Guiding Purposeful Sampling of Tweets using Visual Twitter Analytics (Vista)

Orland Hoeber, University of Regina  
Larena Hoeber, University of Regina  
Laura Wood, University of Windsor  
Ryan Snelgrove, University of Windsor

Twitter is a popular micro-blogging platform that allows individuals and organizations to post short messages in a public, open, and unfiltered medium. As a result of the widespread adoption and the willingness of users to post on wide-ranging topics, Twitter has become a valuable source of information regarding public opinion, citizen reporting, and social interaction. Within many sport communities, Twitter has been embraced and promoted as a mechanism to enhance communication among organizers, management, teams, athletes, fans, and the media (Hambrick, Simmons, Greenhalgh, & Greenwell, 2010; Pegoraro, 2010). From a sport management perspective, Twitter holds a wealth of information that can aid in the understanding of this communication and engagement.

There are a number of challenges that make conducting research with Twitter data rather difficult. Given the short and cryptic nature of the textual element of the tweets, it is often necessary to analyze them manually using traditional content analysis approaches. In addition, retrieving a large number of tweets via the standard Twitter interface is not straightforward, and there is an important temporal aspect to the tweets that must be considered during the analysis. To complicate matters further, for a given topic of interest, it is possible for an extremely large number of tweets to be generated in a very short period of time. As a result, in many cases analyzing Twitter data is a big data problem, exhibiting high volume, velocity, and variety (Russom, 2011).

A common approach among sport management and sport communication research has been to address the big data problems of analyzing tweets using sampling methods. Kassing and Sanderson (2010) and Pegoraro (2010) both used stratified sampling to identify athlete Twitter accounts, and then selected all of the tweets posted during a specific period of time. Hambrick et al. (2010) used stratified random sampling to choose athletes to study, and then selected the most recent 20 tweets posted by these athletes. Blaszka, Burch, Frederick, Clavio, and Walsh (2012) used systematic sampling, choosing every 12th tweet that made use of specific hash tags. While such sampling methods allow the large number of tweets available to be reduced to a collection that can be analyzed manually, new problems are introduced. These include the possibility of missing important or meaningful tweets, the interaction between stakeholders, or the temporal relationships between the tweets and micro-events associated with the phenomenon under investigation.

Visual Twitter Analytics (Vista) was developed as a tool to support the exploration of temporally changing sentiment within sport-related tweets (Hoeber, Hoeber, Wood, Snelgrove, Hugel, & Wagner, 2013). The software extracts data from Twitter based on user-specified queries, performs automatic sentiment analysis (Feldman, 2013), and provides visual timeline representations that allow comparisons of the positive, neutral, and negative sentiments among the tweets. This approach to classifying and visualizing the sentiment was used to analyze a large collection of tweets from the 2013 Le Tour de France, correlating phenomena occurring during the event with changes in the sentiment expressed via Twitter, and contributing to a narrative of the fan experience during the course of the event (Hoeber, Hoeber, Wood, & Snelgrove, 2013).

The purpose of this presentation is to demonstrate a new interactive approach to purposefully sampling Twitter data. We will outline the core features of Vista, along with a number of new features that have been added to support this process. We will discuss how Shneiderman’s (1996) Information Seeking Mantra of “overview first, zoom and filter, then details-on-demand” provides a guiding framework for selecting a purposeful sample based on the interactive discovery of interesting features within the data. An example of this method will be provided, drawing upon over 409,000 tweets collected during the three-week period of the 2013 Le Tour de France making use of the #tdf tag. It will be shown how Vista can provide a visual overview of this large collection of tweets; how high-level patterns within these tweets can be identified with respect to positive, neutral, or negative sentiment; how researchers can zoom into a smaller temporal range in order to study the patterns in greater detail; how the collection of tweets can
be further filtered with sub-queries; and finally how a reasonably small collection of tweets that represent some element of interest discovered during this information-seeking process can ultimately be exported as a sample for further analysis, along with a descriptor of the complex stratification parameters.

The main contributions of this approach beyond other methods that have been used to sample Twitter data are that it is explicitly guided by the researcher via an interactive and dynamic exploratory process, and can be used to purposefully sample Twitter data beyond simple hash tag or user filtering. Additionally, large amounts of Twitter data can be collected and emerging trends can be identified that are worthy of further analysis, rather than specified a priori. The significance of this approach for sport management researchers is that it opens up new avenues for conducting research on Twitter data specifically, and perhaps other large sets of textual data.

References


