Introduction to Data, Text, and Web Mining for Business Analytics Mini-track

Dursun Delen

Oklahoma State University

Hamed M. Zolbanin

University of Dayton

Behrooz Davazdahemami

University of Wisconsin-Whitewater

This mini-track has a total of six accepted papers that are about developing analytics systems for decision support by means of data, text, or web mining. The papers address the gaps from a variety of relevant methodological and practical contexts such as electronic word of mouth, digital economy, fake news detection, social science, and analogical reasoning.

Wood et al. present a computational method that allows researchers and policymakers to identify and understand new business models within the digital economy and track their evolution through analysis of the firms' webpages using clustering methods. The proposed method comprises three distinct components: a tool for scraping content from websites; a clustering mechanism; and an approach for generating a taxonomy from the clusters identified. The authors have showcased their method through analysis of a sample set of websites from the West Country's 150 biggest businesses.

Denisiuk et al. in their study focus on the next generation of office assistants. Particularly, they investigate how named entities can be recognized in a Robotic Process Automation (RPA) through converting screenshots imagery into text and then applying Named Entity Recognition (NER) to extract features that can be used by RPA. The proposed pipeline introduces features that describe the local context of an n-gram, which is shown (through a series of experiments) to be very important for the recognition and categorization of the entity.

In another paper of this mini-track, *Xiao et al.* develop a novel fake news detection framework called Fake News Engagement and Propagation Path (FNEPP) by combining news content, user

engagement, user characteristics, and the news propagation path. The proposed framework comprises two collaborative modules – the engagement module, which captures news content and user engagement, and the propagation path module, which learns global and local patterns of user characteristics and news dissemination patterns. By applying the FNEPP on two real-world datasets, it is shown that it performs superior to the alternative existing methods (DTR, SVM-TS, GRU, and CSI), especially for the detection of fake news in the early stages.

Regarding the inability of current artificial intelligence (AI) systems to learn beyond their comprehension, *Combs et al.* highlight analogical reasoning (AR) as a method to address this limitation. In current AR systems, symbolist, connectionist, or hybrid approaches are usually used to evaluate analogies; however, it is unclear how psychologically-inspired algorithms perform in comparison to NLP-produced algorithms. Using a defined set of metrics (e.g., correctness and prediction scores), this study finds that psychologically-based models have an advantage over NLP-produced algorithms. The authors point out, however, that there is not a clear one-size-fits-all algorithm for all AR problems.

Lu et al. compare the differences between the metrics derived from user-generated content (UGC) on social media and non-social media platforms and examine whether and how such metrics can be used as indicators for operational metrics. They find that topic distributions and weighted service quality (WSQ) vary significantly between the two types of platforms. In addition, the WSQ scores are significant indicators of objective service quality.

Finally, *Yuan et al.* employ natural language processing to measure a new, inherent personality trait: Sensory Processing Sensitivity (SPS). The authors collect both text-based and survey data from 241 participants and find almost perfect agreement among the different methods they use in categorizing a highly sensitive versus non-highly sensitive person. By evaluating the important predictors, they conclude certain features could have great potentials in measuring SPS in written text.