‘Good uni: Quality nightlife’. How harvesting tweets opens up a new world of valuable qualitative data

by Blog Admin July 5, 2012

The qualitative data that is freely available on social media platforms has huge potential. Drawing on his research into what Twitter can tell us about the popularity of universities, **Geraint Johnes** writes that Twitter and Facebook messages could be the key to valuable data.

The quantity of qualitative data generated by social networking platforms such as Facebook and Twitter is huge. These data represent a significant resource for researchers in a wide variety of fields. The development of new techniques aimed at harvesting the data in a form suitable for analysis has, over the last decade or so, been rapid and fruitful.

Using the package available here, it is straightforward to retrieve recent Twitter posts (tweets) that contain terms defined in user-specified searches. It is possible to restrict the data collection exercise by time or by geographical location. These tweets can then be examined using a variety of text analysis tools.

Many such tools are available in the R package (text mining). These allow patterns in the text to be identified, and codified so that they are amenable to statistical investigation. An important subset of these tools comprises methods of sentiment analysis – whereby numerical scores can be assigned to tweets as a means of evaluating the strength and direction of the writer’s reaction to certain key words. These methods of analysis commonly make use of lists of words with positive and negative connotations, and make use both of the presence of such words and their position within the structure of a sentence.

There are many uses to which data of this kind could conceivably be put. In the field of behavioural economics, for example, they could be used to evaluate people’s reactions to various news events, providing a direct test of whether or not people respond symmetrically to good or bad news. In geography, they could be used to measure sentiment about different locales. In politics, they could be used to investigate the impact of events on the fortunes of different parties. In marketing, they can be used to monitor the impact of campaigns.

The research potential of these data becomes all the more apparent when we consider how background data on the individuals writing tweets can be obtained. In some respects, the data that twitter collects about its users are very limited – users’ profiles reveal their location, but little else. We do not, for instance, know about gender, age, or socio-economic group. However, a variety of tools can be used to make inferences about such information. Users may load a profile picture of themselves, they may write a biography of themselves, they may link to other pages (blogs, for instance) which contain more data about themselves, and – crucially – there may be clues in the linguistic structure of their tweets. For example, women and men use different vocabularies, and use punctuation and emoticons differently. Likewise, different age groups use different vocabularies.

In recent work, I have been investigating the potential use of Twitter as a source of data with which the popularity of universities can be monitored. Drawing on data harvested from Twitter in November of last year, I have conducted a sentiment analysis of UK higher education institutions. The tweets themselves concern a variety of features of the universities – not all of which are the concern of traditional measures of university performance. For example, one institution is praised in the phrase “good uni; quality nightlife”. Another is criticised because (presumably owing to concerns about copyright infringement and spread of viruses) its computers do not allow the use of torrent services. But, while these tweets do not reflect conventional notions of academic quality, they do have the merit that they capture stakeholders’ perceptions of the institutions in the round. As such, they are likely to produce a picture of universities’ ‘performance’ that is very different from the picture that is painted by other metrics, yet which is important from other perspectives, such as marketing. Insomuch as institutions are interested in monitoring the way in which they are perceived by stakeholders across all aspects of their provision, this approach has the potential to provide useful summary information.

The results for UK universities are instructive. Post-1992 institutions tend to do better than pre-1992

institutions, other things being equal. This may suggest that the post-1992 universities, which are typically less research intensive, are nonetheless successful in finding ways of generating positive responses amongst their stakeholders. These public perceptions may be due to a wide range of stimuli – most obvious of which is a positive student experience. There are also some interesting (broad) regional patterns, with institutions located in the South West scoring best and those in Yorkshire and Humberside scoring the worst. There were no unusual news stories or weather events associated with these regions at the time of data collection, and the regional patterns observed in the data remain unexplained.

An institution that finds that its 'performance' on this measure is less positive than that of its peers may wish to consider why. In so doing, it may wish to drill down into the qualitative data that underpin the summary measures obtained by the sentiment analysis. In some cases, the differential performance may be purely transitory. In other cases, there may be clues in these data about things that the institution could do that would markedly affect its stakeholders' experience of its offering.

While the application of tools of this kind in the context of academic research is still in its infancy, the potential for social networking data to be used in this way is clearly considerable.

Note: This article gives the views of the author, and not the position of the Impact of Social Sciences blog, nor of the London School of Economics.

Related posts:

Do more tweets mean higher citations? If so, Twitter can lead us to the 'personalised journal'; pinpointing more research that is relevant to your interests.

"But who is going to read 12,000 tweets?!" How researchers can collect and share relevant social media content at conferences

Who gives a tweet? Evaluating microblog content gives us an insight into what makes a valuable academic tweet

Political scientists are limited by their reliance on existing data sets, and there is not enough emphasis on creating new data.

We expect to get information in two clicks, why can't we get data as quickly?