

Introduction to Data, Text, and Web Mining for Business Analytics Minitrack

Dursun Delen
Oklahoma State University

Hamed Zolbanin
University of Dayton

Behrooz Davazdahemami
University of Wisconsin-Whitewater

This mini-track has a total of six 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 data labeling, anomaly detection, automatic procedure generation, ecosystem intelligence, and causality model extraction.

In the first paper, *Bainiaksinaite et al.* develop a framework for automatic data labeling, which is used to create large amounts of annotated data to train domain-specific detection classifiers for financial news events. This framework employs a rule-based approach to annotate the training data and offers two main advantages: Unlike traditional data labeling methods, it helps to filter the relevant news article from the noise; and, it allows an easier transferability to other domains and a better interpretation of models trained on automatically labeled data. The study validates the proposed model using news events from the US Apparel and Footwear industry.

In an application of image mining for ecosystem intelligence, *Basole* combines image recognition, graph modeling, and visualization to introduce a human-assisted knowledge discovery approach for *logomaps*. A logomap is a composite infographic consisting of corporate logos, market segments, and logical hierarchies, typically created by investors and analysts, that seeks to communicate complex, emerging industries to a broader audience. The study illustrates that even a subset of logomaps can provide insights into patterns and trends, including commonality of a core set of firms, disjoint consideration and exclusion of a large portion of the ecosystem, and clustering of market segments.

Ogawa and *Saga* propose a method for constructing a causality model from review data. Textual reviews include evaluation factors, and causality model extraction from such data is important for understanding the evaluation factors and their relationships. While several methods are available for extracting causality models, such as those based on hierarchical latent Dirichlet allocation, the depth of each topic in such hierarchical structures is forcefully pruned even when granularities are different for each

topic. This complicates the interpretation of a hierarchical topic structure. This study develops a hierarchical topic structure with different depths by using Bayesian rose trees and illustrates its accuracy and interpretability using real data.

The paper by *Rojas et al.* uses process mining to detect anomalies in business processes. Anomaly detection refers to discovering behaviors that are not typical or expected in the business process and helps in preventing intrusion and other risks in companies. The study uses a real-world event log from an ITIL-covered incident management process to discuss anomaly detection approaches in business processes. The authors conclude that autoencoders are able to remove the most anomalous behavior from event logs in a way that facilitates the discovery of high-quality process models.

The article by *Geluykens et al.*, after formalizing the problem of procedure generation, presents a novel approach utilizing natural language processing to procedural knowledge recipes. Their approach relies on neural machine translation and the BART model. The proposed approach is showcased by generating cooking recipes given a list of available ingredients. In addition to the existing evaluation metrics for procedure generation performance, the authors have also developed two novel metrics and used them to show the superiority of their proposed approach over a few previous approaches from the literature.

Landolt et al. propose a holistic framework and use it to develop a taxonomy of deep learning techniques in natural language processing applications. Conducting a systematic literature review and inspired by the CRISP-DM methodology, the paper suggests that a deep learning NLP approach can be distinguished by five dimensions, namely NLP application, network architecture, type of embedding, learning technique, and performance evaluation. Multiple characteristics are developed for each dimension to ease up the application of the proposed taxonomy to a variety of DL NLP studies. Lastly, the application of the proposed approach is showcased on a recent relevant study in detail.