Understanding Happiness in Cities using Twitter: Jobs, Children, and Transport

Conference Paper · September 2016

CITATION
1

READS
299

4 authors:

Weisi Guo
The University of Warwick
120 PUBLICATIONS **1,017** CITATIONS
SEE PROFILE

Neha Gupta
The University of Warwick
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Ganna Pogrebna
The University of Warwick
38 PUBLICATIONS **297** CITATIONS
SEE PROFILE

Stephen A. Jarvis
The University of Warwick
231 PUBLICATIONS **3,043** CITATIONS
SEE PROFILE

Understanding Happiness in Cities using Twitter: Jobs, Children, and Transport

Weisi Guo[1,2*], Neha Gupta[1,4], Ganna Pogrebna[1,3], Stephen Jarvis[1,4]

***Abstract*—The demographics and landscape of cities are changing rapidly, and there is an emphasis to better understand the factors which influence citizen happiness in order to design smarter urban systems. Few studies have attempted to understand how large-scale sentiment maps to urban human geography. Inferring sentiment from social media data is one such scalable solution. In this paper, we apply natural language processing (NLP) techniques to 0.4 million geo-tagged Tweets in the Greater London area to understand the influence of socioeconomic and urban geography parameters on happiness. Our results not only verify established thinking: that job opportunities correlate with positive sentiments; but also reveal two insights: (1) happiness is negatively correlated with number of children, and (2) happiness has a U-shaped (parabolic) relationship with access to public transportation. The latter implies that the happiest people are those who have good access to public transport, or such poor access that they use private transportation.**

The number of jobs, children, and transportation availability are every day facets of urban living and individually account for up to 47% of the variations in people's happiness. Our results show that they influence happiness more significantly than long term socioeconomic parameters such as degradation, education, income, housing, and crime. This study will enable urban planners and system designers to move beyond the traditional cost-benefit methodology and to incorporate citizens' happiness.

***Index Terms*—happiness; social media data; sentiment;**

I. Introduction

For the longest part of our existence, human beings have primarily lived in rural environments. This close proximity to nature has fashioned both our social and biological evolution. It is only in the last 200 years (a few generations) that the number of people living in cities has risen from 3% to over 50% of the global population. In the past 50 years and within a single generation, there has been a 6-fold increase in the number of large metropolitan areas [1]. The urbanisation trend has presented new economic and technological opportunities to humanity, but it has also created a set of urban development challenges related to health and happiness. Existing studies have shown that both the benefits and challenges of cities scale super-linearly with the city's size [2], [3], and the growing global urban population certainly exasperates the hidden competition between urban improvements and decay. The question, of how the high density of opportunities (i.e., jobs) and urban threats (i.e., crime and pollution) affect our happiness has become more pertinent than ever.

Surveying citizen happiness is an important area of research [4]–[6]. Qualitatively, the pursuit of happiness are cornerstone philosophies in the governance theory propelled by many ancient cultures. In modern history, quantitative measures such as the Gross National Happiness (GNH) gained traction after 2005. Given the subjective nature of happiness, it is typically measured through self-reported surveys that are validated and normalised against more objective metrics that are widely accepted as ones that support positive sentiment (i.e., income and lifespan). Existing research projects have pursued both qualitative and empirical experiments to understand the sentiment of urban spaces [4]. Indeed, census data is extensively used by governments to create well-being scores (an example can be found for London[1]). However, the data from survey based methods are limited in their resolution (spatial-temporal). Alternative sentiment data collection methods employ wearable monitoring systems such as electro-dermal-activity sensors [7]. These systems will yield precise longitudinal data with high spatial-temporal accuracy. However, their expensive nature means that scaling to the general public and establishing pervasive and non-intrusive sensing remains challenging.

The proliferation of online social interactions has in recent years provided an opportunity to study sentiment

[1]Warwick Institute for the Science of Cities (WISC), University of Warwick, UK. [2]School of Engineering, University of Warwick, UK. [3] Warwick Manufacturing Group, University of Warwick, UK. [4]Department of Computer Science, University of Warwick, UK. *Corresponding Author: weisi.guo@warwick.ac.uk. Funding Acknowledgement: EPSRC Centre for Doctoral Training in Urban Science and Progress - EP/L016400/1, and ESRC Centre for Competitive Advantage in the Global Economy (CAGE).

[1]http://data.london.gov.uk/dataset/london-ward-well-being-scores

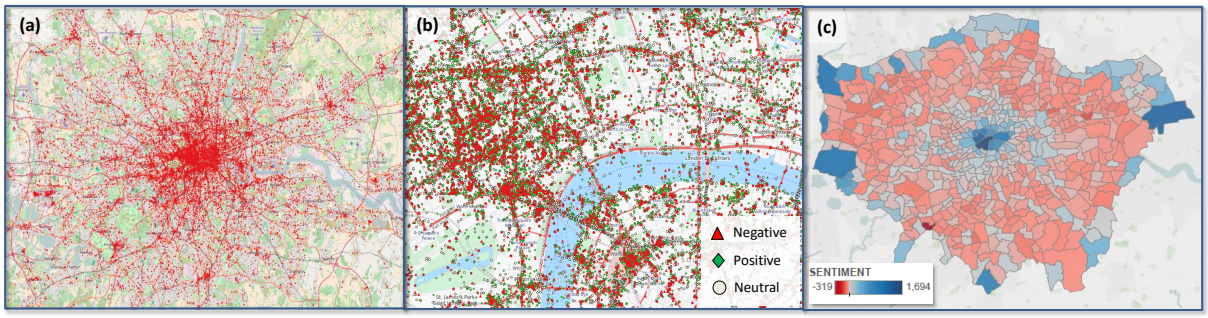

Fig. 1. **Mapping the Sentiment in London:** (**a**) 0.4 million geo-tagged Tweets in Greater London over a 2-weeks period. (**b**) Tweets labelled as negative (red triangle), positive (green diamond), or neutral (pale circle) on a scale of 11. (**c**) Ward level sentiment where dark red indicates negative sentiment and dark blue indicates positive sentiment.

of urban dwellers (residents, workers, tourists, etc.). Social media platforms, such as Twitter, have achieved significant penetration (25% of adult population in the UK), and usage (over 500 million messages per day worldwide). Whilst detecting sentiment using social media data as a proxy incurs bias, it does offer attractive benefits in scalability and there is growing research to validate and benchmark the sentiment labels. Topic based sentiment analysis utilizing natural language processing (NLP) has been extensively exploited in business intelligence [8]. There has also been a growing body of work in applying similar methodologies to examining the sentiment of citizens in urban spaces. In terms of similar research, there have been numerous studies conducted on detecting emotions from Twitter data [9] and creating mood heat maps of city locations [10] as well as comparing between cities [5], [6]. Social media data has the added benefit of not only uncovering real-time high spatial-temporal resolution meta-data, but also reaching across urban demographics to include residents, workers, and tourists.

Despite the growing abundance in urban related sentiment studies through social media data, as far as we are aware, very few research outputs have attempted to understand how large-scale sentiment data (obtained from social media) maps to urban socioeconomic and infrastructure features. As such, without such a mapping, we are no closer to understanding the underlying causes of happiness. Furthermore, without understanding how human beings feel about their urban environment, urban planners are limited to planning services using traditional costbenefit analysis using economic indicators and cannot consider accurately the consequential effects it has on citizen sentiment [11]. This study, as far as we are aware of, is the first attempt to map and correlate large-scale sentiment data to urban geography features, and consequently attempt to understand the main sources of happiness in the city landscape.

II. Methods

A. The Data

The data used in this paper comes from two sources: (1) 0.4 million geo-tagged social media data purchased from Twitter, covering a 2 weeks period (see Fig. 1**a**), and (2) UK government ward-level socioeconomic and urban geographical data (open access) from the London Data Store[2]. In terms of spatial resolution, the analysis in this paper will focus on Greater London, which is made up of 628 wards, and are roughly analogous to a neighbourhood. Many services are delegated to the ward level, including policing; and a range of census statistics are available at the ward level. The ward level census data considers 64 key metrics, including demographics, education, housing, and business statistics.

This paper's focus on using Twitter data (aggregated from all urban dwellers) as a proxy and comparing it to census data (mainly registered residential and business data) means that we are concerned with how **all people in London** (including residents, workers, tourists) feel as a function of the urban geography and its socioeconomic parameters. It is extremely challenging to understand what distribution of the social media data belongs to which demographic, and in this paper we treat all data as equally important (uniform weighting) and do not consider demographic categories within the sentiment data.

[2]http://data.london.gov.uk/dataset/ward-profiles-and-atlas

B. Sentiment Labelling using NLP

In this paper we employ unigram (i.e., keyword) based sentiment analysis. Whilst state-of-the-art methods often include classifying entire sentences using machine learning (e.g. Maximum Entropy, Support Vector Machine), it can be challenging to scale such methods accurately to reflect the diversity and veracity in millions of Twitter users over a large urban area. Therefore, as a first approach, we apply established unigrams to find the polarity of the tweets, and measure a general happiness averaged over a small area (i.e., a ward). This technique was successfully implemented in previous research to analyse sentiment [12], but has not been applied to urban contexts to understand the underlying sources of happiness.

To assign each tweet with a sentiment score we first apply *Tokenization* filtering to remove language noise and transform all text to a common lower case format with no punctuations. We then extract single words or features (unigrams) independently to determine the orientation of the tweet. Researchers in opinion mining have focused on trying to find suitable lexicon for classifying tweets sentiments by annotating tweets for negative or positive polarity (henceforth *happiness*) by recognising words as positive and negative sentiment. We apply the opinion lexicon [13] (full list is approximately 6800 words [3]) to each tweet. Our algorithm calculates the score of each tweet by simply subtracting the number of occurrences of negative words from the number of positive occurrences for each tweet. An example of the sentiment labelled Tweets is shown in Fig. 1**b**, and clustered to ward level in Fig. 1**c**. An interesting trend can be observed: that the happy wards (blue) are either in the centre or on the outer edges of Greater London, and the unhappy wards (red) are in the middle. We will analyse this in greater detail in Section III-D.

C. Metrics for Comparison

In order to conduct cross-dataset comparisons, the coefficient of determination, denoted R^2 is a number that indicates how well the statistical regression model fits the data or in other words: the percentage of variance in the data that can be explained by the proposed model. For a data vector $y = [y_1, y_2, ...y_K]$ (with mean $\overline{y}$) and a predicted data vector using the regression model $\hat{y}$, the residue vector is defined as $e = y - \hat{y}$. The coefficient of determination R^2 is defined as:

$$R^2 \equiv 1 - \frac{\sum_k e_k^2}{\sum_k (y_k - \overline{y})^2}, \qquad (1)$$

[3]https://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

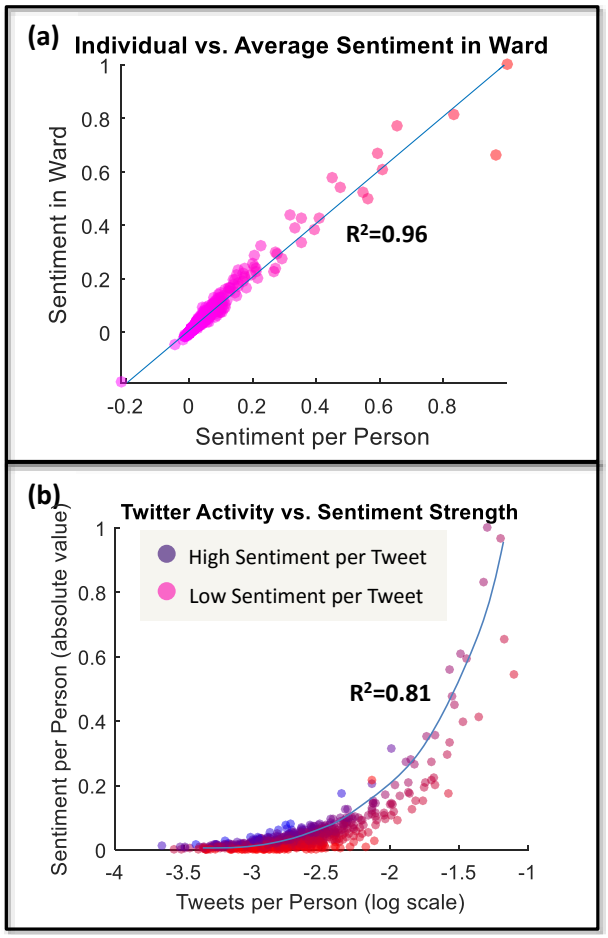

Fig. 2. **Sentiment Data Analysis:** (**a**) Ward level aggregate sentiment can accurately explain 96% of the variance in individual sentiments. (**b**) People who tweet more also express stronger aggregate sentiments, but on average express a lower sentiment per tweet.

where the numerator is the residual sum of squares and the denominator is the total sum of squares. In this paper, we use the adjusted $R^2 = 1 - (1 - R^2)\frac{K-1}{K-P-1}$ to take discount against extra variables P in the model.

III. Results

A. Baseline Sentiment Data

We first present baseline sentiment data results, to gain a better understanding of the sentiment data of individual people, their tweets, and the averaged sentiment of a ward. In order to understand the representativeness of ward-level sentiment relative to individual sentiments in the ward, we plot the average sentiment per person (in the ward) against the aggregate sentiment in the ward in Fig. 2**a**. The results show that a simple linear regression

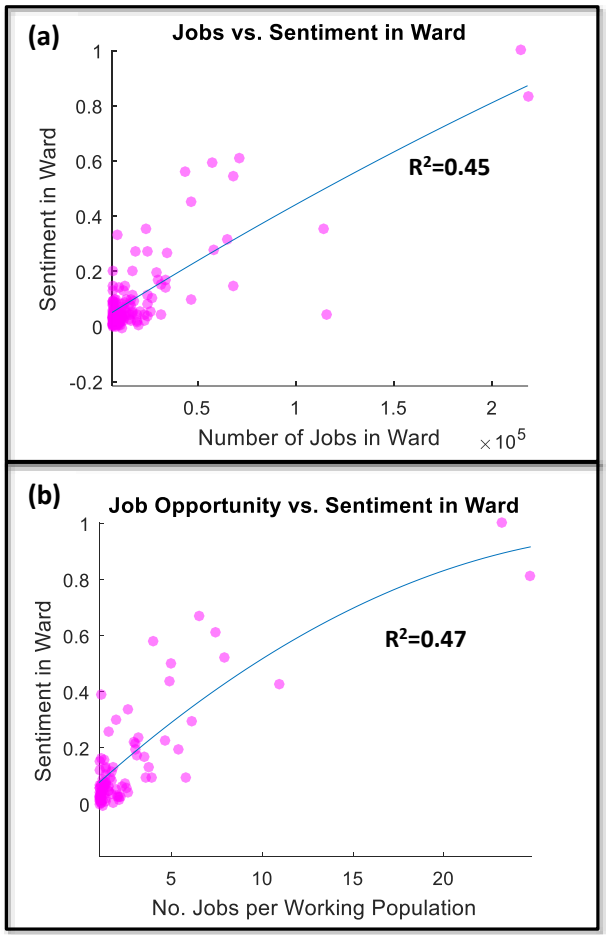

Fig. 3. **Relating Avg. Sentiment per Person to Jobs Opportunities in London:** (**a**) The number of jobs available in a ward is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.45$). (**b**) The number of jobs opportunities (jobs normalised against working population) in a ward is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.47$).

with gradient 1 can relate the ward level sentiment with the average individual sentiment. The regression can accurately explain 96% of the variance in the ward's individual sentiments. The outlier result (Harefield ward in Hillingdon borough) shows that a large discrepancy (negative bias) between individual sentiments and the ward average. This is due to a few people tweeting a high number of negative sentiments. It is also of interest to understand the relationship between the number of tweets and aggregate sentiment of tweets. The results in Fig. 2**b** shows that people who tweet more also express stronger aggregate sentiments (absolute value: either positive or negative), but on average express a lower sentiment per tweet.

The paper will now focus on 3 key areas that were identified through a correlation panel analysis (see Fig. 6 in **Appendix**): (1) Employment Opportunities, (2) Children and Fertility Rate, and (3) Accessibility to Public Transport. In particular, these are areas which affect urban lives on a daily/monthly basis and as such have a direct impact on the sentiment (see Table I in **Appendix**). It is worth mentioning that for the results to be presented below, given the census data lists over 60 urban geography features that can potentially affect happiness, obtaining a coefficient of determination for a single feature that accounts for 33 to 47% of the variations in sentiment is a **significant result**.

B. Employment Opportunities

The two main attributes in employment opportunity measured by the census data are: (i) Number of jobs in a ward (data from businesses) and (ii) Number of jobs normalised against the number of people in the working age (16-64) in a ward. Both sets of employment data are highly positively correlated with each other, as well with other crime and ambulance incident data (see Fig. 6 in **Appendix**). This reinforces the notion that increased opportunities often lead to an increase in the challenges [2], [3]. In terms of how employment relates to online sentiment, Fig. 3**a** shows the number of jobs available in a ward is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.45$). Similarly, Fig. 3**b** shows that the number of jobs normalised against working population is positively correlated with the sentiment in the ward (adjusted $R^2 = 0.47$). The adjusted $R^2 = 0.45 - 0.47$ indicates that the regressions (which both use quadratic functions, $P = 2$) explains for almost 50% of the variance in sentiment variations, and the remaining variations are due to other factors. In other words, this shows that the availability of jobs determines a significant 50% of the expressed sentiment. Yet, the sentiment is correlated with the number of jobs available and not with the number of employed people (see Fig. 6 in **Appendix**). This seems to indicate that the existence of businesses in close proximity promotes positive sentiments.

C. Number of Children

The main attributes in measuring the distribution of children in census data is the number and percentage of children (aged 0-15) in a ward. This percentage is negatively correlated with sentiment, as well with other data such as the general fertility rate (see Fig. 6 in **Appendix**). Fig. 4**a** shows the percentage of population that are children in a ward is negatively correlated with

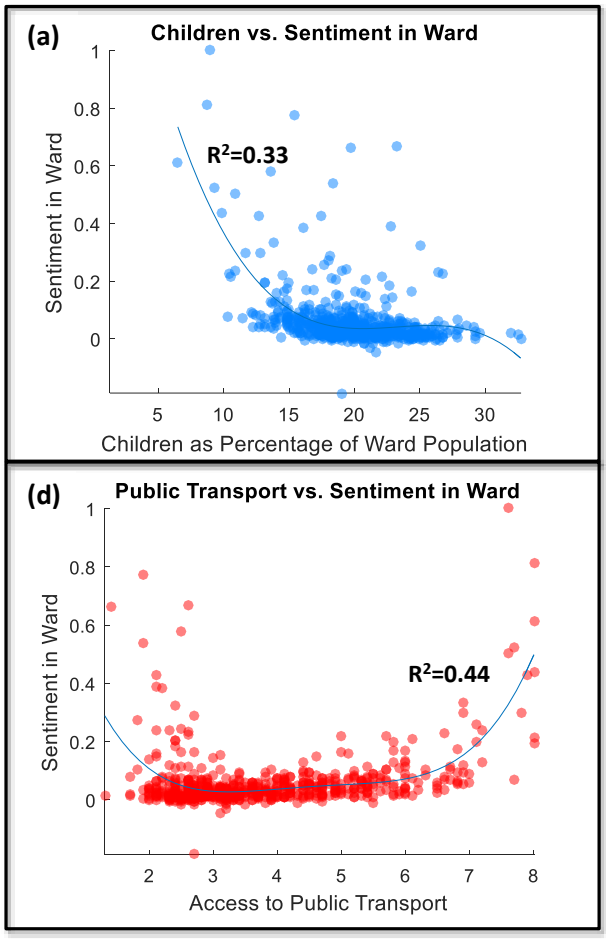

Fig. 4. **Relating Avg. Sentiment per Person to Number of Children and Access to Public Transport in London:** (**a**) The percentage of population that are children in a ward is negatively correlated with the sentiment in the ward (adjusted $R^2 = 0.33$). (**b**) The accessibility to public transport in a ward has a parabolic relationship with the sentiment in the ward (adjusted $R^2 = 0.44$), such that those with good access to public transport are happy and those who are in areas with poor public transport are also happy (rely on personal transport), whilst those that are in between are generally less happy.

the sentiment in the ward (adjusted $R^2 = 0.33$, $P = 3$). This shows that the percentage of children determines a significant 33% of the expressed sentiment. More specifically, it shows that there is a steep decline in sentiment from 5% to 15%, and the relationship saturates thereafter. It is worth noting that the percentage of children does not correlate with other socioeconomic factors such as the deprivation level in the ward, but is negatively correlated with the employment level in the ward. Without inferring causality, the data supports our previous finding that increased job availability leads

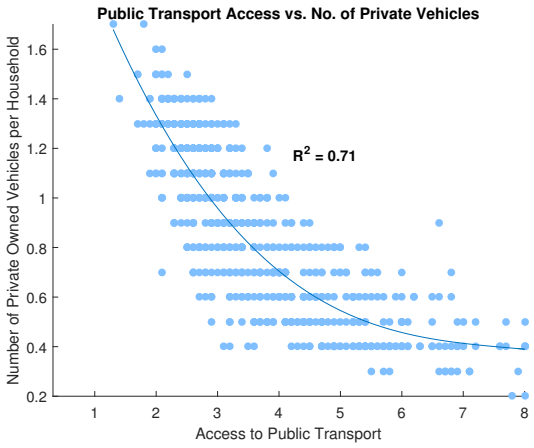

Fig. 5. **Public Transport Access vs. Number of Private Vehicles.** Those with poor public transport access levels (PTALs) own up to 4x more private vehicles per household, and the PTALs explains 71% of the variance in car ownership numbers.

to higher sentiment and a decrease in the percentage of children. We suspect that the wider applicability of this result will depend on the family cultural context.

D. Accessibility to Public Transport

The main attributes in measuring public transport availability in census data is the Public Transport Accessibility Levels (PTAL). It is a detailed and accurate measure of the accessibility of a point to the public transport network, taking into account walk access time and service availability. The method is essentially a way of measuring the density of the public transport network at any location within Greater London. The measure reflects 4 main attributes: (1) walking time to transport access point, (2) reliability of services, (3) number of services, and (4) the average waiting time. It does not consider the speed or utility of the service, crowding effects, and ease or efficiency of interchange. The PTAL methodology was developed for London where a dense integrated public transport network means that nearly all destinations can be reached within a reasonable amount of time. Research using the ATOS (Access to Opportunities and Services) methodology shows that there is a strong correlation between PTALs and the time taken to reach key services i.e., high PTAL areas generally have good access to services and low PTAL areas have poor access to services. Each area is graded between 0 and 6b, where a score of 0 is very poor access to public transport, and 6b is excellent access to public transport.

Fig. 4**b** The accessibility to public transport in a ward has a U-shaped (parabolic) relationship with the

sentiment in the ward (adjusted $R^2 = 0.44$, $P = 4$), such that those with good access to public transport are happy and those who are in areas with poor public transport are also happy (possibly because they rely on personal transportation means), whilst those that are in-between are generally unhappy. Certainly the results in Fig. 5 seem to strongly support this hypothesis. The PTAL values explain for 71% of the variance in the number of private vehicles per household, showing that those with poor public transport access own up to **4 times** more private vehicles per household. Therefore, the availability of public transport explains 44% of the variance in sentiment scores. The wider applicability of this result beyond London is difficult to determine. Yet, we speculate that economies with a high number of privately owned vehicles will exhibit similar patterns, i.e., people are happy when they are either close to public transport or far removed, and struggle when they are in-between the choices.

IV. Conclusions and Discussions

The demographics and landscape of cities are changing rapidly, and there is an emphasis to better understand the factors which influence citizen happiness in order to design smart urban systems. In this paper, we apply natural language processing to 0.4 million geo-tagged tweets in the Greater London area to understand the underlying socioeconomic and urban geography parameters that influence happiness. Our results not only verify established thinking: that job opportunities explain 45-47% of the sentiment variations, but also reveal two additional insights: (1) happiness is negatively correlated with the number of children (accounts for 33% of sentiment variations) and (2) happiness has a U-shaped (parabolic) relationship with access to public transportation (44% of variations). The latter implies that happy people are those who have good access to public transport, or such poor access that they drive (4 times more cars than those who have the best access). The unhappy people are those that rely on, but do not have strong access to public transport. The number of jobs and children, as well as accessibility to public transport are every day facets of urban living (see Table I in **Appendix**) and individually explain up to 47% of the variations in happiness. Our results show that they influence happiness more significantly than more ambient parameters such as degradation, education quality, and crime.

The wider applicability of these results beyond London depends on the context. We expect that the availability of jobs is widely applicable across cultures, whereas the number of children will depend on the culture and the availability to public transport will depend on the ownership level of personal vehicles as well as the culture of transport usage. Future work will focus on creating proprietary sentiment labels for each city by combining meta-data for boosting sentiment analysis accuracy [14]. This will enable large-scale cross-country/city comparisons to be made [4].

The general study of how sentiment is linked to urban features and socioeconomic parameters is useful for urban planners and urban system designers. The results will allow decision makers to move beyond planning services using traditional costbenefit analyses, and enable them to consider the consequences on citizens' happiness. Further research on understanding how these patterns change with different cities and cultures is of interest, as well as how more reliable methods of labelling sentiment to social media data can be applied.

Acknowledgments

The authors would like to acknowledge the EPSRC Doctoral Training Centre (EP/L016400/1), RCUK/EPSRC Grant (EPL023911/1), and the Centre for Competitive Advantage in the Global Economy (CAGE) at the University of Warwick.

Appendix

A linear regression of sentiment vs. ward level socioeconomic and infrastructure metrics is shown in Fig. 6. The linear regression does not uncover more complex parabolic relationships such as those found between sentiment and accessibility to public transportation. Nonetheless it serves as an overview of the first order relationship between all 67 parameters. A categorized table of census data is given in Table I, with challenges that affect citizens daily, yearly, or long-term listed.

TABLE I
Challenges and Factors that Affect Urban Living

Daily / Monthly	*Annual*	*Long Term*
	Pop. Density	Open Space
No. Children	Ethnic Diversity	Fertility
	Rent/Buy, Tax	Housing Types
Jobs	Income/Benefits	Deprivation
	Education	
		Crime
Public Transport	Cars	
	Obesity	Life Expectancy

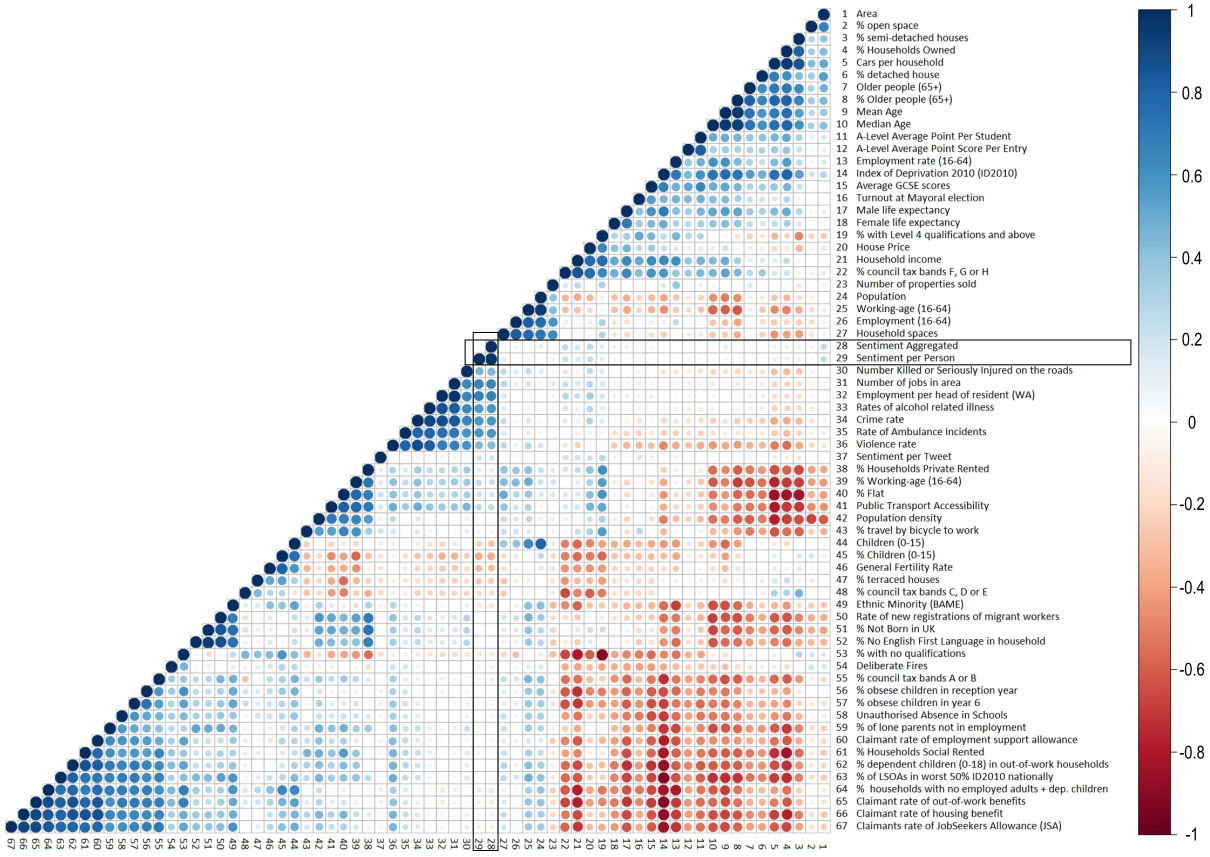

Fig. 6. **Linear Regression Matrix of Sentiment vs. Ward Level Socioeconomic and Infrastructure Metrics.** Sentiment correlations are boxed.

REFERENCES

[1] "World Urbanization Prospects," United Nations, Technical Report, 2014.

[2] L. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G. West, "Growth, innovation, scaling, and the pace of life in cities," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 104, 2007.

[3] L. Bettencourt, "The Origins of Scaling in Cities," *Science*, vol. 340, 2013.

[4] H. Engelbrecht, "Natural capital, subjective well-being, and the new welfare economics of sustainability: Some evidence from cross-country regressions," *Ecological Economics*, 2009.

[5] L. Mitchell, M. Frank, K. Harris, P. Dodds, and C. Danforth, "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place," *PLOS ONE*, vol. 8, 2013.

[6] M. Frank, L. Mitchell, P. Dodds, and C. Danforth, "Happiness and the Patterns of Life: A Study of Geolocated Tweets," *Scientific Reports*, vol. 3, 2013.

[7] E. Kanjo and A. Chamberlain, "Emotions in context: examining pervasive affective sensing systems, applications, and analyses," *Personal and Ubiquitous Computing*, 2015.

[8] E. Qualman, *Socialnomics: How Social Media Transforms the Way We Live and Do Business*. New York, USA: Wiley, 2010.

[9] R. Mitchell and F. Popham, "Greenspace, urbanity and health: Relationships in England," *Journal of Epidemiology and Community Health*, 2007.

[10] T. Lansdall-Welfare, V. Lampos, and N. Cristianini, "Nowcasting the mood of the nation," *Significance*, vol. 9, 2012.

[11] A. Duarte, C. Garcia, G. Giannarakis, S. Limao, A. Polydoropoulou, and N. Litinas, "New approaches in transportation planning: happiness and transport economics," *Economic Research and Electronic Networking*, vol. 10, 2010.

[12] J. Fiaidhi, O. Mohammed, S. Mohammed, S. Fong, and T. H. Kim, "Opinion mining over twitterspace: Classifying tweets programmatically using the R approach," in *ACM Int. Conf. Digit. Inf. Manag. (ICDIM 2012)*, 2012.

[13] M. Hu, B. Liu, and S. M. Street, "Mining and Summarizing Customer Reviews," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.

[14] F. Brave-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis," in *ACM Proceedings on Issues of Sentiment Discovery and Opinion Mining*, 2013.