

Introduction to the Data Analytics, Data Mining and Machine Learning for Social Media Minitrack

Dominique M. Haughton
Bentley University, USA
dhaughton@bentley.edu

Jennifer J. Xu
Bentley University, USA
jxu@bentley.edu

David J. Yates
Bentley University, USA
dyates@bentley.edu

Xiangbin Yan
Harbin Inst. of Technology, China
xyan@hit.edu.cn

Transforming data from social media into useful information, or knowledge, is the focus of this minitrack. The papers at HICSS in 2018 remind our attendees and readers of the many real-world applications of data analytics, data mining, and machine learning for social media. At the fiftieth HICSS, for example, one paper showed how product diversity and service differentiation impact crowd-source provider success. In 2016, we explored critical success factors for online petitions for social causes, and gave voice to other fascinating research topics.

This minitrack begins by exploring predictions made by analyzing more than 1.7 million tweets made during the days leading up to the 2016 national elections to the Upper House in Japan. In Goto & Goto's paper we learn how favorable and unfavorable tweets gathered via social listening can be used for sentiment and predictive analysis. This paper also presents interesting future research directions, e.g., improving the analysis of election data to address overactive social talkers and to merge media posted on Twitter with relevant media from other sources.

The second paper examines hacker behavior in dark forums and identifies its predictors in the light of leadership theory for "communities of practice." Biswas, Mukhopadhyay and Gupta combine techniques from online hacker forum features as well as text mining and sentiment-analysis. Their research demonstrates role-based hacker classification based on a model that is validated using real-world forum data. They also create a novel keyword lexicon of the most significant terms used by hackers derived via the term-frequency-inverse document frequency measure.

Our third paper, "Automated generation of latent topics on emerging technologies from YouTube video content" is by a University of South Florida author team. Daniel and Dutta investigate the efficacy of the Latent Dirichlet Allocation (LDA) algorithm in generating latent topics from discussions related to emerging technologies that have been posted on YouTube. This paper also highlights the utility of data visualization to help tune the value of K as input to the LDA algorithm.

"The 2016 U.S. presidential election on Facebook ...," by Alashri et al., analyzes a dataset of thousands

Facebook posts by candidates and millions of comments to understand the nature of online discourse. Specifically, this paper studies the interaction between political candidates and the public during the 2016 presidential election in the United States. It also outlines a novel method to classify commentators into four groups: strong supporters, supporters, dissenters, and strong dissenters. Finally, comments by each group on policy-related topics are analyzed using sentiment analysis.

The fifth paper, a "Fine grained approach for domain specific seed URL extraction" investigates a novel technique for building domain specific search engines. Mohan et al. propose and evaluate a fine grained approach that includes automatic extraction of seed URLs at the subdomain level, using Wikipedia and Twitter as repositories. Their quantitative evaluation validates their technique for a cybersecurity search engine.

"Yaks versus Tweets: sentiment discrepancy during a social crisis" considers the differences between traditional (Twitter) and anonymous (Yik Yak) social networks during the 'Concerned Student 1950' crisis, which ignited racial tensions at the University of Missouri in 2015. Koohikamali and Gerhart find that people express less total sentiment and more extremity on anonymous social networks. Their results show extremity and length positively influence engagement, but total sentiment negatively influence engagement. These findings provide practical guidance for public safety and law enforcement officials.

Our final paper provides a literature review and proposes research guidance for exploring the role of emotions (vs. facts) in social media. Hyvärinen and Beck find that theories on human emotion are infrequently used to support IS research. Key terms – such as sentiment, emotion and opinion – are not always defined precisely, and sentiment analysis is mostly limited to measuring positivity and negativity rather than considering differentiated emotions. They argue that considering differentiated emotions instead of measuring sentiment, and drawing on theories from other disciplines, is likely to provide valuable insights in IS research.