A Roadmap for Incorporating Online Social Media in Educational Research

HAMID KARIMI
Michigan State University

TYLER DERR
Michigan State University

KAITLIN T. TORPHY
Michigan State University

KENNETH A. FRANK
Michigan State University

JILIANG TANG
Michigan State University

Thanks to advancements in communication and online social media, there has been a surge of useful online educational resources across the Internet. In addition to supplementing educational materials, these resources could be used in varying education research and potentially advance the quality of education. Nevertheless, conducting such research projects requires using big data techniques and approaches to find meaningful resources and harnessing them in an effective way. In this chapter, we present a roadmap for how to incorporate online social media in education research projects. The roadmap consists of three major components: project initialization, data collection, and data utilization. Furthermore, we present some learned lessons, tips, and tricks, as well as case studies from the Teachers in Social Media project (www.teachersinsocialmedia.com/). We believe this chapter can be used as a practical reference point for many researchers whose concern is connecting data to their education research endeavors.

Social media has become an integral part of human life in the 21st century. The number of social media users in 2017 was estimated to be around 2.5 billion individuals (eMarketer, 2019). Social media platforms (e.g., Facebook) have facilitated interpersonal communication, diffusion

Teachers College Record Volume 121, 140312, 2019, 24 pages

0161-4681

of information, the creation of groups and communities, to name a few. As far as education systems are concerned, online social media has transformed and connected traditional social networks within the schoolhouse to a broader and expanded world outside (Wellman, 2001). More specifically, thanks to advancements in communication, educators have access to ample online instructional resources curated and shared across social media platforms. In such expanded virtual space, teachers engage in various activities within their community (e.g., exchange of instructional resources, seek new methods of teaching, engage in online discussions, and so on; Pempek, Yermolayeva, & Calvert, 2009; Torphy & Drake, 2019, this yearbook). Students use social media as well—for example, to supplement educational materials and interact with others (Bagdy, Dennen, Rutledge, Rowlett, & Burnick, 2018; Greenhow, Robelia, & Hughes, 2009; Rutledge, Dennen, & Bagdy, 2019, this yearbook). Educational policy makers take advantage of social media to infer public opinion about new policies (Daly, Liou, Del Fresno, Rehm, & Bjorklund, 2019, this yearbook). Parents seek out resources within social media to supplement their children with educational materials (Calarco, 2011). Hence, education today is closely intertwined with online social media.

It is of great importance for researchers in the education field to understand social media and acquaint themselves with online resources. This chapter is an attempt to present a roadmap for conducting education research using online social media resources. Our activities in the Teachers in Social Media project form the basis of this roadmap. Started in 2015, this project considers the intersection of the cloud to class, the nature of resources within virtual resource pools, and the implications for equity as educational spaces grow. Much of the work coming out of the Teachers in Social Media project concerns instructional and educational resources shared on Pinterest (www.pinterest.com), an image-based public platform that connects more than 250 million users across the globe (TechCrunch, 2018). Covering the technical details of the new approaches and algorithms we have proposed in the Teachers in Social Media project is beyond the scope of this chapter.[1] Instead, we seek to impart our experience and learned lessons through presentation of a roadmap in order to guide and facilitate new academic research projects in this area.

Connecting data from online social media for education research is a part of the broader topic of big data in education research. A considerable number of studies exist that concern utilizing big data to improve the quality of education (Agasisti & Bowers, 2017; Bowers, Bang, Pan, & Graves, 2019; Ho, 2017; Wang, 2016, 2017a; Zeide, 2017). Big data has three distinct aspects: large volume, wide variety, and high velocity (boyd & Crawford, 2012). More recently, veracity has been identified as

another aspect of big data as well (Bello-Orgaz, Jung, & Camacho, 2015; Lukoianova & Rubin, 2014). In this regard, topics such as fake news detection and deception detection have become active and well-recognized research directions (Karimi, Roy, Saba-Sadiya, & Tang, 2018; Karimi & Tang, 2019; Karimi, Tang, & Li, 2018). Next, we briefly discuss these aspects of big data and specially their relations to online social media data.

The first aspect, large volume, defines big data as a tremendous collection of data. Even though there is no clear threshold definition of how big the "big data" are, this aspect contrasts new digital data against traditional "small volume" data sets. In education, we have access to a large volume of data produced by teachers, students, administrators, policy makers, and so on—for example, 200 million math test scores are readily available (Ho, 2017; Reardon, Kalogrides, & Shores, 2019). As far as the focus of this chapter—online social media data—is concerned, big data are generated and curated by millions of online users (Anderson & Rainie, 2012). For instance, in the Teachers in Social Media project, we identified hundreds of thousands of educational pins shared by teacher Pinterest users. Therefore, online social media for education research has a large volume.

The second aspect of big data is wide variety. The data can be different types, such as text, image, video, audio, web pages, maps, and so on. The variety of online social media data stem from the advancements in technology that have enabled online users to produce various data types (Gandomi & Haider, 2015). Usually, different data types of an educational resource offer a complementary view regarding that resource. For instance, we often found that a mathematical resource shared on Pinterest (i.e., a *pin*) is reflected in a digital image accompanied by some textual description, which together characterize that mathematical resource.

The third important aspect of big data is high velocity. The great advancements in telecommunication and the Internet have enabled almost instantaneous transmission of data (Wellman & Haythornthwaite, 2008). Correspondingly, educational resources can now spread very rapidly across social media platforms. For instance, a teacher can prepare a lecture and broadcast it through Facebook livestream to his or her audience, allowing another teacher to use the resource and be informed by the experience within hours.

The last aspect of big data is veracity or trustworthiness of data. It has been well recognized that data can be biased, inaccurate, and implausible (Lukoianova & Rubin, 2014). We need to emphasize the importance of ensuring trustworthy and unbiased data for education research; otherwise, further inferences will be flawed and lead to ill-advised actions. Interested readers are referred to previous studies on veracity and correctness of big data (Bello-Orgaz et al., 2015; Lukoianova & Rubin, 2014; Salganik, 2017).

The present chapter is a commentary on a general guideline to help new researchers use vast online social media data, as a manifestation of "big data," in their education research. This chapter is a complementary reference to the previous literature on using social media in education research (Halverson, 2014, 2018; Hora, Bouwma-Gearhart, & Park, 2017; Wang, 2013, 2016, 2017b). In this chapter, unlike in previous studies, we present a clear roadmap enriched by challenges, solutions, practices, and even administrative experiences from the Teachers in Social Media project. In particular, we present approaches to incorporate online social media in education research from a data science perspective in which we use principled and systematic ways to process and infer meaningful information from heterogeneous and unstructured data (e.g., the collection of instructional pins from Pinterest).

AN OVERVIEW OF THE ROADMAP

Figure 1 illustrates an overview of the major components of the roadmap for incorporating online social media in education research. In the following paragraphs, we briefly review these components.

First, the Teachers in Social Media project team members made several crucial decisions before building the project: (1) we built an effective team while considering the interdisciplinary nature of education research using online social media data, (2) we clearly determined the project's requirements and expectations, and (3) we considered funds and resources, allocating them appropriately.

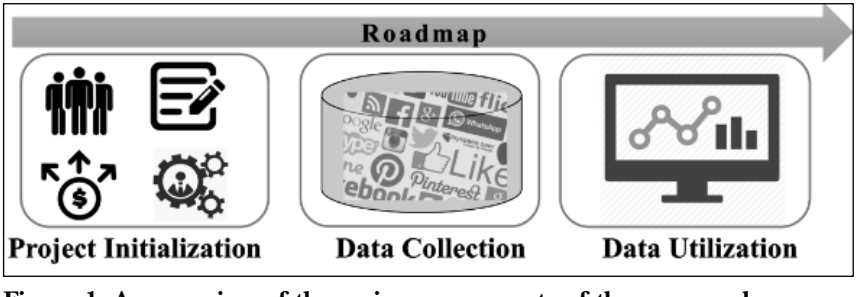

Figure 1. An overview of the major components of the proposed roadmap for incorporating online social media in education research

Second, given the unique nature of the data, education researchers must learn a new method for data collection and building a data archive. Combining disciplinary expertise, the Teachers in Social Media team was able to avoid inefficiencies in time and resources. Later, we present two common approaches for collecting data from online social media sources.

The last component of the roadmap involves presenting techniques and tools that can help us to use the collected data in a better way. More precisely, we present an overview of some machine learning and data mining techniques that help to extract meaningful and complex patterns from collected data that may be difficult to obtain otherwise.

PROJECT INITIALIZATION

Below we highlight insights learned (sometimes the hard way) through interdisciplinary work incorporating social media into our education research project. Specifically, we describe building an interdisciplinary team, defining the roles for each team member recruited, and the importance of wisely allocating funds and resources—all at an early stage to minimize future headaches in managing the project and maximizing efforts to ensure a larger impact on the research community.

BUILDING YOUR TEAM

According to Rossini and Porter (1979), there are four main sociocognitive frameworks, which can range from having an intensive group in which everyone interacts together to generate common group knowledge, to a small group of likeminded individuals. One approach they outlined includes a few individuals (typically of a single discipline) defining a project and consequently bringing in other researchers, often from an outside area of expertise. In another approach, a single team member (for example, the project leader) individually interacts with others, but the team members never get together for discussion or explicitly work together to achieve the common goal. Although it might seem obvious that we should avoid the latter, we consciously made efforts to build a strongly connected team that, through discussions, can collaborate effectively.

When engaging in social science research incorporating big data from virtual space, we suggest connecting with a data scientist early in the project's development. By data scientist, we mean someone who is expert in data mining and machine learning methods and who has a deep understanding of social science (e.g., education) research (Dhar, 2012; Romero & Ventura, 2013; Van der Aalst, 2016). This allows researchers to leverage varied expertise and potentially combine descriptive analyses with causal analyses, and quantitative and qualitative approaches. Often, the project leader is an expert in one of these two research fields, and the formation of interdisciplinary groups and subprojects stimulates learning behavior across the team (Gibson & Vermeulen, 2003).

With project expertise across disciplines, recruiting a team of researchers and students is imperative to provide the analytical capacity necessary

to deal with varied data that are updated iteratively. Graduate (both PhD and master's) students in both fields should work together, across disciplines, to build unique approaches to address analytical challenges. The other key members who should be recruited are undergraduate students. Qualitatively, they can help with surveys, interviews, and, in our case, coding instructional resource content. Quantitatively, undergraduate computer programmers (e.g., junior or senior computer science majors) can provide innovative approaches to data scraping and collection.

REQUIREMENT SPECIFICATION

Using social media data requires flexibility in both the approaches to a particular research question and the kinds of research questions that might be asked. For example, it would not be appropriate to generalize research findings to all teachers, given that we observe behaviors among those teachers engaged within social media. To guide the project in a more fruitful direction and achieve better results, it is essential to perform a through brainstorming of the research questions collaboratively (Bouchard, 1971; Parnes & Meadow, 1959; Rawlinson, 2017).

Incorporating social media data into an interdisciplinary team within education requires researchers to learn one another's field-specific language. Research tasks should be explicitly stated, such as data collection, data preprocessing, and data cleaning, followed by harnessing data mining and machine learning techniques. Through these efforts, one may begin to extract insightful patterns or make predictions with social media data. Last, goal setting and putting together a workflow of ongoing work provide team members with a way to see progress and provide faculty with a way to balance long-term planning, particularly for those interested in software development (Paetsch, Eberlein, & Maurer, 2003).

An important decision to make involves where to find the proper and relevant data. In other words, we need to carefully and wisely identify the social media source(s) where rich instructional resources reside for our research purpose. A thorough investigation of current social media platforms and discussing their pros and cons for education research are beyond the scope of this chapter. However, to make our roadmap more effective, we draw the reader's attention to some very interesting insights and comments about social media in education, in particular the role of different social media platforms in today's education (Frank & Torphy, 2019, this yearbook). Note that there are commercial sources online that allow one to purchase social media data; however, we should avoid such avenues to procure data because of (1) legal and privacy issues, (2) questions about the reliability of the source, (3) extra cost (given that purchasing

social media data costs significantly more than recruiting programmers to your team and obtaining the needed computational resources), and (4) lack of freedom and flexibility in the data that one can obtain.

FUNDING AND RESOURCE ALLOCATION

One of the most important parts of starting a project is determining how to allocate funding and resources. For those who are early in their career, it can be difficult to find the right balance between spending efficiently and overinvesting on a research agenda for the project. However, even for senior researchers, if they have not been exposed to, or explored using, online social media for their research, these uncharted waters can raise many questions. Thus, next we provide suggestions, based on our experiences, to raise awareness of the likely costs involved when using online social media in research.

Here we will discuss the computational related investments that are likely necessary when incorporating online social media and touch on some software packages that can be used for improving communication and team collaboration. First, to perform the data collection and utilization, you will need hardware (i.e., your computational power). It is recommended that you obtain a server (instead of a desktop) for the following reasons: (1) if your project requires collecting a lot of data, you will need a lot of storage space to hold your data, and a server can provide more efficient (in terms of time to process your data) and more reliable (in terms of backing up your data) storage with multiple hard drives; (2) many machine learning techniques have significant improvements in time to completion when a graphics processing unit (GPU) is used (Owens et al., 2008), which you can have on your server; and (3) it can be maintained and shared by the entire programming/quantitative subgroup(s) remotely, as compared with a single desktop. Although our suggestion of obtaining a server might be questioned, in our personal case, we can attest that it led to significant improvement in the workflow of our project. Further, as previously mentioned, having a backup of the data on your own personal hard drives can provide peace of mind and allow you to avoid the catastrophic loss that would occur if a single drive were to fail in your RAID system (Chen, Lee, Gibson, Katz, & Patterson, 1994). We should emphasize that the takeaway message here is the need to store and back up the collected data efficiently and reliably, while keeping in mind that in the future, new storage technologies (e.g., cloud-based server) might emerge and become popular.

As for communication, investing some resources in using communication tools like Slack (www.slack.com) or other future communication

tools is recommended for the ability to quickly chat with your team. Such applications also allow for creating groups, assigning tasks, and sharing files, among other features. In terms of other software packages, using file-sharing services such as Dropbox (www.dropbox.com) can be significantly more effective (as compared with repeatedly sending links or emails), especially when transferring and sharing larger amounts of data or conducting analysis that might change over time. However, public versus private data-sharing spaces should be considered depending on institutional review board parameters or other privacy and security regulations.

LESSONS LEARNED

- The most important takeaway message for building your team is building connections between social and data scientists; this will allow you to leverage varied levels of expertise across disciplines. Furthermore, this will help you construct an interdisciplinary hierarchy for working on multiple subprojects simultaneously.
- We recommend that the project director(s) define a set of clear high-level goals while attempting to revise and possibly modify the project requirements continuously. The need for revision stems from the dynamic and heterogeneous nature of the projects involving online social media data for education research purposes.
- Although it can be difficult to determine the right amount to spend on computational resources, we highlight again the importance of obtaining a server. This will be necessary for education research endeavors if one is seeking to use social media data, and one should ensure that proper steps are taken to back up the data—for example, using a RAID server (Chen et al., 1994).

DATA COLLECTION

To conduct effective and scalable education research within social media, we need to know how to collect data from an online social media source. Generally, two methods are used to collect data from social media platforms: data scraping, and using an application programming interface (API). In the following sections, we explain these two methods.

DATA SCRAPING

In data scraping, a computer program extracts data from a human-readable source, for example, a web page (Mitchell, 2018). In this method, what is being "scraped" is meant for humans and therefore lacks documentation

and a well-defined structure digestible by computer programs. As far as extracting data from online social media is concerned, a particular form of data scraping known as web scraping is typically employed. A web scraper program takes as input a web page (e.g., someone's blog) and attempts to extract data from the markup language used to display the web page, that is, hypertext markup language (HTML).

Recently, some packages and tools have been developed for web scraping. To illustrate how one can perform web scraping, we present a simple case study to retrieve an educational resource from Pinterest.com illustrated in Case Study 1 in this chapter. As shown in this case study, one needs to identify where a resource is located on the web page (first visually, and then in the HTML script). Sometimes, this resembles finding a needle in a haystack. In fact, web scraping is associated with several issues that hinder its practical use:

- The dynamic nature of modern web pages makes it extremely difficult for web scraping programs to find a resource.
- Even in the case of static web pages, a web page's structure might change frequently, thus requiring rewriting and redesigning the entire program.
- Some online social media platforms (and other websites) block web scrapers' access to their content.
- The ownership and copyright of the content might be infringed.
- Web scrapers incur extract programming, processing, and cost overhead.

Because of the aforementioned challenges, web scraping should be your last resort. Interested readers are referred to Keegan, 2019, for more details about web scraping. Next, we present a principled and scalable way to collect data from many existing online social media platforms.

API

API is an interface to send requests to a server (e.g., a social media data storage computer) and obtain the responses in a structured format. The responses, unlike HTML tags, which have been developed to display the content to users, can be accessed and processed easily by a computer program. The current social media platforms offer a rich set of APIs that facilitate data access and collection. In Case Study 2, we demonstrate how one can obtain the same resource in Figure 2 easily.

Case Study 1

Figure 2 illustrates a simple mathematical pin (image) showing a comparison between 2D shapes. This has been shared via the first author's Pinterest account accessible at https://www.pinterest.com/pin/713539134690793610/. The following python code snippet shows a web scraper program attempting to collect the image in Figure 2.

```
import requests
from bs4 import BeautifulSoup
raw_content = requests.get ("https://www.pinterest.com/pin/713539134690793610/").text
html=BeautifulSoup (raw_content, 'html.parser')
resources=html.find_all('SPECIFY THE TAG')
...
```

Using the library *requests*, we can obtain the content of the entire web page. Then the library *BeautifulSoup*, a special library developed for web scraping, parses the content into an HTML-coded structure. Then, we should find an HTML markup tag associated with the desired resource, and so on. Using web scraping, finding the corresponding web address that hosts the image in Figure 2 becomes very complicated.

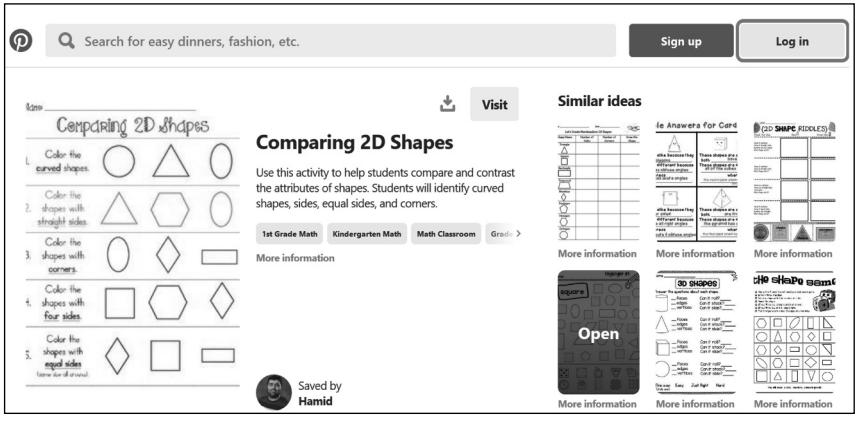

Figure 2. A sample educational resource shared on Pinterest.com

Case Study 2

The following program shows how one can acquire a pin from Pinterest using Pinterest's API. Unlike scraping, the response we obtain is in a fixed and processable format from our simple python program. The program easily saves the image on disk (under the name *resource.jpg*) by downloading the obtained uniform resource locator (URL). Note that to access an online social media's API, one needs to register with the social media platform and receive an authorized access token (this allows you to connect with its server); in the case of Pinterest, this token can be obtained from the following link:

https://www.pinterest.com/oauth/?response_type=token&consumer_id=1431596.

```
import requests
import urllib.request
response =
requests.get('https://api.pinterest.com/v3/pins/713539134690793610/?access_toke=TOKEN')
URL = response.json()['data']['image_large_url']
urllib.request.urlretrieve(URL, "resource.jpg")
```

Figure 2. A sample educational resource shared on Pinterest.com (continued)

LESSONS LEARNED

- Regardless of the method used for data collection, the most important aspects of data collection and storage are data privacy and sharing issues. Although the concept of privacy itself is ambiguous (Solove, 2008), we recommend that researchers take robust measures to protect the data and adhere to online social media regulations and any further project privacy policies (e.g., National Science Foundation privacy policies). We refer the reader to Daniel (2017) and Ho (2017) for more discussion about privacy issues in education research, and to Austin (2018) and Austin et al. (2017) for information on how to curate and publish a data set effectively.

- One of the problems regarding data scraping that we encountered in the Teachers in Social Media project was the complexity of coding a functional and reliable scraping. This is due to the structure of the platform and the purposes for which it was developed: curation of resources.
- We want to draw project developers' attention to two important messages if they opt to use API to collect data from a social media platform. First, it is imperative that you read the terms of service of the API you will be using. You should be mindful of not violating any user privacy policy and technical restrictions enforced by the social media platform (e.g., the number of accounts whereby the API is authenticated). Second, sometimes API calls and returned data structures are subject to change. We experienced this when we collected data from Pinterest. For instance, we discovered that some data fields associated with pins that we collected in the early stages of the project had been removed. You need to keep this in mind while developing a data collection program.

DATA UTILIZATION: TECHNIQUES AND TOOLS

So far, we have discussed how to collect data from online social media and what data can be collected. In this section, we explain how you can utilize the data in an education research project. In this section, we first describe data preprocessing as a crucial step to preparing the data, followed by a discussion on machine learning and, ultimately, how to perform data analysis.

DATA PREPROCESSING

Real-world data sets are often noisy and incomplete, and contain inconsistencies. These challenges make utilizing and mining a data set quite difficult and error prone. Therefore, a crucial step toward a better understanding and analysis of data involves a set of data preprocessing tasks—that is, a set of data mining techniques to transform the data into an analyzable format. (Thoroughly investigating such techniques is beyond the scope of this chapter.) Hence, in this part, we briefly explain the major tasks and provide the reader with additional resources that can be consulted.

Data cleaning. In this task, missing values are filled (Donders, Van Der Heijden, Stijnen, & Moons, 2006; Royston, 2004, 2005; Van Buuren, 2018), noisy values are smoothed (Han, Pei, & Kamber, 2011), outliers are detected and removed (Aggarwal & Yu, 2001; Chandola, Banerjee, & Kumar, 2007; Hodge & Austin, 2004), and inconsistencies are resolved

(Kumar & Chadrasekaran, 2011). Many data cleaning tools are available, such as OpenRefine (http://openrefine.org/), DataWrangler (http://vis.stanford.edu/wrangler/), and NLTK (https://www.nltk.org/). Importantly, in our interdisciplinary team, data cleaning included having conversations about the focus of the research and how to preserve the main data set in its most authentic format without imputation or deletion. Then, for particular analyses, researchers could make decisions that worked best for that research question and context. Note that when incorporating online social media data, the data cleaning is very important because untrained online users generate the content of the data.

Data integration. When multiple data sources or files are collected, we might need to integrate them into a unified data set (Lenzerini, 2002). This task also plays an important role because data might be collected from different social media platforms, a combination of social media data and other sources (e.g., surveys), different communities in a social media platform (e.g., educational resources shared and curated by students and teachers), and so on.

Data normalization. When normalizing your data, they are scaled, and it is ensured that common attributes are represented in similar ranges. This is a very important step for many data analysis algorithms because they are sensitive to large variations in the data. In other words, data mining and machine learning methods are very susceptible to focusing more on the types of data that have a larger magnitude; thus, we need to normalize the data to avoid these numeric issues inherent in the methodologies (Grus, 2019). For examples of normalization and other data science techniques, we direct the readers to the code repository (https://github.com/joelgrus/data-science-from-scratch) from Grus (2019). An example of when normalization would be needed in social media data concerns the number of followers and the number of daily post values for a given user on, say, Twitter or Facebook. It is likely that these values are at least an order of magnitude (or more) different from each other in most cases (i.e., typically people have many more followers than the number of times they post in a given day, for example, 500 and 5, respectively). Data normalization methods include min-max normalization, studentized residual, and normalization by decimal scaling. The python package scikit-learn (https://scikit-learn.org/stable/) offers a rich toolbox for data normalization methods (and has a plethora of other data mining and machine learning methods that are easy to use). Similarly, in our analysis of Pinterest data, we find a large variation in the number of resources pinned by users. The sampling framework or normalization employed on data relates to the specific analysis rather than structuring the main data set.

Data reduction. Big digital data sets can be quite large and likely are of high dimension (i.e., they can have many features and qualities we can extract out, such as age, gender, and vocabulary used on Facebook). In data reduction, we reduce the representations of a data set while ensuring that not too much useful information is lost. Data reduction methods include data cube aggregation (Gray et al., 1997), data aggregation (Ramírez-Gallego et al., 2016), and dimensionality reduction (Sorzano, Vargas, & Montano, 2014), among others. Notably, principal component analysis (Jolliffe, 2011) is a common data reduction method that is available in many packages, including scikit-learn. One thing to note is that most of these methods lose the interpretable meaning of the resulting features, and so other feature selection techniques can be used that simply select a subset of the existing collected set, such as backward feature elimination (Abe, 2010), forward feature construction (Abe, 2010), and so on.

MACHINE LEARNING

There is a rich literature on data analytics in education research (e.g., Agasisti & Bowers, 2017; Baepler & Murdoch, 2010; Baker & Inventado, 2014), a thorough review of which is beyond the scope of this chapter. Instead, we describe a straightforward process and present a discussion about machine learning for education research in the following paragraphs.

Once we preprocess the data, we need to utilize them to extract patterns. These patterns, however, are often quite difficult to identify. For instance, consider a simple instructional resource demonstrated in Figure 3. The task in this mathematical problem is to fill in the blanks with the correct answers in the summations. Now, suppose our goal is to have a program that can perform some useful tasks given similar mathematical resources (images)—for example, checking the correct answers against a student's, and assigning a difficulty score to each image. Then, the very first requirement for such a program involves recognizing the digits in an image. The human brain can easily detect digits in an image even at the kindergarten level (or earlier). However, is extremely challenging to have a fixed "formula" codify what a digit in an image is. We know, however, that such a formula exists; for instance, clearly digit 1 is different from 8. The question now is: How do we describe those differences in a way that a computer program can use to consistently determine digits in an image? This is where machine learning can be helpful.

Machine learning techniques are in broad use today. They recommend books on Amazon, help in sorting emails, find information on Google, and

allow Siri to answer a user's questions. E-mail systems use machine learning tools to remove spam, identify fraudulent emails, and perhaps even suggest responses to emails that you receive (see Jordan &Mitchell, 2015, for more examples and discussions). Fundamentally, machine learning systems are algorithms that can identify the relationships between items of information. These relationships (correlations or patterns) are not explicitly formulated for them; thus, the algorithm engages in inductive reasoning on its own to identify such patterns, thereby performing a useful task (e.g., recognition of digits in an image similar to that in Figure 3).

In machine learning, the machine (i.e., the computer program), as expected, to learns to mimic human coding—for example, the cognitive demand of an educational resource. To make this feasible, we need to provide the machine with a considerable number of samples knowns as the training set. For instance, in the case of digit recognition discussed earlier, each sample in the training set can be a 20×20 pixel image, with each pixel having a different grayscale score. With advancement in communication and the Internet, we can collect samples to construct a large enough training set, and, in some cases, when the data is scarce, we can even synthesize samples (Goodfellow et al., 2014; Sutskever, Martens, & Hinton, 2011). Further, we must generate another database of data points and answers to use as the test set. We then feed the training set into our machine learning algorithm—called a "learner"—and let it go to work. Once the learner is trained, we can use it for its assigned task—for example, solving a "making 10" problem in the test set as illustrated in Figure 3.

We employ various data machine data mining and learning approaches in the Teachers in Social Media project. For instance, using the textual data of collected instructional pins from Pinterest, we identify the topics of pins. By doing so, we can identify what topics students or teachers are more interested in. To make this feasible, we use latent Dirichlet allocation (Blei, Ng, & Jordan, 2003), which is an unsupervised topic modeling approach available in the gensim Python package (https://pypi.org/project/gensim/). We also develop deep neural network models to perform complicated tasks (e.g., classification of mathematical pins according to their cognitive demand). Two popular deep neural network packages include Pytorch (https://pytorch.org/) and TensorFlow (www.tensorflow.org/).

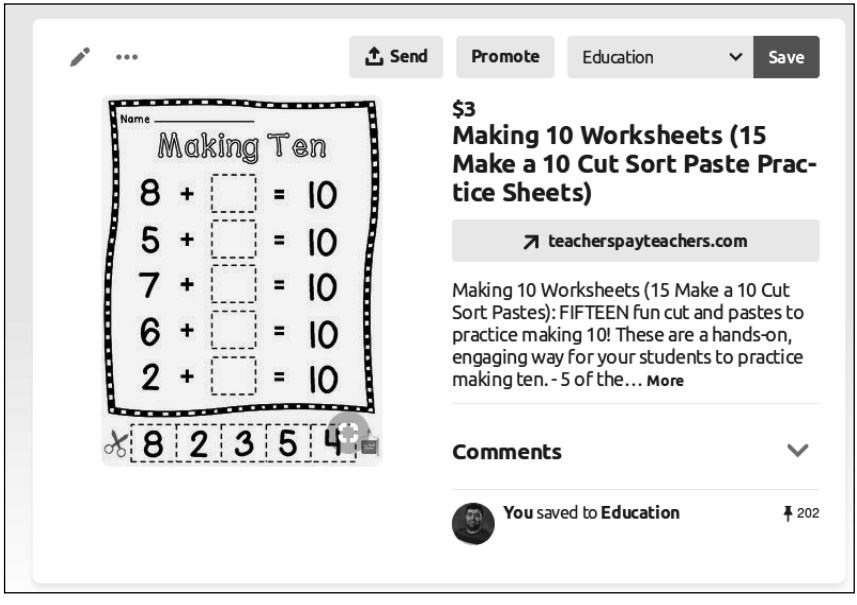

Figure 3. A simple kindergarten-level instructional resource

ANALYSIS

Analyzing the results in a research project is of great importance. This is even more important in our roadmap, given that we are dealing with heterogeneous, large, noisy, incomplete, and sometimes inaccurate data environments. Usually we start analyzing our data with some prior knowledge and assumptions. The data utilization component of our roadmap includes a set of approaches and techniques to explore the data that are aimed at finding meaningful and informative patterns and correlations. Effective analysis of discovered patterns and correlation demands a dedicated effort. For data analysis, one can utilize visualization tools (Ward, Grinstein, & Keim, 2015).

Regardless of the approach and technique of data analysis, we might revise our strategies and decisions in different components and stages of the roadmap. For instance, after analysis, we might discover that the utilized machine learning is incapable of proving informative results; therefore, we need to seek an alternative machine learning algorithm. Another common example is that we detect anomalies and noisy data samples when we carefully analyze the results, so we need to go back to the data preprocessing step and remove or correct the troubling data using the approaches discussed previously. Project directors and decision makers should take

great care during the early stages of the project (e.g., for data source selection and collection) because further changes later on might incur extra cost and require additional energy.

We also note that our work relates to the existing literature of both educational data mining (Baker, 2010; Romero & Ventura, 2013) and learning analytics and knowledge (LAK; Siemens & Baker, 2012). Essentially, educational data mining is what we described here in using data mining techniques specifically for educational data, with a focus on using educational software for student modeling and automation by removing the human in the loop (Siemens & Baker, 2012). A thorough background on educational data mining (although now not current because of the vast changes and widespread adoption of social media) can be found in Baker and Yacef's article (2009). More recently, Baker and Koedinger (2018) discussed an attempt to automatically assess students' understanding to provide them with the content they should be learning based on their performance. In comparison, LAK has roots in both computer science (e.g., semantic web) and sociology/psychology of learning, where the focus is more on how to inform instructors and learners (Siemens & Baker, 2012). However, in comparison, we are specifically focused on using data mining and machine learning techniques for social media–related data in education research. For instance, data backup and team formation strategies are in line with incorporating "big data" for the project. Moreover, the examples and case studies presented in this chapter are all coming from our experiences in response to the research questions in the Teachers in Social Media project. In fact, we utilized this ongoing education research project as a way to convey a principled approach to incorporating online social media in an education research project.

LESSONS LEARNED

- As mentioned, real-world data sets are often noisy and incomplete, so some preprocessing is required. One of the most important lessons learned here is to always double-check that this stage has been performed correctly; incorrectly preprocessing your data can lead to all downstream tasks (such as your analysis and predictions) being unreasonable, or, even worse, incorrect but untraceable after the data preprocessing stage. This is true of almost all research; scientific understanding depends heavily on the quality of the data.
- Traditional machine learning algorithms—for example, the linear model (Bishop, 2006)—cannot perform well on complicated data and tasks while providing transparent and interpretable solutions. On the other hand, advanced machine learning approaches such

as deep neural networks (Bishop, 2006; Nielsen, 2015) can achieve high performance but often lack sufficient explainability and interpretability in exploration. Hence, we recommend that you keep in mind the tradeoff between performance and explainability and choose the machine learning properly.

- Buchanan et al. (2017) presented a set of approaches for data analysis and visualization. Moreover, they experimented with, and illustrated the effectiveness of, teamwork for data analysis and visualization, which is a practice we adhere to in the Teachers in Social Media project and recommend to other educational researchers.

CONCLUSION

Advancements in technology and communication, among many other social aspects of human life, have changed education in the 21st century quite significantly. Many teachers and students rely on curated resources shared on online social media platforms for their daily classroom instruction and practices. Therefore, it is of great importance for education research to incorporate rich data from online social media. Nevertheless, the current literature does not offer a unified and systematic approach for incorporating data from online social media. This chapter is an attempt to fill this gap. In particular, we presented a roadmap on how educational researchers can effectively incorporate into their academic research the rich online social media available. The proposed framework consists of three major components. In the project initialization component, we shed light on initial planning and strategies on how to commence the project effectively. Next, we discussed data collection methods. Finally, we presented some tools and techniques to utilize and explore the data. We hope the presented roadmap can be used as a blueprint to enhance the quality of education research.

NOTE

1. We refer the interested reader to our publications available at: https://www.teachersinsocialmedia.com/our-work-1

ACKNOWLEDGMENTS

Research reported in this paper was supported by the Defense Advanced Research Projects Agency under project ID number (10332.02 RaCHem PHI) and the National Science Foundation (NSF) under grant number IIS1845081.

In memory of Karen King (1971-2019), math educator, early supporter of this work, and our NSF program officer.

REFERENCES

Abe, S. (2010). Feature selection and extraction. In *Support vector machines for pattern classification* (pp. 331–341). London, England: Springer.

Agasisti, T., & Bowers, A. J. (2017). Data analytics and decision making in education: Towards the educational data scientist as a key actor in schools and higher education institutions. In *Handbook of contemporary education economics* (pp. 184–210). Northampton, MA: Edward Elgar.

Aggarwal, C. C., & Yu, P. S. (2001, May). Outlier detection for high dimensional data. In *ACM Sigmod Record* (Vol. 30, No. 2, pp. 37–46). New York, NY: ACM.

Anderson, J., & Rainie, L. (2012). *The future of big data.* Retrieved from Pew Research Center website: http://www.pewinternet.org/2012/07/20/the-future-of-big-data/

Austin, C. C. (2018). A path to big data readiness. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4844–4853). New York, NY: IEEE.

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V. K., Murphy, F., Nurnberger, A., . . . Whyte, A. (2017). Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, *18*(2), 77–92.

Baepler, P., & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching and Learning*, *4*(2). Article 17.

Bagdy, L. M., Dennen, V. P., Rutledge, S. A., Rowlett, J. T., & Burnick, S. (2018). Teens and social media: A case study of high school students' informal learning practices and trajectories. In *Proceedings of the 9th International Conference on Social Media and Society* (pp. 241–245). New York, NY: ACM.

Baker, R. S. (2010). Data mining for education. *International Encyclopedia of Education*, *7*(3), 112–118.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. Larusson & B. White (Eds.), *Learning analytics* (pp. 61–75). New York, NY: Springer.

Baker, R. S., & Koedinger, K. R. (2018). Towards demonstrating the value of learning analytics for K-12 education. In D. Niemi, R. D. Pea, B. Saxberg, & R. E. Clark (Eds.), *Learning analytics in education* (pp. 49–62). Charlotte, NC: Information Age.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM Journal of Educational Data Mining*, *1*(1), 3–17.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45–59.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York, NY: Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bouchard, T. J., Jr. (1971). Whatever happened to brainstorming. *The Journal of Creative Behavior*, *5*(3), 182–189.

Bowers, A. J., Bang, A. H., Pan, Y., & Graves, K. E. (2019). *Education Leadership Data Analytics (ELDA): A White Paper Report on the 2018 ELDA Summit.* Teachers College, Columbia University, New York, NY.

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Buchanan, V., Lu, Y., McNeese, N., Steptoe, M., Maciejewski, R., & Cooke, N. (2017). The role of teamwork in the analysis of big data: A study of visual analytics and box office prediction. *Big Data*, *5*(1), 53–66.

Calarco, J. M. (2011). "I need help!": Social class and children's help-seeking in elementary school. *American Sociological Review, 76*(6), 862–882.

Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. *ACM Computing Surveys, 14*, 15.

Chen, P. M., Lee, E. K., Gibson, G. A., Katz, R. H., & Patterson, D. A. (1994). RAID: High-performance, reliable secondary storage. *ACM Computing Surveys, 26*(2), 145–185.

Daly, A. J., Liou, Y.-H., Del Fresno, M., Rehm, M., & Bjorklund, P., Jr. (2019). Educational leadership in the Twitterverse: Social media, social networks and the new social continuum. *Teachers College Record, 121*(14). Retrieved from https://www.tcrecord.org/Content.asp?ContentId=23044

Daniel, B. K. (2017). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology, 50*(1), 101–113.

Dhar, V. (2012). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087–1091.

eMarketer. (2019). Number of social media users worldwide from 2010 to 2021 (in billions). *Statista—The Statistics Portal*. Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Frank,* K. A., & Torphy, * K. T. (2019). Social media, who cares? A dialogue between a millennial and a curmudgeon. *Equal authorship. *Teachers College Record, 121*(14). Retrieved from https://www.tcrecord.org/Content.asp?ContentId=23064

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Gibson, C., & Vermeulen, F. (2003). A healthy divide: Subgroups as a stimulus for team learning behavior. *Administrative Science Quarterly, 48*(2), 202–239.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680). Red Hook, NY: Curran Associates.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., . . . Pirahesh, H. (1997). Data cube: A relational aggregation operator generalizing group-by, cross-tab, and subtotals. *Data Mining and Knowledge Discovery, 1*(1), 29–53.

Greenhow, C., Robelia, B., & Hughes, J. E. (2009). Learning, teaching, and scholarship in a digital age: Web 2.0 and classroom research: What path should we take now? *Educational Researcher, 38*(4), 246–259.

Grus, J. (2019). *Data science from scratch: First principles with python*. Sebastopol, CA: O'Reilly Media.

Halverson, R. (2014). Data-driven leadership for learning in the age of accountability. In A. J. Bowers, A. R. Shoho, & B. G. Barnett (Eds.), *Using data in schools to inform leadership and decision making* (pp. 255–267). Charlotte, NC: Information Age.

Halverson, R. (2018). A distributed leadership perspective on information technologies for teaching and learning. In *Second handbook of information technology in primary and secondary education* (pp. 1–17). Cham, Switzerland: Springer.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Waltham, MA: Elsevier.

Ho, A. (2017). Advancing educational research and student privacy in the "big data" era. In *Workshop on big data in education: Balancing the benefits of educational research and student privacy* (pp. 1–18). Washington, DC: National Academy of Education.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126.

Hora, M. T., Bouwma-Gearhart, J., & Park, H. J. (2017). Data driven decision-making in the era of accountability: Fostering faculty data cultures for learning. *The Review of Higher Education*, *40*(3), 391–426.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science* (pp. 1094–1096). Berlin, Germany: Springer.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1546–1557). Stroudsburg, PA: Association for Computational Linguistics.

Karimi, H., & Tang, J. (2019). Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 3432–3442). Stroudsburg, PA: Association for Computational Linguistics.

Karimi, H., Tang, J., & Li, Y. (2018). Toward end-to-end deception detection in videos. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 1278–1283). New York, NY: IEEE.

Keegan, B. C. (2019). *Web data scraping*. Retrieved from GitHub website: https://github.com/CU-ITSS/Web-Data-Scraping-S2019

Kumar, R., & Chadrasekaran, R. M. (2011). Attribute correction-data cleaning using association rule and clustering methods. *International Journal Data Mining & Knowledge Management Process*, *1*(2), 22–32.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 233–246). New York, NY: ACM.

Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, *24*(1), 4–15.

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. Sebastopol, CA: O'Reilly Media.

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). USA: Determination Press.

Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., & Phillips, J. C. (2008). GPU computing. *Proceedings of the IEEE*, *96*(5), 879–899.

Paetsch, F., Eberlein, A., & Maurer, F. (2003). Requirements engineering and agile software development. In *WET ICE 2003. Proceedings. Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003* (pp. 308–313). doi:10.1109/ENABL.2003.1231428

Parnes, S. J., & Meadow, A. (1959). Effects of "brainstorming" instructions on creative problem solving by trained and untrained subjects. *Journal of Educational Psychology*, *50*(4), 171–176.

Pempek, T. A., Yermolayeva, Y. A., & Calvert, S. L. (2009). College students' social networking experiences on Facebook. *Journal of Applied Developmental Psychology*, *30*(3), 227–238.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., . . . Herrera, F. (2016). Data discretization: Taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(1), 5–21.

Rawlinson, J. G. (2017). *Creative thinking and brainstorming*. London, England: Routledge.

Reardon, S. F., Kalogrides, D., & Shores, K. (2019). The geography of racial/ethnic test score gaps. *American Journal of Sociology*, *124*(4), 1164–1221.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Rossini, F. A., & Porter, A. L. (1979). Frameworks for integrating interdisciplinary research. *Research Policy, 8*(1), 70–79.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal, 4*(3), 227–241.

Royston, P. (2005). Multiple imputation of missing values: Update of ice. *The Stata Journal, 5*(4), 527–536.

Rutledge, S. A., Dennen, V. P., & Bagdy, L. M. (2019). Exploring adolescent social media use in a high school: Tweeting teens in a bell schedule world. *Teachers College Record, 121*(14). Retrieved from https://www.tcrecord.org/Content.asp?ContentId=23038

Salganik, M. J. (2017). *Bit by bit: social research in the digital age*. Princeton, NJ: Princeton University Press.

Siemens, G., & Baker, R. S. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). New York, NY: ACM.

Solove, D. J. (2008). *Understanding privacy* (Vol. 173). Cambridge, MA: Harvard University Press.

Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. arXiv:1403.2877. arXiv preprint.

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1017–1024).

TechCrunch. (2018). Number of monthly active Pinterest users from September 2015 to September 2018 (in millions). *In Statista—The Statistics Portal*. Retrieved from https://www.statista.com/statistics/463353/pinterest-global-mau/

Torphy, K. T., & Drake, C. (2019). Educators meet the Fifth Estate: The role of social media in teacher training. *Teachers College Record, 121*(14).

Van Buuren, S. (2018). *Flexible imputation of missing data*. Boca Raton, FL: Chapman and Hall/CRC.

Van der Aalst, W. (2016). Data science in action. In *Process Mining* (pp. 3–23). Berlin, Germany: Springer.

Wang, Y. (2013). Social media in schools: A treasure trove or hot potato? *Journal of Cases in Educational Leadership, 16*(1), 56–64.

Wang, Y. (2016). Big opportunities and big concerns of big data in education. *TechTrends, 60*(4), 381–384.

Wang, Y. (2017a). Education policy research in the big data era: Methodological frontiers, misconceptions, and challenges. *Education Policy Analysis Archives, 25*(94), 1–21.

Wang, Y. (2017b). The social networks and paradoxes of the opt-out movement amid the Common Core State Standards implementation: The case of New York. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, 25*(34), 1–24.

Ward, M. O., Grinstein, G., & Keim, D. (2015). *Interactive data visualization: foundations, techniques, and applications*. Natick, MA: AK Peters/CRC Press.

Wellman, B. (2001). Computer networks as social networks. *Science, 293*(5537), 2031–2034.

Wellman, B., & Haythornthwaite, C. (Eds.). (2008). *The Internet in everyday life*. Hoboken, NJ: Wiley.

Zeide, E. (2017). The structural consequences of big data-driven education. *Big Data, 5*(2), 164–172.

HAMID KARIMI is a fourth-year PhD student of computer science at the Department of Computer Science and Engineering (CSE) at Michigan State University (MSU) and a member of the Data Science and Engineering Lab. He obtained his BS (2010) in computer engineering from the University of Isfahan and his MS (2012) in information technology from Urmia University. His PhD advisor is Dr. Jiliang Tang, an assistant professor of computer science and engineering. His research interests are in machine learning, data mining, natural language processing, social network analysis, and misinformation detection. Mr. Karimi has published several papers in the top conferences and journals on networking, social network analysis, big data, and machine learning, and he has several other papers under review. He received the best paper award for his paper presented at the 2018 IEEE-ACM International Conference on Advances in Social Networks Analysis and Mining, which is one of the top conferences in social network analysis; the MSU Engineering Leadership Fellowship for directing undergraduate summer research and education in the College of Engineering at MSU; and the Best Student Service award from the CSE department at MSU. Moreover, he is an active member of the Teachers in Social Media project at MSU, where he develops data mining and machine learning models and algorithms to characterize instructional resources in online social media. For more information about Mr. Karimi's professional and academic activities please refer to his web page: http://cse.msu.edu/~karimiha/.

TYLER DERR is a fourth-year PhD student in the Department of Computer Science and Engineering at Michigan State University. He obtained his MS in computer science and dual BS degrees in computer science and mathematical sciences from The Pennsylvania State University in 2015 and 2013, respectively. His research interests are in data mining and social network analysis. He has published several papers in these domains at top international conferences, such as ASONAM, CIKM, and ICDM, while receiving travel awards to present his work. Professionally, he serves as a regular reviewer and program committee member for numerous journals and conferences. He is also a member of the Teachers in Social Media project at Michigan State University. For more information about Mr. Derr's professional and academic activities please refer to his web page.

KAITLIN T. TORPHY is the lead researcher and founder of the Teachers in Social Media project at Michigan State University. This project considers the intersection of cloud to class, the nature of resources within virtual resource pools, and implications for equity as educational spaces grow increasingly connected. Dr. Torphy conceptualizes the emergence of a teacherpreneurial guild in which teachers turn to one another for instructional content and resources. She has expertise in teachers' engagement across virtual

platforms, teachers' physical and virtual social networks, and education policy reform. Dr. Torphy was a co-PI and presenter for an American Education Research Association conference convened in October 2018 at Michigan State University on social media and education. She has published work on charter school impacts, curricular reform, and teachers' social networks, and has presented work regarding teachers' engagement within social media at the national and international levels. Her other work examines diffusion of sustainable practices across social networks within The Nature Conservancy. Dr. Torphy earned a PhD in education policy and a specialization in the economics of education from Michigan State University in 2014 and is a Teach for America alumna and former Chicago Public Schools teacher.

KENNETH A. FRANK received his PhD in measurement, evaluation, and statistical analysis from the School of Education at the University of Chicago in 1993. He is MSU Foundation Professor of Sociometrics, professor in Counseling, Educational Psychology and Special Education; and adjunct (by courtesy) in Fisheries and Wildlife and Sociology at Michigan State University. His substantive interests include the study of schools as organizations, social structures of students and teachers and school decision making, and social capital. His substantive areas are linked to several methodological interests: social network analysis, sensitivity analysis and causal inference (http://konfound-it.com),and multilevel models. His recent publications include agent-based models of the social dynamics of the implementation of innovations in organizations, and the implications of social networks for educational opportunity. Recent publications include: *Frank, K. A., & *Xu, R. (2018) Implementation of evidence based practice in human service organizations: Implications from agent-based models. *Journal of Policy Analysis and Management, 37*(4), 4867–4895. *Coequal first authors; and Frank K. A., Lo, Y., Torphy, K., & Kim, J. (2018). Social networks and educational opportunity. In B. Schneider (Ed.), *Handbook of the sociology of education in the 21st century* (pp. 297–316). Handbooks of Sociology and Social Research. Cham, Switzerland: Springer.

JILIANG TANG is an assistant professor in the Department of Computer Science and Engineering at Michigan State University. Before that, he was a research scientist at Yahoo Research. He earned his PhD from Arizona State University in 2015. He has broad interests in social computing, data mining, and machine learning. He was the recipient of the Best Paper Award in ASONAM 2018, the 2018 Criteo Faculty Award, the Best Student Paper Award in WSDM2018, the Best Paper Award for KDD2016, the runner up for the Best KDD Dissertation Award in 2015, and the best paper shortlist of WSDM2013. He is now associate editor of ACM TKDD, ICWSM, and Neurocomputing. He has published his research in highly ranked journals and top conference proceedings, and his work has received thousands of citations and extensive media coverage.