E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Consumer Data Research Centre

An ESRC Data Investment

2014

e·on

Harnessing the power of social media to determine moving house flows and customers' energy needs.

William Trimble[1], Dr Paul Norman[1], Dr Nick Malleson[1] & Adam Rawson[2]

[1]University of Leeds, [2]E.ON UK plc

Project Background

The extraction and analysis of social media data has enabled researchers to derive new and intriguing datasets on a vast scale. This study extracted text strings (Tweets) from the micro-blogging platform Twitter to form a dataset. Twitter has been used in studies since its conception in 2006, yet it remains as a little used instrument for academic research. Existing research using Twitter datasets include determining the sentiment towards a certain subject matter, such as films and political preference. Other studies have involved epidemiology and earthquake location. The aim of this study is to determine if there is a correlation between Twitter data and E.ON data on the subject of moving house, to encourage the development of using Twitter as a data source. Secondarily, E.ON's online marketing strategies will be assessed.

Data and Methods

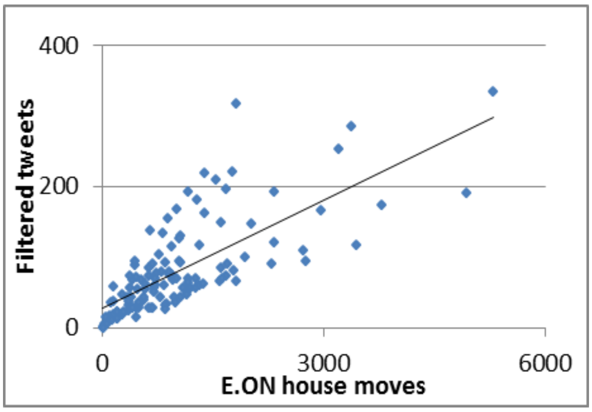
E.ON and Twitter datasets were received covering 3rd June – 25th July 2014. The moving house data supplied by E.ON originated from known customer moves in Great Britain. Geo-located Tweets, derived from a concise search related to moving house, were supplied by the University of Leeds. Both datasets were received and processed in Microsoft Excel. This involved discarding expendable metadata, and geo-locating postcode and place name locations to a coordinate system. A Geographical Information System (ArcGIS) was used to visually present the data on a postcode area scale, used to enable choropleth mapping of comparable regions. Correlation analysis, over time and by area, between the E.ON and Twitter datasets was conducted to fulfil the main aim of this study. Sentiment analysis using an online programme assessed customer attitudes towards E.ON's Twitter page.

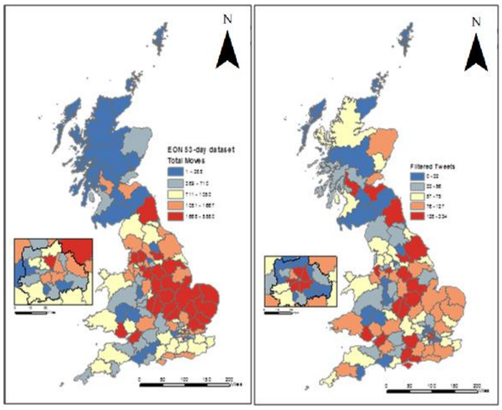
Key Findings

The correlation between the datasets was analysed in two ways. Firstly the correlation over time was calculated by segregating the datasets into eight, one week blocks. The Twitter dataset was separated into raw Tweets – derived from the initial search, and filtered Tweets – derived from manual elimination of Tweets not relating moving house. The Pearson's correlation coefficient found a positive correlation of 0.852 between the raw Tweets and the E.ON house moves data ($p<0.05$). Filtering the Tweets strengthened the correlation to 0.922 ($p<0.05$). The improvement of 0.07 due to filtering demonstrates that more advanced methods of processing could yield greater improvements. Correlation by postcode area involved segregating the datasets by their geographical location. This type of correlation sought to identify if there was a correlation between the datasets on a small geographic scale. The results found a strong positive correlation of 0.73 and 0.74 between the filtered Tweets and raw Tweets respectively ($p<0.05$) (figure 1). The choropleth map (figure 2) shows, despite some fuzziness within the middle quantiles, a broad degree of similarity. Duplicate analysis reveals that 52-54% of areas in the lowest quantile, and 68% of areas within the highest quantile from both datasets matched. Thus this analysis has shown that geographical and time variations both exhibit strong positive correlations between E.ON's real world dataset and a dataset derived from Twitter. Sentiment analysis showed that E.ON ranks third in terms of customer satisfaction on Twitter in relation to the 'big six' UK energy companies. Improvements such as homepage improvements and greater personal interaction were recommended.

Value of the Research

The outcome of this research will allow E.ON to use this study as a platform to develop the use of Twitter datasets. The use of simple methods in this study demonstrates the potential to utilise more advanced methods of data collection, such as a support vector machine (SVM). This study has also developed a personal understanding of how Tweets can be extracted as a data source, and how geographical elements play an essential role.





Figures 1 and 2 – comparing the spatial distribution of home movers with filtered Tweets