Bryant University

Bryant Digital Repository

1-5-2021

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

Jeffry S. Babb
West Texas A&M University, jbabb@wtamu.edu

Kevin D. Mentzer
Bryant University, kmentzer@bryant.edu

David J. Yates
Bentley College, dyates@bentley.edu

Follow this and additional works at: https://digitalcommons.bryant.edu/cis-conf-proceedings

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

Jeffry S. Babb
West Texas A&M University
jbabb@wtamu.edu

Kevin D. Mentzer
Bryant University
kmentzer@bryant.edu

David J. Yates
Bentley University
dyates@bentley.edu

Abstract

Social media continues to influence the way we work, play, and interact with each other. We are on the other side of a U.S. presidential election with a sitting president who has clearly shown the power of utilizing social media to drive public discourse. Never before has a U.S. president communicated more directly with the public. Our online communities continue to struggle with the conflicting challenge of being free and open, allowing anyone to express their thoughts and ideas, while at the same time working to combat misinformation and deep fakes. We praise social media when it is used as a tool to rally "our side" or confirm our bias while condemning it when it is used by groups we believe are harmful to the public good. In the midst of all of this, social media scholars continue on various paths of sense-making to better understand the impact social media has on society. Finally, we note that this minitrack is broad and inclusive. We welcome papers on techniques and tools to transform unstructured data into information that can be used to better drive theory, while also accepting papers on the impact that social media has on our society.

1. Sessions and papers at a glance

Overall, this minitrack contributes five papers to the fifty-fourth HICSS, reflecting a 35.7% acceptance rate.

Improving News Popularity Estimation via Weak Supervision and Meta-active Leaning [1] is by a team from the University of Alberta. This work proposes and evaluates a framework by which an active learning selection policy is shown to predict the long-term popularity of mobile news articles without the need for ground-truth observation. The framework utilizes weakly supervised labels from user-defined heuristics to generate initial popularity estimates. It then applies a selection policy to repair these weak predictions. The framework also queries end users before the final model is produced. Through a series of experiments utilizing three real-world datasets, the authors show that their framework improves prediction performance

by 28% while reducing the volume of required ground truth by 32%.

The authors from Montclair State University behind our next paper, *User Demographics and Censorship on Sina Weibo* [2], heeded our call from 2020 to expand beyond social media platforms that have dominated the literature (namely Twitter and Facebook) and instead used Sina Weibo data to better understand who might be targeted for censorship on that platform. They find that gender, whether an account is verified (the user pays for mobile and security features) or not, and the user's geographic location all seem to play a role in censorship. Their results suggest that verified male users are more likely to be censored than females or users who are not verified. And users from provinces such as Hong Kong, Macao, and Beijing are censored at a rate that is almost the same as posts from overseas (see Figure 1).
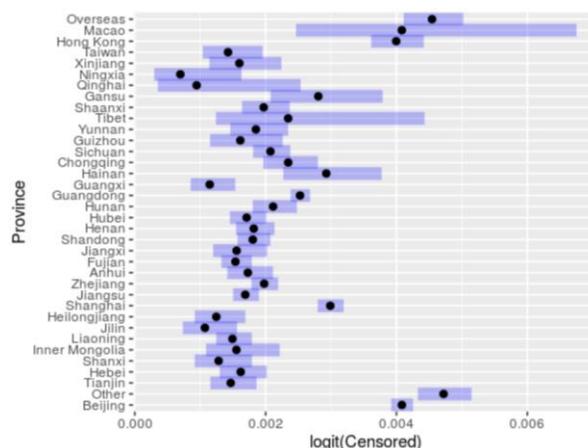
Figure 1. Marginal Effect that a Post is Censored Based on Location (Source [2]).

The topic of COVID-19 was introduced to this minitrack by authors from the National University of Ireland Maynooth in the paper *Inferring the Relationship between Anxiety and Extraversion from Tweets during COVID-19 – A Linguistic Analytics Approach* [3]. Through the use of linguistic analytics, the authors were able to understand personality traits and trait anxiety to better understand the impact

COVID-19 is having on a set of individuals. Their findings confirm previous studies, which show that extraverted individuals in normal times experience lower state anxiety than more introverted individuals. However, there seemed to be no significant differences between individuals after the pandemic announcement in March 2020. These findings suggest that the COVID-19 pandemic affects individuals regardless of their extraversion trait disposition. Furthermore, our collective anxiety is converging to a steady state as we adjust to a "new normal" during the pandemic.

In our fourth paper, *A Group Recommendation Model using Diversification Techniques* [4], colleagues from Universidade Federal da Bahia improve upon current models of recommendation systems. Instead of relying on a group preferences matrix, the authors use diversification techniques to exploit various aggregation techniques resulting in improvements over established models. Using the algorithm shown here, they evaluated diversity and accuracy goals for group recommendations. Their results demonstrate that their proposed approach achieves a 1.8% improvement in diversity and yields a 3.8% improvement in precision.

Our fifth and final paper, titled *Predicting Question Deletion and Assessing Question Quality in Social Q&A Sites using Weakly Supervised Deep Neural Networks* [5], comes to us from San José State University. By learning from past deletions by site monitors, this work trains a neural network to predict question quality. Ghosh also explores the efficiency of different input representations, optimization functions, and neural network models for predicting question quality. The results highlight that combining natural language features with word embeddings can result in better performance (higher recall and F-scores) than word embeddings alone. When assessing question quality, his model obtained a total positivity rate of 0.841 and a precision of 0.514. This research, which uses an LSTM model, represents an important step towards automating assessing the quality of questions posed by community members on a social Q&A site.

2. Future directions in the field

The world is experiencing a Coronavirus surge, which continues to shape society in ways that will take years to unfold. With classrooms limiting or eliminating the face-to-face experience, with work and entertainment becoming remote, and with social activities being severely restricted, we will likely see a significant number of papers emerge on the impact of these drastic societal shifts. How are people responding to lockdowns? What is the impact of the COVID-19 pandemic on short- and long-term psychological health, and how can we utilize signals from social media to better understand these difficult topics? How has social media changed due to new forms of electronically enabled human interaction that have emerged to replace our face-to-face interactions? To this end, a partial list of topics we encourage scholars to consider for future submissions to this minitrack include:

- Analysis of emerging social media platforms
- Discovery, collection, and extraction of social media data
- Unstructured data mining of digital or social media
- Identification of, and response to, social media manipulation
- Impact of online privacy policies on data collection or use
- Opinion mining, sentiment analysis, and recommendation analysis
- Identifying and profiling influential participants, subgroups, and communities
- Online reactions to offline events
- Cloud computation on social media data
- Predictive and forecasting analytics based on social media content
- Trend analysis to identify emerging topics, ideas, and shifts
- Visual analysis of online media structure, usage, and content
- Comparison of structured vs. unstructured techniques
- Social search, retrieval, and ranking
- Social innovation and social entrepreneurship through digital media

References

[1] M. Nashaat and J. Miller, "*Improving News Popularity Estimation via Weak Supervision and Meta-active Leaning*," in Proc. of 54th Hawaii International Conference on System Sciences, Wailea, HI, USA, January 2021.

[2] W. Kenney and C. S. Leberknight, "*User Demographics and Censorship on Sina Weibo*," in Proc. of 54th Hawaii International Conference on System Sciences, Wailea, HI, USA, January 2021.

[3] D. Gruda and A. Ojo, "*Inferring the Relationship between Anxiety and Extraversion from Tweets during COVID-19 – A Linguistic Analytics Approach*," in Proc. of 54th Hawaii International Conference on System Sciences, Wailea, HI, USA, January 2021.

[4] A. Chagas de Oliveira and F. Araújo Durão, "*A Group Recommendation Model Using Diversification Techniques*," in Proc. of 54th Hawaii International Conference on System Sciences, Wailea, HI, USA, January 2021.

[5] S. Ghosh, "*Predicting Question Deletion and Assessing Question Quality in Social Q&A Sites using Weakly Supervised Deep Neural Networks*," in Proc. of 54th Hawaii International Conference on System Sciences, Wailea, HI, USA, January 2021.