Advances in Analytics and Data Science

Murugan Anandarajan
Chelsey Hill
Thomas Nolan

Practical Text Analytics

Maximizing the Value of Text Data

Springer

Advances in Analytics and Data Science

Volume 2

Series Editors
Ramesh Sharda
Oklahoma State University, Stillwater, OK, USA

Hsinchun Chen
University of Arizona, Tucson, AZ, USA

More information about this series at http://www.springer.com/series/15876

Murugan Anandarajan • Chelsey Hill
Thomas Nolan

Practical Text Analytics

Maximizing the Value of Text Data

Murugan Anandarajan
LeBow College of Business
Drexel University
Philadelphia, PA, USA

Chelsey Hill
Feliciano School of Business
Montclair State University
Montclair, NJ, USA

Thomas Nolan
Mercury Data Science
Houston, TX, USA

ISSN 2522-0233 ISSN 2522-0241 (electronic)
Advances in Analytics and Data Science
ISBN 978-3-319-95662-6 ISBN 978-3-319-95663-3 (eBook)
https://doi.org/10.1007/978-3-319-95663-3

Library of Congress Control Number: 2018955905

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my aunts and uncle—MA
To my angel mother, Deborah—CH
To my dad—TN

Preface

The oft-cited statistic that "80% of data is unstructured" reminds industry leaders and analytics professionals about the vast volume of untapped text data resources. Predictably, there has been an increasing focus on text analytics to generate information resources. In fact, the expected growth in this market is projected to be $5.93 billion by 2020![1]

Whenever businesses capture text data, they want to capitalize on the information hidden within. Beginning with early adopters in the intelligence and biomedical communities, text analytics has expanded to include applications across industries, including manufacturing, insurance, healthcare, education, safety and security, publishing, telecommunications, and politics. The broad range of applied text analytics requires practitioners in this field.

Our goal is to democratize text analytics and increase the number of people using text data for research. We hope this book lowers the barrier of entry for analyzing text data, making it more accessible for people to uncover value-added text information.

This book covers the elements involved in creating a text mining pipeline. While analysts will not use every element in every project, each tool provides a potential segment in the final pipeline. Understanding the options is key to choosing the appropriate elements in designing and conducting text analysis.

The book is divided into five parts. The first part provides an overview of the text analytics process by introducing text analytics, discussing the relationship with content analysis, and providing a general overview of the process.

Next, the chapter moves on to the actual practice of text analytics, beginning with planning the project. The next part covers the methods of data preparation and preprocessing. Once the data is prepared, the next step is the analysis. Here, we describe the array of analysis options. The part concludes with a discussion about reporting options, indicating the benefits of various choices for convincing others about the value of the analysis.

[1] http://www.marketsandmarkets.com/PressReleases/text-analytics.asp

The last part of the book demonstrates the use of various software programs and programming languages for text analytics. We hope these examples provide the reader with practical examples of how information hidden within text data can be mined.

Philadelphia, PA, USA — Murugan Anandarajan
Montclair, NJ, USA — Chelsey Hill
Houston, TX, USA — Thomas Nolan

Acknowledgments

The authors wish to acknowledge the invaluable contributions of several individuals to the preparation of the manuscript of this book.

Diana Jones, Director of the Center for Business Analytics and the Dornsife Office for Experiential Learning, at the LeBow College of Business, Drexel University, for her chapter on *Storytelling Using Text Data.*

Jorge Fresneda Fernandez, Assistant Professor of Marketing at the Martin Tuchman School of Management, New Jersey Institute of Technology, for his chapters on *Latent Semantic Analysis (LSA) in Python* and *SAS Visual Text Analytics.*

We thank Diana and Jorge for their expertise and invaluable contributions to this book.

We also thank Irena Nedelcu, Rajiv Nag, and Stacy Boyer, all of the LeBow College of Business, Drexel University, for providing valuable comments on various chapters.

Our appreciation to Matthew Amboy and his team at Springer who made the publication of this book possible.

Murugan Anandarajan
Chelsey Hill
Thomas Nolan

Contents

About the Authors

Murugan Anandarajan is a Professor of MIS at Drexel University. His current research interests lie in the intersections of crime, IoT, and analytics. His work has been published in journals such as *Decision Sciences*, *Journal of Management Information Systems*, and *Journal of International Business Studies*. He co-authored eight books, including *The Internet and Workplace Transformation* (2006) and its follow-up volume, *The Internet of People, Things and Services* (2018). He has been awarded over $2.5 million in research grants from various government agencies including the National Science Foundation, the US Department of Justice, the National Institute of Justice, and the State of PA.

Chelsey Hill is an Assistant Professor of Business Analytics in the Information Management and Business Analytics Department of the Feliciano School of Business at Montclair State University. She holds a BA in Political Science from the College of New Jersey, an MS in Business Intelligence from Saint Joseph's University, and a PhD in Business Administration with a concentration in Decision Sciences from Drexel University. Her research interests include consumer product recalls, online consumer reviews, safety and security, public policy, and humanitarian operations. Her research has been published in the *Journal of Informetrics* and the *International Journal of Business Intelligence Research*.

Tom Nolan completed his undergraduate work at Kenyon College. After Kenyon, he attended Drexel University where he graduated with an MS in Business Analytics. From there, he worked at Independence Blue Cross in Philadelphia, PA, and Anthem Inc. in Houston, TX. Currently, he works with all types of data as a Data Scientist for Mercury Data Science.

List of Abbreviations

ANN	Artificial neural networks
BOW	Bag-of-words
CA	Content analysis
CTM	Correlated topic model
df	Document frequency
DM	Data mining
DTM	Document-term matrix
HCA	Hierarchical cluster analysis
idf	Inverse document frequency
IoT	Internet of Things
KDD	Knowledge discovery in databases
KDT	Knowledge discovery in text
kMC	k-means clustering
kNN	k-nearest neighbors
LDA	Latent Dirichlet allocation
LSA	Latent semantic analysis
LSI	Latent semantic indexing
NB	Naive Bayes
NLP	Natural language processing
NN	Neural networks
OM	Opinion mining
pLSI	Probabilistic latent semantic indexing
RF	Random forest
SA	Sentiment analysis
sLDA	Supervised latent Dirichlet allocation
STM	Structural topic model
SVD	Singular value decomposition
SVM	Support vector machines
TA	Text analytics
TDM	Term-document matrix
tf	Term frequency
tfidf	Term frequency-inverse document frequency
TM	Text mining

List of Figures

List of Tables

Chapter 1
Introduction to Text Analytics

Abstract In this chapter we define text analytics, discuss its origins, cover its current usage, and show its value to businesses. The chapter describes examples of current text analytics uses to demonstrate the wide array of real-world impacts. Finally, we present a process road map as a guide to text analytics and to the book.

Keywords Text analytics · Text mining · Data mining · Content analysis

1.1 Introduction

Recent estimates maintain that 80% of all data is text data. A recent article in *USA Today* asserts that even the Internal Revenue Service is using text, in the form of US citizens' social media posts, to help them make auditing decisions.[1] The importance of text data has created a veritable industry comprised of firms dedicated solely to the storage, analysis, and extraction of text data. One such company, Crimson Hexagon, has created the world's largest database of text data from social media sites including one trillion public social media posts spanning more than a decade.[2]

1.2 Text Analytics: What Is It?

Hearst (1999a, b) defines text analytics, sometimes referred to as text mining or text data mining, as the automatic discovery of new, previously unknown, information from unstructured textual data. The terms text analytics and text mining are often used interchangeably. Text mining can also be described as the process of deriving

[1] Agency breaking law by mining social media. (2017, 12). USA Today, 146, 14–15.

[2] http://www.businessinsider.com/analytics-firm-crimson-hexagon-uses-social-media-to-predict-stock-movements-2017-4

M. Anandarajan et al., *Practical Text Analytics*, Advances in Analytics and Data Science 2, https://doi.org/10.1007/978-3-319-95663-3_1

high-quality information from text. This process involves three major tasks: information retrieval (gathering the relevant documents), information extraction (unearthing information of interest from these documents), and data mining (discovering new associations among the extracted pieces of information).

Text analytics has been influenced by many fields and has made significant contributions to many disciplines. Modern text analytics applications span many disciplines and objectives. In addition to having multidisciplinary origins, text analytics continues to have applications in and make advancements to many fields of study. It intersects many research fields, including:

- Library and information science
- Social sciences
- Computer science
- Databases
- Data mining
- Statistics
- Artificial intelligence
- Computational linguistics

Although many of these areas contributed to modern day text analytics, Hearst (1999a, b) posited that text mining came to fruition as an extension of data mining. The similarities and differences between text mining and data mining have been widely discussed and debated (Gupta and Lehal 2009; Hearst 2003). Data mining uses structured data, typically from databases, to uncover "patterns, associations, changes, anomalies, and significant structures" (Bose 2009, p. 156). The major difference between the two is the types of data that they use for analysis. Data mining uses structured data, found in most business databases, while text mining uses unstructured or semi-structured data from a variety of sources, including media, the web, and other electronic data sources. The two methods are similar because they (i) are equipped for handling large data sets; (ii) look for patterns, insight, and discovery; and (iii) apply similar or the same techniques. Additionally, text mining draws on techniques used in data mining for the analysis of the numeric representation of text data.

The complexities associated with the collection, preparation, and analysis of unstructured text data make text analytics a unique area of research and application. Unstructured data are particularly difficult for computers to process. The data itself cover a wide range of possibilities, each with its own challenges. Some examples of the sources of text data used in text mining are blogs, web pages, emails, social media, message board posts, newspaper articles, journal articles, survey text, interview transcripts, resumes, corporate reports and letters, insurance claims, customer complaint letters, patents, recorded phone calls, contracts, and technical documentation (Bose 2009; Dörre et al. 1999).

1.3 Origins and Timeline of Text Analytics

Text-based analysis has its roots in the fields of computer science and the social sciences as a means of converting qualitative data into quantitative data for analysis. The field of computer science is in large part responsible for the text analytics that we know today. In contrast, the social sciences built the foundation of the analysis of text as a means of understanding literature, discourse, documents, and surveys. Text analytics combines the computational and humanistic elements of both fields and uses technology to analyze unstructured data text data by "turning text into numbers." The text analytics process includes the structuring of input text, deriving patterns within the structured data, and evaluating and interpreting the output.

Figure 1.1 presents a timeline of text analytics by decade. In the 1960s, computational linguistics was developed to describe computer-aided natural language processing (Miner et al. 2012). Natural language processing techniques are outlined in Chaps. 5 and 6. During this decade, content analysis, the focus of Chap. 2, emerged in the social sciences as a means of analyzing a variety of content, including text and media (Krippendorff 2012).

In the late 1980s and early 1990s, latent semantic indexing, or latent semantic analysis, introduced in Chap. 6 arrived as a dimension reduction and latent factor identification method applied to text (Deerwester et al. 1990; Dumais et al. 1988). At this time, knowledge discovery in databases developed as a means of making sense of data (Fayyad et al. 1996; Frawley et al. 1992). Building on this advancement, Feldman and Dagan (1995) created a framework for text, known as knowledge discovery in texts to do the same with unstructured text data.

Data mining emerged in the 1990s as the analysis step in the knowledge discovery in databases process (Fayyad et al. 1996). In the 1990s, machine learning methods, covered in Chaps. 7 and 9, gained prominence in the analysis of text data (Sebastiani 2002). Around that time, text mining became a popular buzzword but lacked practitioners (Hearst 1999a, b). Nagarkar and Kumbhar (2015) reviewed text mining-related publications and citations from 1999 to 2013 and found that the number of publications consistently increased throughout this period.

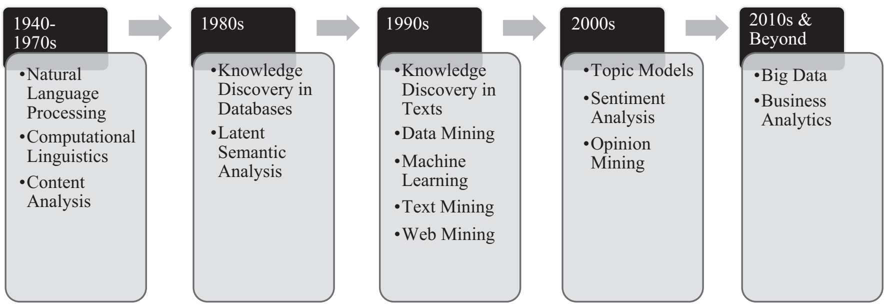

Fig. 1.1 Text analytics timeline

Building on a probabilistic form of latent semantic analysis (Hofmann 1999) introduced the late 1990s, topic models, discussed in Chap. 8, were created in the early 2000s with the development of the latent Dirichlet allocation model (Blei et al. 2002, 2003). Around the same time, sentiment analysis (Nasukawa and Yi 2003) and opinion mining (Dave et al. 2003), the focus of Chap. 10, were introduced as methods to understand and analyze opinions and feelings (Liu 2012).

The 2010s were the age of big data analytics. During this period, the foundational concepts preceding this time were adapted and applied to big data. According to Davenport (2013), although the field of business analytics has been around for over 50 years, the current era of analytics, Analytics 3.0, has witnessed the widespread use of corporate data for decision-making across many organizations and industries. More specifically, four key features define the current generation of text analytics and text mining: foundation, speed, logic, and output. As these characteristics indicate, text analysis and mining are data driven, conducted in real time, rely on probabilistic inference and models, and provide interpretable output and visualization (Müller et al. 2016). According to *IBM Tech Trends Report* (2011), business analytics was deemed one of the major trends in technology in the 2010s (Chen et al. 2012).

One way to better understand the area of text analytics is by examining relevant journal articles. We analyzed 3,264 articles, 2,315 published in scholarly journals and 949 published in trade journals. In our article sample, there are 704 journals, 187 trade journals and 517 scholarly journals. Each article belongs to one or more article classifications. The articles in our sample have 170 distinct classifications based on information about them such as industry, geographical location, content, and article type. Figure 1.2 displays the total number of text analytics-related

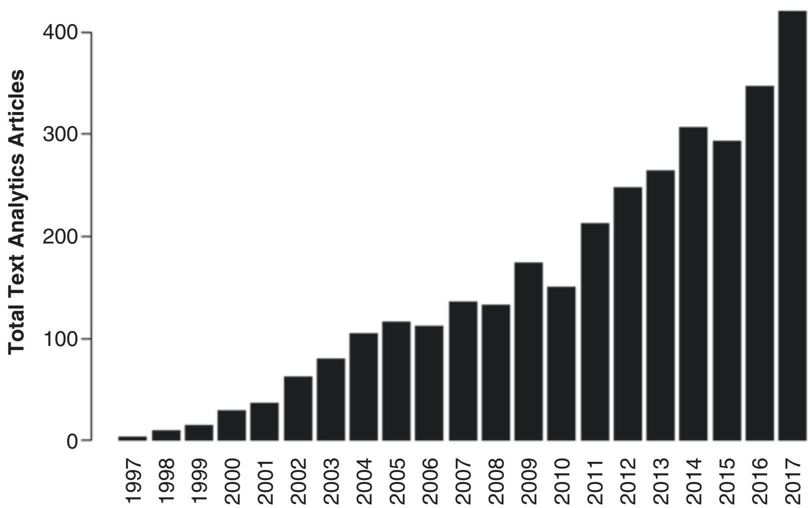

Fig. 1.2 Frequency of text analytics articles by year

articles over time. As the figure illustrates, there has been a considerable growth in the publications since 1997, with 420 articles about text mining published in 2017.

1.4 Text Analytics in Business and Industry

Text analytics has numerous practical uses, including but not limited to email filtering, product suggestions, fraud detection, opinion mining, trend analysis, search engines, and bankruptcy predictions (Talib et al. 2016). The field has a wide range of goals and objectives, including understanding semantic information, text summarization, classification, and clustering (Bolasco et al. 2005). We explore some examples of text analytics and text mining in the areas of business and industry. Text analytics applications require clear, interpretable results and actionable outcomes to achieve the desired result. Indeed, the technique can be used in almost every business department to increase productivity, efficiency, and understanding.

Figure 1.3 is a word cloud depicting the most popular terms and phrases in the text analytics-related articles' abstracts and titles. As the figure shows, research and applications in this area are as diverse as the area's history. The close relationship between data mining and text analytics is also evident in the figure.

Fig. 1.3 Word cloud of the titles and abstracts of articles on text analytics