Summer 7-24-2020

A Mixed-Method Approach to Investigating Difficulty in Data Science Education

Sydney Shearer  
*Juniata College*, shearse18@juniata.edu

Ellie Strauss  
*Bates College*, estrauss@bates.edu

Ethan Hawk  
*Valparaiso University*, ethan.hawk@valpo.edu

Sasha Lioutikova  
*Yale University*, sasha.lioutikova@yale.edu

Marius Orehovschi  
*Colby College*, moreho21@colby.edu

*See next page for additional authors*

Follow this and additional works at: [https://scholar.valpo.edu/sires](https://scholar.valpo.edu/sires)

**Recommended Citation**

Shearer, Sydney; Strauss, Ellie; Hawk, Ethan; Lioutikova, Sasha; Orehovschi, Marius; Vazquez, Frankie; Clark, Linda; Kinnaird, Katherine; Sandstede, Bjorn; Schmitt, Karl; and Wertz, Ruth, "A Mixed-Method Approach to Investigating Difficulty in Data Science Education" (2020). *Summer Interdisciplinary Research Symposium*. 69.  
[https://scholar.valpo.edu/sires/69](https://scholar.valpo.edu/sires/69)

This Poster Presentation is brought to you for free and open access by the Office of Sponsored and Undergraduate Research at ValpoScholar. It has been accepted for inclusion in Summer Interdisciplinary Research Symposium by an authorized administrator of ValpoScholar. For more information, please contact a ValpoScholar staff member at scholar@valpo.edu.
Authors
Sydney Shearer, Ellie Strauss, Ethan Hawk, Sasha Lioutikova, Marius Orehovschi, Frankie Vazquez, Linda Clark, Katherine Kinnaird, Bjorn Sandstede, Karl Schmitt, and Ruth Wertz

This poster presentation is available at ValpoScholar: https://scholar.valpo.edu/sires/69
The purpose of this study was to define a methodology to identify disconnect between students and instructors in data science classrooms through analyzing qualitative data. A combined qualitative and quantitative approach was used for analysis of survey data from three institutions. As a whole, the methods used throughout this research process provide direction for researchers in interpretation and analysis of the survey data in an efficient and time-sensitive manner. Although the research was applied to data science classrooms, this method has the potential to be applied into other fields and areas of study when performed with coordination between a field expert and a data scientist.

Dataset

The dataset was collected during the COVID-19 pandemic, resulting in a smaller dataset. The survey was conducted in data science classrooms about aspects of class.

Who was surveyed?
- Students (304 responses)
- Faculty/instructors (112)
- Teaching assistant (63)

Where was this conducted?
- Brown University
- Smith College
- Valparaiso University

What was asked?
- Six Likert Scale Questions
- Four Open-Ended Questions

Methods

- Natural Language Processing
  - Removed stop words
  - Lemmatization
  - Tokenization
  - Bags of bigrams
- Quality analysis
  - Proportion of relevant words in responses
- Manual Content Analysis
  - Categories created from inductive reasoning, responses categorized
- TF-IDF analysis
  - Frequency of bigrams
  - Manually labeled into categories
  - Corpora
  - 7 separate corpora used
  - 4 from open-ended questions
  - 3 from respondent roles
  - Responses used as documents
  - Bigrams used as terms

TF-IDF Analysis of Bigrams

TF-IDF: “Term Frequency, Inverse-Document Frequency”
- Term Frequency (TF): times a term appears in a document
- Inverse-Document Frequency (IDF): logarithm of the inverse of the number of documents in which the term appears
- To calculate, multiply TFIDF by IDF to obtain a weight for each term for each document
- Finds which terms are important within the documents by filtering out words that do not give insight

Conclusions/Limitations

- Manual content analysis with an automated algorithmic process gives in-depth understanding to datasets
- Data scientists should collaborate with field experts to gain insight of dataset
- Without collaboration, an in-depth analysis could not be achieved
- Data was collected during COVID-19 pandemic, resulting in a smaller dataset

Acknowledgements

- Valparaiso University
- EPIC
- TRIPODS+X-EDU NSF Grant
- DMS #1839257, 1839259, 1839270
- Ruth Wertz, Valparaiso University
- Linda Clark, Brown University
- Katherine M. Kinnaird, Smith College
- Karl R. B. Schmitt, Valparaiso University
- Bjorn Sandstede, Brown University