An Introduction to Social Analytics: Concepts and Methods

Shaila M. Miranda, PhD
Expected Publication Date: July, 2018

Proposal

Focus of the Book
Social Analytics deals with the collection, management, and analysis of social data. Social media offer a rich source of social data, but are not the only source of social data. Other sources include company reports and press releases, news media coverage, analyst reports, and conventional websites. Social data typically assume three forms: (1) data about who connects with whom; (2) data about the content of a conversation, typically textual information; (3) the non-textual cues transmitted during the conversation, e.g., non-verbals, typically photographs and videos. All these forms of data typically are unstructured (not in a database or formatted in a database-ready fashion) prior to being subjected to social analytics. The scope of this book is limited to the first two types of data, focusing on concepts, methods, and tools related to social network analysis and text analysis and mining.

This book will introduce readers to concepts and methods relevant to social analytics. The objective of the book is two-fold. First, it will develop the reader's understanding of what they can do with social analytics and of the concepts underlying social analytics. The reader will have the vocabulary necessary to engage with social analytics tools and teams. Second, readers will attain at least a rudimentary familiarity with tools and techniques for social analytics. We will alternate between learning social analytics methods with the help of easy-to-use software and learning to implement those methods using R. Readers therefore will get the most out of this book if they engage with it in a hands-on fashion.
**Prior Knowledge Assumed**

Ideally, the reader will have at least a rudimentary knowledge of statistics. I also assume a fair amount of skill with MS Excel as we will use that to parse and clean text data, visualize networks, and combine relational and conversational data into more complex visuals.

**Target Audiences**

The book targets three audiences. The first is MBA or MS-MIT students in courses related to social analytics. These courses often are taught by MIS departments, but also may be offered by marketing departments, where social media use for advertising and customer relationship management is a growing concern.

The second are PhD students in disciplines ranging from MIS and Marketing to Sociology and Political Science, for whom social media provide research data. If not already being offered, methods classes with a significant component on gathering, storing, and analyzing the unstructured data that is the basis of Social Analytics will be a valuable addition to such PhD programs.

The third audience for the book are analysts and researchers conducting business analysis or scholarly research with social media data. The concepts, methods, and tools covered by the book also will be useful in helping these constituents update their analytics and research skills for the era of big, unstructured data.

**Key Challenges in the Course**

The field of Data Analytics still is evolving. Within this field, the discipline of Social Analytics is particularly novel. Social Analytics draws upon conceptual and methodological advances at the intersection of social and computer sciences and, as I have operationalized it, focuses on social network analysis and text analysis. While myriad books cover the source material for Social Analytics, very few put the material together into a single text. Given the skyrocketing cost of higher education, having students purchase multiple books usually is not a viable option. Often, the style of coverage in such books is intimidating to readers not immersed in the area. The rare book that combines both social network analysis and text analysis does not do so with sufficient technical depth for a Master of Science curriculum.
**Book Approach and Features**

Two preliminary questions need to be addressed. *First, why is it necessary to cover both SNA and text analysis?* While the focus of the course I teach and the book is not exclusively on social media data, it is social media data that currently makes a Social Analytics course and social analytic skills desirable. To best understand social media conversations, one must understand both the structure and content of the conversations. This requires SNA and text analysis skills. Concurrent analysis of social networks and text provides insights such as how sentiment expressed contributes to actors’ network centrality or to partitioning a network into sub-groups, as is depicted in the graph below representing the grouping of actors discussing the Keystone Pipeline on Twitter relative to the sentiment predominantly expressed in their tweets (red=negative; blue=positive; grey=neutral).

![Graph showing network analysis](image)

Further, SNA techniques are used in unsupervised learning algorithms applied to text analysis. Developing an understanding of SNA prior to applying it to text analysis tasks enables students to apply the SNA-based unsupervised learning with greater understanding of the technique.

*Second, why not use best-of-breed books?* One could argue that instructors would be better served using the Wasserman and Strauss for SNA and the Ingersoll et al. for text analysis. There are three reasons why this might not be a desirable course of action. First, it would be impossible to do justice to either book – much less both books – within a semester-long course. Given that, the combined cost of the two (or other combinations) books to the
students would not be justified. Second, the depth of best-of-breed books, pace, and manner of coverage would be daunting to college of business students. These books tend to expect a level of comfort with mathematical notation and/or programming that is atypical of such students. Third, the disparate styles of coverage across the two books would steepen the learning curve for students.

My pedagogical approach is to introduce the reader to the concepts underlying the various social analytics methods using easy-to-use tools such as NodeXL and LIWC, which require no programming background. Using the tools makes the concepts tangible. These easy-to-use tools are limited in their ability to handle large, complex, real-world data though and have limited appeal to prospective recruiters. So, once the reader is familiar with the concepts, I then introduce R packages that provide equivalent and advanced capabilities relative to the basic tools. Having taught a Social Analytics course twice using the approach of iterating between the accessible low-end tools and R, I have found students react well to it. They consistently have advised me to continue to use this approach rather than introducing students to the social analytics concepts and to R programming together.

I use R (instead of SAS Enterprise Miner or IBM SPSS Modeler) for three reasons. First, R is versatile for both social network analysis and a range of text analysis and mining tasks. Since it is open source, tools implementing novel methods tend to be developed first within the R environment, providing R users with a first-mover advantage. A competing environment would be Python. However, while Python would handle tasks specific to Social Analytics, R is able to perform a wider range of statistical analysis tasks than Python and therefore is a more useful environment for students to master than Python. Second, being open source, R also is free. This has the dual advantage of keeping down the costs of higher education and permitting students to continue to develop their skills beyond the course. Third, the availability of Microsoft R Server now ensures the commercial viability of R, and therefore its appeal to prospective student recruiters.

Since few students/readers will already have facility with R, in the appendices, I have provided enough of an overview of R to get readers started. Rather than provide readers with complete documentation on any command or package (which is available off the
internet), I have chosen to introduce select commands and packages to get readers started. I have chosen also to introduce some of the idiosyncratic things you can do in R (compared with SPSS or SAS). I also provide pointers to R tutorials available elsewhere.

Through the book, I try to motivate techniques introduced with examples of their use in business and non-business contexts. Chapters conclude with exercises that challenge students/readers to synthesize and apply the knowledge and skills they have developed through the chapter. Where I have been able to anticipate problems readers may encounter while using the tools covered, I have attempted to document those problems and provide solutions.

The pace of evolution of social analytics methods and tools makes a low-cost eBook an ideal solution for students and other audiences. An eBook offers the additional advantage of permitting students to easily copy the code necessary for executing social analytics tasks in R.

**Table of Contents**

**Chapters**

**Part I: Introduction**

1. What is social analytics?
   a. The meaning of “social”
   b. The meaning of “analytics”
   c. Skills required for conducting social analytics
   d. What can we do with social analytics
   e. Focus of the book

2. The need for social analytics
   a. Business use cases (2+ examples of describing a social space and predicting organizational outcomes)
   b. Public policy use case

**Part II: Social Network Analysis**

3. Introduction to social network analysis (SNA)
   a. Origins and applications
b. Vocabulary and concepts

c. Factors accounting for connections

d. Types of social networks

e. Setting up social network data with NodeXL

f. Visualizing social networks with NodeXL

g. Graph metrics with NodeXL

h. Ego metrics with NodeXL

4. Visualizing social networks and SNA metrics with R

a. SNA packages

b. Opening, preparing, and inspecting SNA data

c. Visualizing networks with R

d. Graph metrics with R

e. Ego metrics with R

5. Advanced SNA concepts

a. Understanding homophily and subgroup analysis with NodeXL

b. Cluster analysis with NodeXL

c. Distributional properties of network data

d. Gathering social media data with NodeXL

6. Network subgroup analysis with R

a. Detecting homophily with R

b. Community detection with R

Part III: Text Analysis

7. Introduction to text analysis and mining

a. Origins and applications (e.g., sentiment analysis)

b. Vocabulary and concepts

c. Overview of different approaches

d. Count-based approaches with word cloud generators

e. Dictionary-based approaches with LIWC

8. Basic text analysis with R

a. Text analysis packages
b. Text data objects in R

c. Pre-processing text data

d. Sentiment analysis with R and non-R dictionaries

e. Simple and comparison word clouds

9. Text mining concepts and methods

a. Text mining tasks

b. Text mining methods

c. A simple supervised learning task (sentiment analysis) with WEKA

d. A simple unsupervised learning task (clustering) with WEKA

10. Information extraction with R

a. Natural language processing packages

b. Tagging text for different types of entities

c. Analyzing tagged text

11. Machine learning for text analysis with R

a. Machine learning packages

b. Cluster analysis

c. Network analysis

Part IV: Conclusion

12. Ethical issues

a. Data ownership

b. Consequences of analytics use to individuals and society

Appendices

Appendix A: Introduction to R (data structures, basic programming)

Appendix B: Statistical analysis with R (descriptives, basic inferential statistics)

Appendix C: Scraping Twitter with R

Appendix D: Scraping Facebook with R

Additional E: R resources
Bio

Shaila Miranda is Associate Professor of MIS at the University of Oklahoma's Price College of Business. She obtained her doctorate from the University of Georgia, focusing on Management Information Systems with a minor in Computer Science. She also has an M.A. in Sociology from Columbia University, and a Master of Management Studies and a B.A. in Psychology from the University of Bombay.

Professor Miranda’s research, which has been published in *MIS Quarterly* and *Information Systems Research*, most recently applies the core methods of Social Analytics to investigate topics such as how Fortune 50 companies describe their use of social media in press releases and how traditional news media versus social media cover public policy issues. Professor Miranda has served as Senior Editor and Associate Editor for *Information Systems Research* and as Associate Editor for *MIS Quarterly*. 