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

This virtual minitrack for HICSS-55 recognizes that most global collaboration systems, social media, and information systems of all types, generate enormous amounts of textual data, including system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this structured, semi-structured, and 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. This minitrack explores the tools, techniques, and insights generated from the analysis of text data.

1. Introduction

Building on the success of our annual tutorial on Text Analytics and the corresponding minitrack, we are pleased to introduce the five papers selected for the minitrack on Text Analytics at the virtual HICSS-55. Global collaboration systems and information systems of all types generate enormous amounts of structured, semi-structured, and 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 using commercial and open source tools is an increasingly important research approach for a growing interdisciplinary group of scholars, practitioners, government officials, and international organizations.

2. Minitrack Topics and Themes

The minitrack on Text Analytics is designed to provide an interactive forum for interdisciplinary researchers to discuss the critical issues of text mining and to contribute to the ongoing focus on big data at HICSS. Our minitrack invites papers that apply theoretical and applied text-mining approaches to a wide variety of substantive domains, including, but not limited to:

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

And addressing methodological challenges 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 growth of our HICSS community. We received eleven submissions and after our peer review, accepted five excellent papers. These papers highlight various important aspects of this emerging research community, including one Best Paper nomination. We are excited about our virtual minitrack and look forward to discussions of these papers.

URI: https://hdl.handle.net/10125/79424
978-0-9981331-5-7

In the following sections, we present a summary of the HICSS-55 Text Analytics minitrack and our five papers to be presented during two virtual panels. Our first panel of three papers focuses on machine learning in text analytics. The second panel focuses on model development in language detection and language learning. In the remaining time, we will discuss ways to strengthen our growing and vibrant HICSS community.

3. Paper 1: New Threats to Privacy-Preserving Text Representations

Our first paper is grounded in the privacy concerns from users of information systems and the mandates to data publishers to protect privacy by anonymizing the data before sharing it with data consumers. This study sees the goal of privacy-preserving representation as protecting user privacy while ensuring the utility and accuracy of the published data. This study highlights the role played by privacy-preserving embeddings, which are usually functions that are encoded to low-dimensional vectors to protect privacy while preserving important semantic information about an input text. In this study, the researchers demonstrate that these embeddings still leak private information, even though the low dimensional embeddings encode generic semantics. The researchers develop two classes of attacks, i.e., adversarial classification attack and adversarial generation attack, to study the threats for these embeddings. In particular, the threats are (1) these embeddings may reveal sensitive attributes letting alone if they explicitly exist in the input text, and (2) the embedding vectors can be partially recovered via generation models. The experimental results of the study show that this approach can produce higher-performing adversary models than other adversary baselines.

4. Paper 2: On the use of Machine Learning and Deep Learning for Text Similarity and Categorization and its Application to Troubleshooting Automation

Our second paper sees troubleshooting as a labor-intensive task that includes repetitive solutions to similar problems. The researchers argue that this task can be partially or fully automated using text-similarity matching to find previous solutions, lowering the workload of technicians. They develop a systematic literature review to identify the best approaches to solve the problem of troubleshooting automation and classify incidents effectively. The researchers also identify promising approaches and point in the direction of a comprehensive set of solutions that could be employed in solving the troubleshooting automation problem.

5. Paper 3: What to Prioritize? Natural Language Processing for the Development of a Modern Bug Tracking Solution in hardware Development

Our third paper addresses the problem of managing large numbers of incoming bug reports and finding the most critical issues in hardware development is time consuming, but crucial in order to reduce development costs. In this paper, the researchers present an approach to predict the time to fix, the risk and the complexity of debugging and resolution of a bug report using different supervised machine learning algorithms namely Random Forest, Naive Bayes, SVM, MLP and XGBoost. Further, the researchers investigate the effect of the application of active learning and evaluate the impact of different text representation techniques, namely TF-IDF, Word2Vec, Universal Sentence Encoder and XLNet on the model's performance. The evaluation shows that a combination of text embeddings generated through the Universal Sentence Encoder and MLP as classifier outperforms all other methods and is well suited to predict the risk and complexity of bug tickets.

6. Paper 4: Generating Vocabulary Sets for Implicit Language Learning Using Masked Language Modeling

The fourth paper argues that a well-balanced language curriculum must include both explicit vocabulary learning and implicit vocabulary learning. However, most language learning applications focus on explicit instruction. Students require support with implicit vocabulary learning because they need enough context to guess and acquire new words. Traditional techniques aim to teach students enough vocabulary to comprehend the text, thus enabling them to acquire

new words. Despite the wide variety of support for vocabulary learning offered by learning applications today, few offer guidance on how to select an optimal vocabulary study set. This paper proposes a novel method of student modeling with masked language modeling to detect words that are required for comprehension of a text. It explores the efficacy of using deep learning via a pre-trained masked language model to model human reading comprehension and presents a vocabulary study set generation pipeline (VSGP). Promising results show that masked language modeling can be used to model human comprehension and the pipeline produces reasonably sized vocabulary study sets that can be integrated into language learning systems.

7. Paper 5: Towards Automated Moderation: enabling Toxic Language Detection with Transfer Learning and Attention-Based Models

Finally, our fifth paper, which is also our Best Paper nominee, argues that our world is more connected than ever before. Sadly, however, this highly connected world has made it easier to bully, insult, and propagate hate speech on the cyberspace. Even though researchers and companies alike have started investigating this real-world problem, the question remains as to why users are increasingly being exposed to hate and discrimination online. In fact, the noticeable and persistent increase in harmful language on social media platforms indicates that the situation is, actually, only getting worse. Hence, in this work, we show that contemporary ML methods can help tackle this challenge in an accurate and cost-effective manner. Our experiments demonstrate that a universal approach combining transfer learning methods and state-of-the-art Transformer architectures can trigger the efficient development of toxic language detection models. Consequently, with this universal approach, we provide platform providers with a simplistic approach capable of enabling the automated moderation of user-generated content, and as a result, hope to contribute to making the web a safer place.

8. Towards a Text Mining Community

We believe the Text Mining minitrack has makes an important contribution to HICSS. It 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 55th HICSS represent what we see as an important emergent trend, which we believe will remain for many years to come.