Qualitative analysis of Twitter data: Searching for and analyzing discourses of gender in Le Tour de France tweets

Larena Hoeber¹, Orland Hoeber²
¹ Faculty of Kinesiology and Health Studies, University of Regina
² Department of Computer Science, University of Regina

Abstract
Social media platforms, such as Twitter, hold a wealth of knowledge for the study of social issues expressed in the context of sport. Because of the large volume of data and its qualitative nature, it can be difficult to discover and extract relevant tweets for a topic of interest, especially if the topic is emergent. To address this problem, we have developed a visual analytics approach for the exploration and study of Twitter data, called Vista. A case study is used to show how Vista can be used to identify and explore gender discourses within a mega-sport event, providing visual overviews of the patterns of tweet behaviour as well as access to the individual tweets for qualitative analysis.

Keywords: sport, Twitter, search, gender

Citation: Editor will add citation with page numbers in proceedings and DOI.
Copyright: Copyright is held by the author(s).
Acknowledgements: 
Research Data: In case you want to publish research data please contact the editor.
Contact: larena.hoeber@uregina.ca

1 Introduction
Studying the discourses of sport fans provides a valuable window into human behaviour and attitudes toward social issues such as race, ethnicity, gender, sexuality, and class. The opinions sport fans express via social media often contain strong sentiment, and may include topics that would not normally be discussed in person. Further, it has been found that media sport audiences are unique from other audiences in that they are “more actively involved both emotionally and cognitively when consuming media sport” (Frandsen, 2013, p. 21). Among the various social media platforms, Twitter is embraced as an important communications tool for a broad range of stakeholders in sport, including athletes, fans, governing bodies, sponsors, and traditional media outlets (Hambrick et al., 2010; Pegoraro, 2010; Wenner 2014). Thus, discourse analysis of sport-related topics on Twitter can provide a rich lens for identifying and understanding the attitudes, beliefs, and deep-seated biases of such stakeholders (Hardin, 2014).

Twitter has become a popular social networking platform, in which users can easily post short (140 character) messages that are by default public and available for anyone to read. The value of using such data to study public opinion at the intersection of sport and social issues is three-fold: (1) the open, accessible, and unfiltered features of the data, (2) the common practice of tagging important elements of the tweets using hashtags, and (3) the vast amount of data that is available to be explored. While researchers are interested in Twitter data sets because of these features, the analysis can pose significant challenges due to the textual (qualitative) nature of the data, the common use of short and cryptic language, the high volume of data that is available to be analyzed, and the important temporal link between real-world events and the public tweeting about these events.

Because reading and making sense of vast amounts of textual data is time consuming and mentally taxing (even when using qualitative analysis software such as NVivo), many studies in the sport communication and media contexts use sampling strategies (e.g., Blaszka et al., 2012; Hambrick et al., 2010; Kassing & Sanderson, 2010; Pegoraro, 2010). While this reduces the large volume of data to a more manageable amount that can be analyzed using traditional qualitative approaches, the temporal link to real-world events may be lost and important tweets that are highly relevant to the topic of study may be missed (Hoeber et al., 2014). Mahrt and Scharkow (2013) and Tinati et al. (2014) critiqued the inappropriateness of these small-scale approaches for studying large-scale Twitter data sets. Of note, Tinati et al. (2014, p. 6) advocated for the use of “technical capabilities with in-depth qualitative research methods” to enhance the rigor and theoretical development of research involving Twitter data. We propose that such technical capabilities should simultaneously leverage the capabilities of automatic information processing and human judgment (Shneiderman & Plaisant, 2010), and should take a visual
and interactive approach in order to enhance the researcher’s ability to perceive and interpret, reason about, make sense of, and explore within the data (Ward et al., 2010).

2 Purpose
The purpose of this presentation is to demonstrate and discuss a case study of a visual analytics approach for qualitatively analyzing gender discourses within a sport context using Twitter data. The method to be followed and an outline of the case study are provided below.

3 Method
In order to address some challenges of studying Twitter data, we propose taking a visual analytics approach. Combining data processing and machine learning techniques with information visualization and human-computer interaction methods allows for the creation of visual analytics software that supports data exploration, analytic reasoning, information synthesis, hypothesis development and testing, and human decision-making (Keim et al., 2010; Thomas & Cook, 2006). The objectives are to take advantage of the computer’s computation and storage capabilities to extract and infer relevant information, present it to visually to the analyst, and support their cognitive and analytic processes through interactive exploration.

Visual Twitter Analytics (Vista) was developed to support the exploration and study of temporally changing sentiment within sport-related tweets (Hoeber et al., 2013). The software extracts data from the live Twitter stream based on user-specified queries, performs automatic sentiment classification on the tweets (Feldman, 2013), stores these in a database, provides visual timeline representations that allow comparisons of the positive, neutral, and negative sentiments among the tweets, and shows the geographic locations of the tweets. The system automatically extracts the most commonly used hashtags, terms, user mentions, and authors within the collection of tweets, providing a sparkline representation of the distribution of each of these over the temporal range in which the data is shown. The system is highly interactive, supporting temporal zooming, adjustment of the temporal aggregation, tweet inspection, sub-querying, and timeline comparisons. A screenshot of the core interface in the midst of a data analysis scenario is provided in Figure 1.

4 Case Study
We collected over 400,000 tweets during the 2013 Le Tour de France, a three-week men’s cycling event. One of our areas of interest relates to discourses of gender in sport-based Twitter data (Bruce & Hardin, 2014; Meân, 2014). It may be argued that there is nothing particularly unique or special about this event to warrant an examination of gender. However, since sport is a gendered institution (Young & White, 2007), discourses of gender are likely to be expressed in everyday discussions related to the event, athletes, and supporting figures of the event (e.g., wives, girlfriends, fans).

Identifying relevant hashtags and terms, related to gender discourses, before a sport event begins is sometimes difficult because one may not know when those terms or conversations related to Figure 1: Vista Interface. The temporally changing sentiment of tweets over a one-week period of the #tdf dataset (last five days of the race and two days after) highlights the cyclical nature of fan engagement during a multi-day sporting event. Performing a sub-query to explore how women are discussed shows two positive spikes, one neutral spike, and the ongoing discussion of women after the finish of the race.
these concepts will be used during a large-scale, multi-day sport event. However, having access to the entire collection of tweets for an event allows us to interactively explore if gender-specific hashtags and terms were used, and to interactively explore how other related terms were used. The sub-query and content-based visualization features of Vista enable the exploration of the data set for comments related to terms such as ‘men’, ‘women’, ‘boys’, and ‘girls’. The timelines of these tweets can be compared to that of the event in general, providing context and enabling an analysis in relation to micro-events that occurred during the sporting event. For example, a careful examination of Figure 1 reveals that the first positive spike in the use of the term ‘women’ in the context of “#tdf” [Tour de France] occurred after the finish of the race on July 17, but the second spike occurred during the race on July 20. The temporal aspect of the data may lead to further insight into the relationships that exist within the data. Additionally, the system supports the inspection and exploration of the data with respect to the use of hashtags, terms, user mentions, and authors. The sparklines beside each of these provide an overview of when these concepts were discussed within the larger timeline of the data. These can be used to further filter the tweets to focus on a concept of interest that may not have been known at the time (e.g., ‘podium girls’; ‘women’s #TDF’), but emerges from the visualization of the data. Once we identify a relatively small collection of relevant tweets using the interactive filtering mechanism, we can then analyze them individually using qualitative processes, such as coding and categorizing or deconstruction, to examine the nature of the gender discourses.

5 Conclusion
The primary contribution of Vista is the visual and interactive approach to exploring Twitter data sets. By providing researchers with tools to interactively explore a large Twitter data set we can purposely select the most relevant tweets, rather than randomly selecting tweets that may or may not relate to our lines of inquiry. In this particular situation Vista is useful for identifying and exploring gender discourses within a mega-sport event, providing visual overviews of the patterns of tweet behaviour as well as access to the individual tweets for qualitative analysis.

Although Vista offers some tools for exploring and isolating tweets from a large data set, we have encountered some unresolved challenges with the analysis of discourses. These challenges include dealing with the multiple meanings in language and the cryptic nature of tweets due to the 140 character limit. For example, if we were to search for the term ‘race’ among the Tour de France tweets, as indicative of discourses related to a person’s racial identity or ethnicity, the search results would likely include tweets related to the organization of the sport event (e.g., it is organized as a race) or the competitiveness of the event (e.g., athletes racing each other). These challenges are more pronounced when examining the intersectionality of discourses related to gender, race, and sexuality, and hence the reason why we are focusing on gender discourses at this point. Future work could focus on developing tools for delimiting search results (e.g., tweets related to ethnicity but not about the competitiveness of the athletes), thus allowing researchers further opportunities to purposefully select the most meaningful tweets to their lines of inquiry. In addition, deciphering and analyzing comments is complicated because there is little context in which to infer meaning. Providing a mechanism to conveniently review other tweets from the same author may allow for potential ambiguities to be resolved.

References


**Table of Figures**

Figure 1: Vista Interface............................................................................................................ Error! Bookmark not defined.