Talk of the City: Our Tweets, Our Community Happiness

Daniele Quercia Diarmuid Ò Séaghdha Jon Crowcroft
The Computer Laboratory, University of Cambridge, UK
dq209@cl.cam.ac.uk, do242@cam.ac.uk, jac22@cl.cam.ac.uk

Abstract

The literature of urban sociology and that of psychology have separately established two relationships: the first has linked characteristics of a community to its residents' well-being, the second has linked well-being of individuals to their use of words. No one has hitherto explored the potential transitive relationship - that between characteristics of a community and its residents' use of words. We test this relationship by performing three steps. We consider Twitter users in a variety of London census communities; extract the subject matter of their tweets using "topic models"; and study the relationship between topics and community socio-economic well-being. We find that certain topics are correlated (positively and negatively) with community deprivation. Users in more deprived community tweet about wedding parties, matters expressed in Spanish/Portuguese, and celebrity gossips. By contrast, those in less deprived communities tweet about vacations, professional use of social media, environmental issues, sports, and health issues. We finally show that monitoring the subject matter of tweets not only offers insights into community well-being, but it is also a reasonable way of predicting community deprivation scores.

1 Introduction

Urban sociologists have found a link between the characteristics of a community and well-being of its residents. There are specific characteristics of a built environment that promote physical activity and, consequently, residents' well-being. For example, some places are designed in such a way that physical activity is encouraged (e.g., playgrounds, open spaces, bike lanes), while other places are designed and managed in such a way that physical activity is unattractive (e.g., streets without sidewalks, crime-infested parks). Physical activity is not the only factor that impact community health: other factors include population density (Lopez 2004), land use diversity (Cervero and Duncan 2003), and block size (Boer and al. 2007). Given the strong connection between built environment and community health, there is a growing literature on the development and evaluation of measures that capture community characteristics (Sallis 2009).

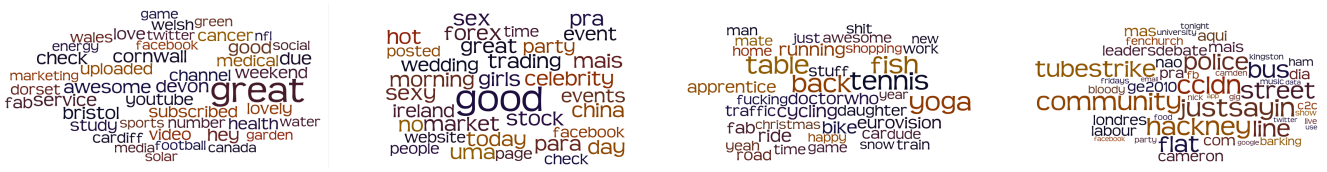
At the same time, researchers in social psychology have found a link between well-being of individuals and their use of words. To a certain extent, "our words reflect ourselves" (Henrich, Heine, and Norenzayan 2003). The number of first-person pronouns (e.g., I, my) in speech or writing often correlates with narcissism and with the personality trait of "Neuroticism" (Stirman and Pennebaker 2001; Weintraub 1989). Second-person pronouns (e.g., you) and third-person pronouns (e.g., she, they) are markers of social engagement and negatively correlate with depression (Rude, Gortner, and Pennebaker 2004). Furthermore, words that express positive emotions (e.g., good, happy) are used more by people who are satisfied with their lives (Pennebaker and King 1999).

Considering the two links 'community' $\rightarrow$ 'well-being' $\rightarrow$ 'words', a transitive link might follow: that between characteristics of a community and its residents' use of words. We test this hypothesized correspondence by making the following contributions:

- We crawl tweets produced by Londoners and obtain census data - socio-economic well-being scores called Index of Multiple Deprivation (*IMD*) scores - for their communities, and we extract the subject matter of tweets using topic modeling.
- We study the relationship between topics and socio-economic well-being. We identify discussion topics that show significant correlation with *IMD*. We also show that, knowing the topical distribution of tweets in a community, it is possible to predict the unseen community's *IMD* score.

2 Our Analysis

Datasets. To identify tweets from neighborhood residents, we consider 573 Twitter profiles whose user-specified locations are London neighborhoods (e.g., Brixton, Notting Hill) (Quercia et al. 2012). To then identify tweets not only from residents but also from visitors, we also consider the geo-referenced tweets collected by Cheng et al. (Cheng, Caverlee, and Lee 2011). They collected Twitter updates (single tweets) that report location information. We take the



(a) Residents (well-off) (b) Residents (deprived) (c) Visitors (well-off) (d) Visitors (deprived)

Figure 1: Tag clouds of topical words for *residents* ((a) and (b)) and visitors ((c) and (d)) of London communities. Word size is proportional to corresponding correlation coefficients with deprivation scores.

228,625 tweets that fall into the wider area of Greater London. Finally, 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. This is a composite score based on income, employment, education, health, crime, housing, and the environmental quality of each community (Noble *et al.* 2008). The higher a community's *IMD* score, the more socially deprived the community (e.g., Brixton); whilst the lower the score, the less deprived the community (e.g., Notting Hill).

Topical Analysis. The corpus of tweets was preprocessed with a standard pipeline (e.g., text converted to lowercase, tokenized, specific English stopwords removed). Tweets are by definition short even before filtering, making it difficult to model each individual tweet accurately; as a result we aggregate all tweets for a given user in one "document" (we are interested in the topical distribution at community level after all). We then extract the subject matter of tweets using topic modeling, a state-of-the-art method for identifying thematically related clusters of words from a document collection that has previously been applied to many different text types including Twitter data. More specifically, we use the MALLET implementation of *LDA* with learning parameters set to their default values[1].

Analysis of residents' tweets. To understand whether residents in deprived areas and residents in less deprived areas talk about different things, we will first study the relationship between the topics covered by Twitter users and their communities' *IMD* scores. The number of topics is set to 100. Since *LDA* returns a profile's topical distribution over these 100 topics (e.g., the prevalence of *topic 1* in the profile is, say, 10% and of that of *topic 2* is 90%), we first normalize a profile's topical distribution by computing what we call *normalized fraction* of topic t in the profile: $z_t = \frac{p_t - \mu_t}{\sigma_t}$, where p_t is the original (*LDA-computed*) prevalence of topic t in the profile; μ_t is the prevalence of topic t, averaged across *all* profiles; and σ_t is the corresponding standard deviation. We then compute the Pearson product-moment correlation between the normalized fraction of each topic and *IMD* across all profiles. Pearson's correlation $r \in [-1, 1]$ is a measure of the linear relationship between two random variables. Table 1 reports only the correlation coefficients that are statistically significant. For each topic (row), the table reports: the

[1]http://mallet.cs.umass.edu/

Residents of Well-off Communities		
t_r	ρ	**Top words**
9	-0.12	bristol devon service due cornwall
13	-0.10	wales cardiff welsh number dorset
18	-0.09	social media twitter facebook marketing
19	-0.09	green energy garden water solar
37	-0.11	video youtube uploaded channel subscribed
56	-0.09	game sports football nfl canada
66	-0.11	lovely great weekend love fab
67	-0.11	health cancer care study medical
71	-0.12	awesome great good hey check
Residents of Socially-deprived Communities		
t_r	ρ	**Top words**
26	0.11	wedding event events ireland party
32	0.12	no para pra uma mais
38	0.12	hot sex sexy celebrity girls
44	0.09	check facebook posted website page
49	0.12	great good day today morning
82	0.09	people time it's good don't
85	0.12	market trading forex stock china

Table 1: List of statistically significant topics extracted from tweets of *residents* in less and more deprived London communities ($p < 0.05$ at most).

topic identifier t_r in [1,100] range (t_r stands for topic generated by **r**esidents); the Pearson correlation coefficient ρ between the topic's prevalence and deprivation (*IMD*) scores; and the list of top (most frequent) words associated with the topic.

From the correlations in Table 1 (whose top-words the first two tag clouds in Figure 1 graphically depict), we find that the less a community is socially-deprived, the more the discussion of its residents are about vacation resorts (topics t_v 9 and 13) and vacation time (topic 66), professional use of social media (topics 18), environmental issues (topic 19), sports (topic 56), health issues (topic 67). By contrast, the discussions in deprived communities are about the Royal wedding (topic 26), matters expressed in Spanish/Portuguese (topic 32), and celebrity gossip (topic 38).

There are few results left uncommented, and they might seem puzzling at first. To gain insight into the matter, we have built some diagnostic tools and performed further anal-

Visitors of Well-off Communities		
t_v	ρ	**Top words**
1	-0.11	car road traffic stuff daughter
12	-0.11	home work time just train
13	-0.20	back tennis table fish yoga
14	-0.11	fucking awesome yeah man shit dude mate game
25	-0.13	#eurovision fab #apprentice #doctorwho
41	-0.14	bike cycling ride running
42	-0.10	christmas snow happy new year shopping
Visitors of Socially-deprived Communities		
t_v	ρ	**Top words**
2	0.21	que por con los pero para del una las
4	0.22	que não mas pra com aqui dia mais londres
11	0.10	use twitter google data email app facebook
22	0.17	#fb c2c fenchurch barking bloody ham
27	0.12	tonight gig show live party music camden
30	0.35	bus street #tubestrike line flat police
33	0.12	university kingston food nick fridays
37	0.37	community #ccldn hackney #justsayin
38	0.21	labour #ge2010 cameron #leadersdebate

Table 2: List of statistically significant topics extracted from tweets of *visitors* (and possibly residents) in less and more deprived London communities ($p < 0.05$ at most).

ysis. First, residents of both types of communities express positive emotions in their tweets (topics 49 and 71). Indeed, both topics are generally positive and event-based. However, topic 49 in socially-deprived communities is about work ("meeting", "working", "office" , "exciting", "forward", "client"), while topic 71 in well-off communities is slangier and is about having a good time ("awesome", "it's", "i'm", "crazy", "tonight", "party"). Second, residents in both more and less deprived communities tend to talk about social media (topics 37 and 44). However, topic 37 in well-off communities is about online video ("video", "youtube", "uploaded", "trailer"), while topic 44 in socially-deprived communities is about social-networking sites ("facebook", "posted", "photos", "join", "twitter", "friends"). So residents in less deprived communities tend to talk about content consumption, while those in more deprived communities tend to talk about social-networking interactions. Third, topic 85 (market trading) in deprived communities is associated with Twitter profiles that are financial news digests. There are two partial explanations for this, and both might be contributing. First, research in individual well-being has consistently found that "people who care a lot about becoming rich tend to be less happy, on average, than those for whom getting rich is less important" (Bok 2010). Second, Rotherhithe is a not-so-well-off area that, being close to the financial districts, hosts a considerable number of analysts. Finally, as for topic 82 ('people', 'time', 'good') in deprived communities, the most strongly associated words (using t-test statistic) reflect "self-help" and "personal development".

Analysis of residents' and visitors' tweets. We now try to understand whether visitors (people who happen to be) in deprived areas and visitors in less deprived areas talk about different things. To this end, we will study the relationship between the topics expressed in geo-referenced tweets (not profiles) and *IMD* scores. After preprocessing, the dataset consists of 3,422 user "documents" and 1.2 million words. We set the number of topics to 100 and, by inspection, learn that some of the resulting topical clusters are related with each other. Thus we set the number of topics to 50, which results into more "semantically orthogonal" clusters. The choice of 50 or 100 creates more or less redundancy in the topic space but does not affect the overall interpretation of the correlation coefficients we will present next.

Table 2 reports only the correlation coefficients that are statistically significant. From these correlations (which the last two tag clouds in Figure 1 graphically depict), we find that the less a community is socially-deprived, the more its visitors talk about: family matters (topic t_v 1 and 12); sport (topic 13, which shows a correlation of $\rho = 0.20$) and, more specifically, cycling (topic 41); popular TV programs (topic 25); and Christmas shopping (topic 42). Topic 14 reports slang expressions generally used by youngsters - these individuals are likely to be visitors (and not residents) of central (well-off) areas, and their tweets could not be captured by our previous analysis of residents. By contrast, visitors of more deprived communities, tend to, much like their residents, express themselves in Spanish (topic t_v 2) and Portuguese (topic 4) and talk about the use of social-networking sites (topic 11). They also mention strikes and police matters (topic 30, which has a correlation of $\rho = 0.35$) and tended to be vocal in the past general election (hashtag #ge2010 in topic 38).

Based on the top-words associated with topic 22, we find that this topic is related to travelers on the *c2c* train line, which goes from Fenchurch Street railway station through (socially-deprived) east London to the Essex area. We have also identified individuals who talked about gigs in Camden Town (topic 27). This area is well-know for its markets of "alternative" clothing and for its music venues that are strongly associated with alternative culture. Based on its *IMD*, Camden is ranked 15^{th} most deprived in London out of 33 local authorities and 74^{th} most deprived in England out of 326 local authorities. Those interested in Kingston University's Friday celebrations (topic 33) are also represented. The university is located near Wimbledon (London suburban area). The final group of people tweeted in the area of Hackney and is associated with CityCamp London (topic 37). CityCamp is an international unconference series and online community focused on innovation for municipal governments and community organizations. In London, its events are hosted in Hackney, whose wards remain among the 10% most deprived in the country. Yet, the area also hosts a considerable number of technology spin-off companies (e.g., Last.fm).

3 Predicting Well-being

We now turn to predicting a community's *IMD* from the topics discussed in Twitter. We build a regression model that predicts *IMD* as a linear combination of the normalized fractions z_t of all topics. The extent to which the regression predicts IMD_i is reflected in a measure called R^2 - the higher R^2, the better the fit of the model. In our case, $R^2 = .32$ for residents' tweets, suggesting that the topics discussed by a community's residents do indeed explain a large share or variance of the community's *IMD* (roughly 32% of it). The correlation coefficients in Table 1 are admittedly weak, yet an $R^2 = .32$ reflects a good model in social science, especially if one measures something personal and full of variation (e.g., well-being). The linear regression of the residents' and visitors' tweets has an R^2 as high as .49, which, for the purposes of explorative research, is considered a medium-high coefficient, suggesting that the topics discussed by a community's visitors explain a very high variation in *IMD* - 49% of it. Using fewer topics as predictors - say, the eleven topics that are statistically significant - still results into R^2 being as high as .44, suggesting that the predictive ability does not significantly depend on the number of predictors used.

4 Discussion

Theoretical Implications. The implications of this study go well beyond monitoring the subject matter of tweets in two major ways. First, we have shown that Twitter has some connection with objective physical reality: the characteristics of physical communities have noticeable effects on what Londoners talk about online. This comes as no surprise for some and, at the same time, might catch popular press pundits off-guard, especially those who have claimed that social-networking sites like Twitter "dehumanize" community life (Irvine 2009; Lanier 2010). Second, topical analysis of tweets might represent a novel way of studying people's perceptions about their physical communities. The most widely-used way of collecting people's perceptions is self-reporting, that is, asking residents how they perceive their neighborhoods. However, self-reporting has been found to produce unreliable results (Sampson and Raudenbush 2004). To tackle this problem, researchers have been proposing alternatives. Clarke *et al.*, for example, evaluated the use of Google Street View to capture community characteristics at scale (Clarke and et al. 2010). Our study has proposed yet another way of collecting community (social) characteristics based on unobtrusive collection and analysis of user-generated content.

Practical Implications. This study also suggests that, with topics extracted from tweets, one is able to predict census well-being data, and that opens up the possibility of tracking the emotional health of local communities at scale. This possibility supports the vision behind "smart cities": new information and communication technologies will be needed to promote healthy and socially sustainable communities and, more generally, to better manage complex urban systems.

Limitations. This study has two 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 topics discussed by 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 (Butcher 2009), 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, so a cross-lag analysis to potentially observe causal relationships is in order.

Acknowledgment. We thank EPSRC for its financial support through the Horizon Digital Economy Research grant (EP/G065802/1).

References

[Boer and al. 2007] Boer, R., and al., e. 2007. Neighborhood design and walking trips in ten U.S. metropolitan areas. *American Journal Preventive Medicine*.

[Bok 2010] Bok, D. 2010. *The Politics of Happiness: What Government Can Learn from the New Research on Well-Being*. Princeton University Press.

[Butcher 2009] Butcher, M. 2009. Twitter CEO dismisses claims that social networking 'dehumanises'. TechCrunch Europe.

[Cervero and Duncan 2003] Cervero, R., and Duncan, M. 2003. Walking, bicycling, and urban landscapes: Evidence from the San Francisco Bay Area. *American Journal of Public Health*.

[Cheng, Caverlee, and Lee 2011] Cheng, Z.; Caverlee, J.; and Lee, K. 2011. Exploring millions of footprints in location sharing services. In *Proceedings of ICWSM*.

[Clarke and et al. 2010] Clarke, P., and et al. 2010. Using Google Earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health & Place*.

[Henrich, Heine, and Norenzayan 2003] Henrich, J.; Heine, S.; and Norenzayan, A. 2003. Psychological aspects of natural language. use: our words, our selves. *Annual Review Psychology*.

[Irvine 2009] Irvine, C. 2009. Twitter CEO dismisses claims that social networking 'dehumanises'. The Telegraph.

[Lanier 2010] Lanier, J. 2010. *You Are Not a Gadget: A Manifesto*. Knopf.

[Lopez 2004] Lopez, R. 2004. Urban sprawl and risk for being overweight or obese. *American Journal Public Health*.

[Noble *et al.* 2008] Noble *et al.*, M. 2008. The English Indices of Deprivation 2007. The Department of Communities and Local Government.

[Pennebaker and King 1999] Pennebaker, J., and King, L. 1999. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*.

[Quercia et al. 2012] Quercia, D.; Ellis, J.; Capra, L.; and Crowcroft, J. 2012. Tracking "gross community happiness" from tweets. In *Proceedings of the* 15^{th} *ACM CSCW*.

[Rude, Gortner, and Pennebaker 2004] Rude, S.; Gortner, E.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*.

[Sallis 2009] Sallis, J. 2009. Measuring physical activity environments: a brief history. *American Journal Preventive Medicine*.

[Sampson and Raudenbush 2004] Sampson, R. J., and Raudenbush, S. W. 2004. Seeing Disorder: Neighborhood Stigma and the Social Construction of Broken Windows. *Social Psychology Quarterly*.

[Stirman and Pennebaker 2001] Stirman, S., and Pennebaker, J. 2001. Word Use in the Poetry of Suicidal and Nonsuicidal Poets. *Psychosomatic Medicine*.

[Weintraub 1989] Weintraub, W. 1989. *Verbal Behavior in Everyday Life*. New York: Springer.