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

Dominique M. Haughton
Bentley University
dhaughton@bentley.edu

Kevin D. Mentzer
Bryant University
kmentzer@bryant.edu

David J. Yates
Bentley University
dyates@bentley.edu

Abstract

Transforming data from social media into useful information, or knowledge, is the focus of this minitrack. The papers at HICSS in 2020 remind our attendees and readers of the many applications of data analytics, data mining, and machine learning to social media. At the fifty-second HICSS, for example, one paper showed how itinerary recommendations for distributed events (e.g., festivals, conferences, and conventions) might be improved by using publically-available social media about the event combined with other published information (e.g., points of interest and schedules). In 2018, this minitrack explored how text mining and sentiment analysis of messages and threads could be used to determine the level of expertise of cybersecurity hackers participating in online forums, and also gave voice to many other fascinating research topics.

1. Sessions and papers at a glance

The first session begins with our best-paper nominee, *Detecting Political Bots on Twitter during the 2019 Finnish Parliamentary Election*, by a team from Aalto University. This study investigates the influence of social media bots on the Finnish political Twittersphere during the period leading up to the Finnish parliamentary election of 2019. To identify the bots, the authors extend existing models with the use of user-level metadata and state-of-art classification models. Their results demonstrate the authors' model as a suitable instrument for detecting Twitter bots. They find that although there are many bot accounts following Finnish politicians, it is unlikely that this phenomena is a result of foreign entities' attempting to influence Finnish parliamentary elections.

Context Map Analysis of Fake News in Social Media: A Contextualized Visualization Approach is by a team from Virginia Tech and University of Hawaii. This study proposes and evaluates a novel Context Map approach to create a connected network of n-grams by considering the frequency in which they are used together in the same context. They combine network optimization techniques with embedded representation models to generate a visualization interface with clearer and more accurate interpretation potential than previous approaches. Finally, they apply their Context Map method to analyze fake news in social media, exploring news article veracity, political orientation, and context analysis.

Understanding the Mood of Social Media Messages is a collaboration between a team from CSIRO and the Australian War Memorial. Analysis of public opinion and sentiment is usually based on sentiment or emotion processing using machine learning techniques or referencing a curated lexicon of words to measure the emotive intensity being expressed. These approaches can be limited by the sparsity problem, where the lexicon words are not present in the text being processed, and context issues, where the lexicon words have different meanings. This study develops a novel technique based on word embeddings to mitigate these issues and presents a case study showing its application, where the mood expressed on social media is about the Centenary of Armistice in Australia in 2018.

Measuring and Unpacking Affective Polarization on Twitter: The Role of Party and Gender in the 2018 Senate Races is the last paper in the first session. This study classifies Twitter users as Liberal or Conservative to better understand how the two groups use social media during a major election. The authors assess how the Twittersphere felt about in-group party versus out-group party candidates. When they further break these findings down based on the candidates' gender, they find that male senatorial candidates are talked about more positively than female candidates. However, they find that Conservatives talk more positively about female Republican candidates than they do about Republican male candidates. Female candidates of the out-group party are talked about the least favorably.

The second paper session begins with *Generalized Blockmodeling of Multi-Valued Networks* by a team from Sandia National Labs and Cornell University. This research presents an extension to generalized blockmodeling where there are more than two types of objects to be clustered based on valued network data. The authors use the ideas in homogeneity blockmodeling to develop an optimization model to

URI: https://hdl.handle.net/10125/64039
978-0-9981331-3-3

perform the clustering of the objects and the resulting partitioning of the ties so as to minimize the inconsistency of an empirical block with an ideal block. Two case studies are presented: The Southern Women dataset and a larger example using a subset of the IMDb movie dataset.

The New Window to Athletes' Soul – What Social Media Tells Us About Athletes' Performances is by a team from University of St. Gallen. One factor that is shaping the sports industry is the pervasive use of social media. On the one hand, social media is used as a powerful medium for distributing and obtaining news, engaging in topical discussions, and empowering brands. On the other hand, social media has become a platform for athletes to interact with peers, share opinions, thoughts, and feelings. Millions of followers, tweets, and likes later, researchers, practitioners, and athletes ask whether social media has an impact on an athlete's performance. The authors answer this question using professional tennis as the context of their study.

Using Data Analytics to Filter Insincere Posts from Online Social Networks, A Case Study: Quora Insincere Questions is from Texas A&M, San Antonio. Because of their significance in everyday life and public visibility, abuses of Online Social Networks (OSNs) for different purposes are common. Driven by goals such as marketing and financial gain, some users use OSNs to post misleading or insincere content. In this context, the authors utilize a real-world dataset posted by Quora on Kaggle to evaluate different mechanisms and algorithms to filter insincere and spam content. The authors evaluate different preprocessing and analysis models. Furthermore, they analyze the cognitive effort users make in writing their posts and whether that effort can improve the prediction accuracy.

The third and final paper session begins with *Evaluation of VI Index Forecasting Model by Machine Learning for Yahoo! Stock BBS using Volatility Trading Simulation* by a collaboration among Nara Institute of Science & Technology, Ritsumeikan University, and Yahoo Japan. The risk avoidance is very crucial in investment and asset management. One commonly used index as a risk index is the VI index. Suwa et al. (2017) analyzed stock bulletin board messages and predicted it rise. In this study, the authors develop a simulation of trading Nikkei stock index options using intra-day data and verify the validity of the VI index prediction model proposed by Suwa et al. In a period from November 18, 2014, to June 29, 2016 using a long straddle strategy.

Success Factors of Donation-Based Crowdfunding Campaigns: A Machine Learning Approach is by a team from University of Texas, Rio Grande Valley and Dakota State University. This study identifies key drivers of donation-based crowdfunding campaign success using a machine learning approach. Based on an analysis of crowdfunding campaigns from Gofundme.com, the authors show that their models are able to predict the average daily amount received at a high level of accuracy using variables available at the beginning of the campaign and the number of days it has been posted. Among the six machine learning algorithms the authors compare, support vector machine (SVM) performs best in predicting campaign success.

Identifying Citation Sentiment and its Influence while Indexing Scientific Papers is a by a team from Rutgers University and University of Washington. For scientific papers it is often assumed that the sentiment associated with citation instances is inherently positive. As a result, most of the existing indexes focus only on the frequency of citation. In this paper, the authors highlight the importance of considering sentiment of citation while calculating ranking indexes for scientific literature. Using various baselines, the authors analyze the impact of their index on the Association for Computational Linguistics (ACL) Anthology collection of papers. This study thus contributes toward building a more sentiment-sensitive ranking that better captures the influence and usefulness of research papers.

As minitrack co-chairs, we are delighted that this year's papers hail from such a wide range of disciplines and utilize a variety of methodologies used by contemporary social media scholars.

2. Future directions in the field

As scholars in this area, we see quite a bit of research that focuses on Twitter. We encourage future authors to explore alternative social media platforms. This effort should not be limited to what is traditionally thought of as social media (Facebook, Twitter, and Instagram being three of the largest platforms), but should explore how technology-enabled conversation is unfolding in other areas such as embedded in video games as well as on video sharing sites such as YouTube and TikTok. Since these platforms appeal to different audiences, we expect that in-depth or cross-platform studies would be both interesting and rich. Moreover, research that explores the use of video, images, and other non-textual elements could lead to new and exciting discoveries.

Finally, with models of value relying on machine learning and artificial intelligence (AI), we would be negligent to simply blame the "algorithms" for bias or social injustice, whether it be discriminatory lending, police profiling, etc. How we collectively respond to these and other challenges will greatly influence the adoption of "intelligent" models as well the level of fairness and equity built into the underlying designs. Unpacking and understanding machine learning and AI algorithms and also the consequences of their use is therefore of vital importance.