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

URI: http://hdl.handle.net/10125/49961
ISBN: 978-0-9981331-1-9
(CC BY-NC-ND 4.0)

- 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: On the Patent Claim Eligibility Prediction Using Text Mining Techniques

With the widespread of computer software in recent decades, software patent has become controversial for the patent system. Of the many patentability requirements, patentable subject matter serves as a gatekeeping function to prevent a patent from preempting future innovation. Software patents may easily fall into the gray area of abstract ideas, whose allowance may hinder future innovation. However, without a clear definition of abstract ideas, determining the patent claim subject matter eligibility is a challenging task for examiners and applicants. In this research, in order to solve the software patent eligibility issues, we propose an effective model to determine patent claim eligibility by text-mining and machine learning techniques. Drawing upon USPTO issued guidelines, we identify 66 patent cases to design domain knowledge features, including abstractness features and distinguishable word features, as well as other textual features, to develop the claim eligibility prediction model. The experiment results show our proposed model reaches the accuracy of more than 80%, and domain knowledge features play a crucial role in our prediction model.

4. Paper 2: Enhancing Scientific Collaboration Through Knowledge Base Population and Linking for Meetings

Recent research on scientific collaboration shows that distributed interdisciplinary collaborations report comparatively poor outcomes, and the inefficiency of the coordination mechanisms is partially responsible for the problems. To improve information sharing between past collaborators and future team members, or reuse of collaboration records from one project by future researchers, this paper describes systems that automatically construct a knowledge base of the meetings from the calendars of participants, and that then link reference to those meetings found in email messages to the corresponding meeting in the knowledge base. This is work in progress in which experiments with a publicly available corporate email collection with calendar entries show that the knowledge base population function achieves high precision (0.98, meaning that almost all knowledge base entities are actually meetings) and that the accuracy of the linking from email messages to knowledge base entries (0.90) is already quite good.

5. Paper 3: Text Mining Narrative Survey Responses to Develop Engagement Scale Items

A sixteen-item employee engagement scale was supplemented with items developed from literature review, from related scales, and from text mining narrative responses to an open-ended question about employee performance. The text mining procedure is described and may be useful to other scale developers. Possible modifications and extensions of the method are suggested. Some items derived from text mining performed as well as those developed using traditional methods.

6. Paper 4: Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction: A Lesson of History

Topic modeling is often perceived as a relatively new development in information retrieval sciences, and new methods such as Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation have generated a lot of research. However, attempts to extract topics from unstructured text using Factor Analysis techniques can be found as early as the 1960s. This paper compares the perceived coherence of topics extracted on three different datasets using Factor Analysis and Latent Dirichlet Allocation. To perform such a comparison a new extrinsic evaluation method is proposed. Results suggest that Factor Analysis can produce topics perceived by human coders as more coherent than Latent Dirichlet Allocation and warrant a revisit of a topic extraction method developed more than fifty-five years ago, yet forgotten.

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 50th Anniversary HICSS Conference represent what we see as an important emergent trend, which we believe will remain for many years to come.