Big Crisis Data
Social Media in Disasters and Time-Critical Situations

Social media is an invaluable source of time-critical information during a crisis. However, emergency response and humanitarian relief organizations that would like to use this information struggle with an avalanche of social media messages that exceeds human capacity to process. Emergency managers, decision makers, and affected communities can make sense of social media through a combination of machine computation and human compassion-expressed by thousands of digital volunteers who publish, process, and summarize potentially life-saving information.

This book brings together computational methods from many disciplines: natural language processing, semantic technologies, data mining, machine learning, network analysis, human-computer interaction, and information visualization, focusing on methods that are commonly used for processing social media messages under time-critical constraints, and offering more than 500 references to in-depth information.

CARLOS CASTILLO is a researcher on social computing. He is a web miner with a background on information retrieval, and has been influential in the areas of web content quality and credibility. He has co-authored more than seventy publications in top-tier international conferences and journals, a monograph on adversarial web search, and a book on information and influence propagation.
Dedicated to the people who spend countless hours in front of digital devices helping others, sharing their time, energy, and skills.
Big Crisis Data
Social Media in Disasters and Time-Critical Situations

CARLOS CASTILLO
Contents

Preface xi
Acknowledgments xi

1 Introduction 1
1.1 “Sirens going off now!! Take cover . . . be safe!” 2
1.2 What Is a Disaster? 3
1.3 Information Flows in Social Media 5
1.4 The Data Deluge 8
1.5 Requirements: “Big Picture” Versus “Actionable Insights” 9
1.6 Organizational Challenges 11
1.7 Scope and Organization of This Book 13
1.8 Further Reading and Online Appendix 16

2 Volume: Data Acquisition, Storage, and Retrieval 18
2.1 Social Media Data Sizes 18
2.2 Data Acquisition 22
2.3 Postfiltering and De-Duplication 28
2.4 Data Representation / Feature Extraction 29
2.5 Storage and Indexing 31
2.6 Research Problems 32
2.7 Further Reading 34

3 Vagueness: Natural Language and Semantics 35
3.1 Social Media Is Conversational 36
3.2 Text Preprocessing 37
3.3 Sentiment Analysis 41
3.4 Named Entities 42
3.5 Geotagging and Geocoding 44
### Contents

3.6 Extracting Structured Information 46  
3.7 Ontologies for Explicit Semantics 47  
3.8 Research Problems 48  
3.9 Further Reading 49  

4 Variety: Classification and Clustering 51  
4.1 Content Categories 52  
4.2 Supervised Classification 57  
4.3 Unsupervised Classification / Clustering 63  
4.4 Research Problems 66  
4.5 Further Reading 66  

5 Virality: Networks and Information Propagation 68  
5.1 Crisis Information Networks 69  
5.2 Cascading of Crisis Information 73  
5.3 User Communities and User Roles 76  
5.4 Research Problems 78  
5.5 Further Reading 78  

6 Velocity: Online Methods and Data Streams 79  
6.1 Stream Processing 80  
6.2 Analyzing Temporal Data 81  
6.3 Event Detection 83  
6.4 Event-Detection Methods 85  
6.5 Incremental Update Summarization 90  
6.6 Domain-Specific Approaches 92  
6.7 Research Problems 94  
6.8 Further Reading 94  

7 Volunteers: Humanitarian Crowdsourcing 96  
7.1 Digital Volunteering 97  
7.2 Organized Digital Volunteering 99  
7.3 Motivating Volunteers 102  
7.4 Digital Volunteering Tasks 104  
7.5 Hybrid Systems 107  
7.6 Research Problems 108  
7.7 Further Reading 109  

8 Veracity: Misinformation and Credibility 110  
8.1 Emergencies, Media, and False Information 111  
8.2 Policy-Based Trust and Social Media 113  
8.3 Misinformation and Disinformation 114
## Contents

8.4 Verification Practices 115  
8.5 Automatic Credibility Analysis 117  
8.6 Research Problems 121  
8.7 Further Reading 122  

9 Validity: Biases and Pitfalls of Social Media Data 123  
9.1 Studying the “Offline” World Using “Online” Data 124  
9.2 The Digital Divide 126  
9.3 Content Production Issues 128  
9.4 Infrastructure and Technological Factors 129  
9.5 The Geography of Events and Geotagged Social Media 130  
9.6 Evaluation of Alerts Triggered from Social Media 134  
9.7 Research Problems 135  
9.8 Further Reading 136  

10 Visualization: Crisis Maps and Beyond 138  
10.1 Crisis Maps 138  
10.2 Crisis Dashboards 142  
10.3 Interactivity 145  
10.4 Research Problems 149  
10.5 Further Reading 150  

11 Values: Privacy and Ethics 152  
11.1 Protecting the Privacy of Individuals 153  
11.2 Intentional Human-Induced Disasters 156  
11.3 Protecting Citizen Reporters and Digital Volunteers 157  
11.4 Ethical Experimentation 158  
11.5 Giving Back and Sharing Data 159  
11.6 Research Problems 161  
11.7 Further Reading 162  

12 Conclusions and Outlook 164  
12.1 The Quality of Crisis Information 165  
12.2 Peer Production of Crisis Information 166  
12.3 Technologies for Crisis Communications in Social Media 167  
12.4 User-Generated Images, Video, and Aerial Photography 167  
12.5 Outlook 168  

Bibliography 171  
Index 209  
Terms and Acronyms 211
Preface

Social media is an invaluable source of time-critical information during a crisis. However, emergency response and humanitarian relief organizations that would like to use this information struggle with an avalanche of social media messages – often exceeding human capacity to process.

Emergency managers, decision makers, and affected communities can make sense of social media through a combination of machine computation and human compassion. Machine computation takes many forms, including natural language processing, semantic technologies, data mining, machine learning, network analysis, human-computer interaction, and information visualization. Human compassion is expressed by thousands of digital volunteers who publish, process, and summarize potentially life-saving information.

This book brings together computational methods from many disciplines, focusing on methods that are commonly used for processing social media messages under time-critical constraints, and offering over 500 references to in-depth information.

Researchers and computer science students can read this book as an extended survey of methods to be improved, extended, or built upon through research. It can also be used in an integrative, applied course or seminar on mining the real-time Web.

Developers and practitioners can read this book as an overview of composable state-of-the-art methods that can be used to architect solutions for handling time-critical social media data. The discussion uses examples from current social media platforms, which of course may merge, become abandoned, or disappear in the future, but every effort has been made to make the discussion platform-agnostic.

Emergency relief and humanitarian response are fascinating topics that should attract some of the best minds in the scientific and technical
This book is an invitation for computer scientists and technologists who want to apply their skills to help disaster-affected communities by providing information, a basic need during disaster response.

Check out the website at www.BigCrisisData.org
Acknowledgments

The Qatar Computing Research Institute (QCRI) supported me during most of the writing of this book. Sapienza University of Rome was also kind to host me during part of the writing.

Special thanks to Patrick Meier for introducing me to “Big Crisis Data” concepts, including digital humanitarianism, and for his contagious passion for social innovation.

My colleagues Marcelo Mendoza and Bárbara Poblete, coauthors in Mendoza et al. (2010) and Castillo et al. (2011, 2013), were the first to get me interested in information credibility during disasters, after their experience with the earthquake in 2010 in Chile. I want to thank Muhammad Imran, Sarah Vieweg, and Fernando Diaz for our work together and joint survey (Imran et al., 2015) which formed the starting point for Chapters 1, 2, 3, and 6.

I am very thankful to PhD students who codeveloped many of the ideas in this book, including Aditi Gupta, Alexandra Olteanu, Hemant Purohit, Jakob Rogstadious, Soudip Chowdhoury, and Irina Temnikova during her postdoc. Thanks to Leysia Palen for her advice and all her contributions to this topic over more than a decade, both directly and through her students. Thanks to Jaideep Srivastava for his support and guidance during my last year at QCRI, and for coining the “machine computation and human compassion” phrase.

I asked colleagues to review early drafts of this book: Ken Anderson, Fabricio Benevenuto, Luis Capelo, Fernando Diaz, Hamed Haddadi, Muhammad Imran, Ponnurangam Kumaraguru, Alexandra Olteanu, Leysia Palen, Jürgen Pfeffer, Robert Power, Hemant Purohit, Kate Starbird, and Ingmar Weber. I am very thankful for their expert advice and detailed feedback, and of course I am responsible for all errors and omissions in this book.
Acknowledgments

Cambridge University Press editor Lauren Cowles was patient and persistent, and her dedication was invaluable for this project.

Last but not least, I would like to thank my wife Fabiola for her unconditional support during the writing of this book and almost two decades of joint adventures.