Text 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

Global collaboration systems, social media platforms, and information systems of all types, generate voluminous unstructured textual data. Genres of these data include system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this quantity, quality, and diversity of data is a researcher's dream, for many analysts, this unstructured data presents tremendous challenges. This is especially for researchers limited to traditional qualitative methods. Computational text mining and big data analytics is an increasingly important technique for an interdisciplinary group of scholars, practitioners, government officials, and international organizations. These techniques are being used by a wide range of interdisciplinary researchers, represented in part by the papers in this year's Text Analytics minitrack.

1. Introduction

Computational text mining continues to grow in importance to scholarly researchers and practitioners alike. This growth is driven in part by the continued availability of voluminous amounts of text-based data, flowing from collaboration systems and information systems of all types, and from the availability of powerful open source and commercials computational tools.

The annual HICSS program has contributed substantially to this growth. For the past several years, we have contributed to the emerging HICSS theme on big data and analytics, with both our annual tutorial on text mining opportunities and challenges, and our minitrack on big data analytics in text mining.

This year at HICSS 53, we are pleased to present six papers selected papers for the this minitrack through our highly competitive processes. These six papers include a range of contributions which include both methodological innovations, and an exploration of substantively interesting subject matter.

2. Minitrack Topics and Themes

The minitrack on Text Analytics brings together a global community of interdisciplinary researchers to discuss technological and methodological innovation in computational text mining through an interactive examination of theoretical and applied papers in a wide variety of substantive domains, including, but not limited to, analysis of 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 Analytics Minitrack, we are pleased with the results of our call for papers. We have accepted six papers that highlight various important aspects of this emerging community, which will be presented over two sessions.

URI: https://hdl.handle.net/10125/63831
978-0-9981331-3-3

3. Paper 1: Trends in U.S. Foreign Policy Prioritizations, 2000-2019

In our initial paper, Frederic Lestina uses the open source software language R, to demonstrate several ways to exploit a critical source of US government data, congressional transcripts. This paper examines trends in U.S. foreign policy priorities by recent U.S. presidents, using transcripts from Congressional foreign appropriations committees from 2000 to 2019. Textual analysis of the transcripts shows a divergence in distribution of key phrases, suggesting a possible shift in foreign policy focus by president. Differences in key phrases were also found during the two terms of the Bush and Obama presidencies, suggesting a shift in foreign policy priorities even under the same president. Although the limitations of this paper's methodology preclude finding any conclusive shift in foreign policy priorities by president, this paper demonstrates the feasibility of applying basic text-mining techniques in answering social science questions where data can be found in text-based sources.

4. Paper 2: Supporting Interview Analysis with Autocoding

Our second paper, by Andreas Kaufmann, along with co-authors Ann Barcomb and Dirk Riehle, take us deeper into our methodological exploration in this minitrack. Interview analysis is a technique employed in qualitative research. Researchers annotate (code) interview transcriptions, often with the help of Computer-Assisted Qualitative Data Analysis Software (CAQDAS). The tools available today largely replicate the manual process of annotation. In this article, the authors demonstrate how to use natural language processing (NLP) to increase the reproducibility and traceability of the process of applying codes to text data. They integrated an existing commercial machine--learning (ML) based concept extraction service into an NLP pipeline independent of domain specific rules. They then applied their prototype in three qualitative studies to evaluate its capabilities of supporting researchers by providing recommendations consistent with their initial work. Unlike rule-based approaches, this process can be applied to interviews from any domain, without additional burden to the researcher for creating a new ruleset. This work using three example data sets shows that this approach shows promise for a real—life application, but further research is needed.

5. Paper 3: Towards an Integrative Approach for Automated Literature Reviews Using Machine Learning

In the last paper of our first session, Christoph Tauchert and co-authors Marco Bender, Neda Mesbah, and Peter Buxmann continue our methodological focus, by highlighting the role machine learning can play in automating the process of developing integrative literature reviews. Due to a vast amount of scientific publications which are mostly stored as unstructured data, complexity and workload of the fundamental process of literature reviews increase constantly. Based on previous literature, these authors develop an artifact that partially automates the literature review process from collecting articles up to their evaluation. This artifact uses a custom crawler, the word2vec algorithm, LDA topic modeling, rapid automatic keyword extraction, and agglomerative hierarchical clustering to enable the automatic acquisition, processing, and clustering of relevant literature and subsequent graphical presentation of the results using illustrations such as dendrograms. Moreover, the artifact provides information on which topics each cluster addresses and which keywords they contain. The authors evaluate their artifact based on an exemplary set of 308 publications. Their findings indicate that the developed artifact delivers better results than previously known approaches and can be a helpful tool to support researchers in conducting literature reviews.

6. Paper 4: Using Computational Text Mining to Understand Public Priorities for Disability Policy Towards Children in Canadian National Consultations

In our second session, we begin with a paper by minitrack chair Derrick L. Cogburn, and co-authors, Keiko Shikako-Thomas, Jonathan Lai who present a model for extracting public policy preferences from transcripts of national government led consultations for complex policy domains. In this paper, the policy domain under review is national and international disability policy. Identifying policy preferences from public consultations presents a challenge to national and local governments. Computational text mining approaches provide a useful strategy for analyzing the large-scale textual data emerging from these policy processes. In this study, we developed an inductive and deductive text mining approach to understand disability-related policy priorities. This approach is applied to data from the nationwide disability policy consultation conducted in 2016 by the Government of Canada. This process included 18 town hall meetings, 9 thematic roundtables, and online submissions from 92 stakeholders. Transcripts of these consultations were made available to researchers. Three broad research questions were asked of this data, focused on key themes; differences by city size and type of consultation; and impact of two global policy frameworks. The study identified a number of key themes and saw differences by city size. The study also identified content related to both the CRPD and CRC.

7. Paper 5: Non-Exhaustive, Overlapping, k-mediods for Document Clustering

Next, Eric Kerstens helps us understand the role non-exhaustive, overlapping, k-mediods can play in document clustering. Manual document categorization is time consuming, expensive, and difficult to manage for large collections. Unsupervised clustering algorithms perform well when documents belong to only one group. However, individual documents may be outliers or span multiple topics. This paper proposes a new clustering algorithm called non-exhaustive overlapping k-medoids inspired by k-medoids and non-exhaustive overlapping k-means. The proposed algorithm partitions a set of objects into k clusters based on pairwise similarity. Each object is assigned to zero, one, or many groups to emulate manual results. The algorithm uses dissimilarity instead of distance measures and applies to text and other abstract data. Neo-k-medoids is tested against manually tagged movie descriptions and Wikipedia comments. Initial results are primarily poor but show promise. Future research is described to improve the proposed algorithm and explore alternate evaluation measures.

8. Paper 6: An Investigation of Predictors of Information Diffusion in Social Media: Evidence from Sentiment Mining of Twitter Messages

Finally, in our nominee for best paper, we present a paper by Mohammad Salehan and Dan Kim which demonstrates the ability to exploit an increasingly popular, albeit challenging, data source, Twitter. In this paper, the authors explore predictors of information diffusion through a sentiment analysis of Twitter data. 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. We propose a research model and test it using data collected from Twitter.

9. Towards a Text Mining Community

This minitrack has generated great interests and along with our annual tutorial on problems and opportunities in text mining and big data analytics, this minitrack has the potential to help support 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.