Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

# Text Mining in Big Data Analytics

Derrick L. Cogburn
American University
dcogburn@american.edu

Michael J. Hine
Carleton University
mike.hine@carleton.edu

Normand Peladeau
Provalis Research
Peladeau@provalisresearch.com

Victoria Y. Yoon
Virginia Commonwealth U.
vyyoon@vcu.edu

## Abstract

*This mini-track recognizes the reality that global collaboration systems, social media, and information systems of all types, generate enormous amounts of unstructured textual data, including: system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this unstructured textual data is readily available, it presents tremendous challenges to researchers trying to analyze these large bodies of text with traditional methods. Text mining in big data analytics is an increasingly important technique for an interdisciplinary group of scholars, practitioners, government officials, and international organizations. For example, the American Association for the Advancement of Science (AAAS) launched a new competition in 2014 on Big Data and Analytics within its highly competitive senior executive branch fellowship program.*

## 1. Introduction

Building on the success of the inaugural "Text Mining in Big Data Analytics" mini-track at HICSS-50, we are pleased to introduce the selected papers for the second iteration of our mini-track. The mini-track is built on the successful HICSS tutorials on Text Mining we have organized since HICSS 48. At HICSS 50, we had over 71 registered participants for the text mining tutorial, including 19 doctoral students. We had over 20 participants in the mini-track and accepted four excellent papers, which were well received by the participantsThis mini-track recognizes the reality that global collaboration systems, social media, and information systems of all types, generate enormous amounts of unstructured textual data, including: system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this unstructured textual data is readily available, it presents tremendous challenges to researchers trying to analyze these large bodies of text with traditional methods. Text mining in big data analytics is an increasingly important technique for an interdisciplinary group of scholars, practitioners, government officials, and international organizations. For example, the American Association for the Advancement of Science (AAAS) launched a new competition in 2014 on Big Data and Analytics within its highly competitive senior executive branch fellowship program.

## 2. Minitrack Topics and Themes

The mini-track on Text Mining in Big Data Analytics is designed to provide an interactive forum by bringing together researchers to discuss the critical issues of text mining and to contribute to the growing big data focus at HICSS, and invites papers that apply text-mining approaches to a wide variety of substantive domains, including, but not limited to theoretical and applied approaches to analyzing various genres of textual data:

- Blog posts
- Social media analysis
- Email archives
- Published articles
- Websites
- Meeting transcripts
- Speeches
- Online discussion forums
- Online communities
- Computer logs

And addressing methodological challenges, such as:

- Automated acquisition and cleaning data
- Working on distributed, high-performance computers
- Overcoming API limitations
- Using LDA, LSA, and other techniques

- Robust Natural Language Processing (NLP) techniques
- Text summarization, classification, and clustering.

As co-chairs of the HICSS Text Mining Mini-Track, we are pleased with the results of this initial offering. We have accepted four papers that highlight various important aspects of this emerging community.

## 3. Paper 1: From Facebook to the Streets: Russian Troll Ads and Black Lives Matter Protests

Our first paper focuses on understanding the relationship between social media and political mobilization. It analyzes the dynamics of the Russian State's coordinated trolling campaign against the United States beginning in 2015. Using the May 2018 release of all Russian Troll Facebook advertisements, this study constructs a topic model of the content of these ads. The relationship between ad topics and the frequency of Black Lives Matter protests is examined.

## 4. Paper 2: An Investigation of Predictors of Information Diffusion in Social Media: Evidence from Sentiment Mining of Twitter Messages

Social media have facilitated information sharing in social networks. Previous research shows that sentiment of text influences its diffusion in social media. Each emotion can be located on a three-dimensional space formed by dimensions of valence (positive–negative), arousal (passive / calm–active / excited), and tension (tense–relaxed). While previous research has investigated the effect of emotional valence on information diffusion in social media, the effect of emotional arousal remains unexplored. This study examines how emotional arousal influences information diffusion in social media using a sentiment mining approach. This paper proposes a research model and tests it using data collected from Twitter.

## 5. Paper 3: Analyzing Trends and Topics in Internet Governance and Cybersecurity Debates Found in Twelve Years of IGF Transcripts

Internet Governance research generates substantial and innovative, interdisciplinary global scholarship. What are key topics and themes in this research area, and how do they relate to cybersecurity? This paper answers these questions by analyzing transcripts from twelve years of the UN Internet Governance Forum (IGF), asking: (1) What key themes, topics, and entities are discussed at IGF? (2) Which issues have remained consistent at IGF, and which have changed? And (3) to what extent is the NIST Cybersecurity Framework represented at IGF? The CRISP-DM approach to text mining shows human rights as the most dominant IGF theme, followed by freedom of expression, with disability being a persistent issue. During entity extraction cybersecurity emerges prominently, as does blockchain and IoT. Topic Modeling illustrates the resilience of human rights, but also identifies the IANA transition, accessibility, and "fake news." Finally, the NIST cybersecurity framework is represented clearly in the data.

## 6. Paper 4: Towards Computational Assessment of Idea Novelty

The final paper in the minitrack explores a way to assess the novelty of ideas in textual data. In crowdsourcing ideation websites, companies can easily collect large amount of ideas. Screening through such volume of ideas is very costly and challenging, necessitating automatic approaches. It would be particularly useful to automatically evaluate idea novelty since companies commonly seek novel ideas. Three computational approaches were tested, based on Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and term frequency–inverse document frequency (TF-IDF), respectively. These three approaches were used on three set of ideas and the computed idea novelty was compared with human expert evaluation. TF-IDF based measure correlated better with expert evaluation than the other two measures. However, the results show that these approaches do not match human judgement well enough to replace it.

## 7. Towards a Text Mining Community

We believe this new mini-track has great potential to stimulate the creation of a robust, interdisciplinary text mining research community within HICSS. Given the amount of unstructured textual data generated by widespread collaboration systems and technologies, such a research community would be invaluable. The text mining papers at this 52nd Anniversary HICSS Conference represent what we see as an important emergent trend, which we believe will remain for many years to come.