. The Guardian, November 2010.