# Sentiment Analysis

*Mining Opinions, Sentiments, and Emotions*

Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes. This fascinating problem is increasingly important in business and society. It offers numerous research challenges but promises insight useful to anyone interested in opinion analysis and social media analysis.

This book gives a comprehensive introduction to the topic from a primarily natural language processing point of view to help readers understand the underlying structure of the problem and the language constructs that are commonly used to express opinions and sentiments. It covers all core areas of sentiment analysis; includes many emerging themes, such as debate analysis, intention mining, and fake-opinion detection; and presents computational methods to analyze and summarize opinions. It will be a valuable resource for researchers and practitioners in natural language processing, computer science, management sciences, and the social sciences.

Bing Liu is a professor of computer science at the University of Illinois at Chicago. His current research interests include sentiment analysis and opinion mining, data mining, machine learning, and natural language processing. He has published extensively in top conferences and journals, and his research has been cited on the front page of the *New York Times*. He is also the author of two books: *Sentiment Analysis and Opinion Mining* (2012) and *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (first edition, 2007; second edition, 2011). He currently serves as the Chair of ACM SIGKDD and is an IEEE Fellow.

# Sentiment Analysis

*Mining Opinions, Sentiments, and Emotions*

BING LIU

*University of Illinois at Chicago*

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107017894



First published 2015

Printed in the United States of America

*A catalog record for this publication is available from the British Library.*

*Library of Congress Cataloging in Publication Data*
Liu, Bing, 1963–
Sentiment analysis : mining opinions, sentiments, and emotions / Bing Liu.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-01789-4 (hardback)
1. Natural language processing (Computer science) 2. Computational linguistics. 3. Public opinion – Data processing. 4. Data mining. I. Title.
QA76.9.N38L58 2015
006.3′12–dc23 2014036113

ISBN 978-1-107-01789-4 Hardback

# Contents

# Preface

Opinion and sentiment and their related concepts, such as evaluation, appraisal, attitude, affect, emotion, and mood, are about our subjective feelings and beliefs. They are central to human psychology and are key influencers of our behaviors. Our beliefs and perceptions of reality, as well as the choices we make, are to a considerable degree conditioned on how others see and perceive the world. For this reason, our views of the world are very much influenced by others' views, and whenever we need to make a decision, we often seek out others' opinions. This is true not only for individuals but also for organizations. From an application point of view, we naturally want to mine people's opinions and feelings toward any subject matter of interest, which is the task of *sentiment analysis*. More precisely, sentiment analysis, which is also called *opinion mining*, is a field of study that aims to extract opinions and sentiments from natural language text using computational methods.

The inception and rapid growth of sentiment analysis coincide with those of social media on the web, such as reviews, forum discussions, blogs, and microblogs, because for the first time in human history, we now have a huge volume of opinion data recorded in digital forms. These data, also called *user-generated content*, prompted researchers to mine them to discover useful knowledge. This naturally led to the problem of sentiment analysis or opinion mining because these data are full of opinions. That these data are full of opinions is not surprising, because the primary reason why people post messages on social media platforms is to express their views and opinions, and therefore sentiment analysis is at the very core of social media analysis. Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, web mining, and information retrieval. In fact, the research has spread from computer science to management science and social science because of its importance to business and society as a whole. In recent years, industrial activities surrounding sentiment analysis have also thrived. Numerous start-ups have emerged. Many large corporations, for example, Microsoft, Google, Hewlett-Packard, and Adobe, have also built their

own in-house systems. Sentiment analysis systems have found applications in almost every business, health, government, and social domain.

Although no silver bullet algorithm can solve the sentiment analysis problem, many deployed systems are able to provide useful information to support real-life applications. I believe it is now a good time to document the knowledge that we have gained in research, and, to some extent, in practice, in a book. Obviously, I don't claim that I know everything that is happening in the industry, as businesses do not publish or disclose their algorithms. However, I have built a sentiment analysis system myself in a start-up company and served clients on projects involving social media data sets in a large variety of domains. Over the years, many developers of sentiment analysis systems in the industry have also told me roughly what algorithms they were using. Thus, I can claim that I have a reasonable knowledge of practical systems and their capabilities and firsthand experience in solving real-life problems. I try to pass along those nonconfidential pieces of information and knowledge in this book.

In writing this book, I aimed to take a balanced approach, analyzing the sentiment analysis problem from a linguistic angle to help readers understand the underlying structure of the problem and the language constructs commonly used to express opinions and sentiments and presenting computational methods to analyze and summarize opinions. Like many natural language processing tasks, most published computational techniques use machine learning or data mining algorithms with the help of text-specific clues or features. However, if we only focus on such computational algorithms, we will miss the deep insights of the problem, which in turn will hinder our progress on the computational front. Most existing machine learning algorithms are black boxes. They do not produce human-interpretable models. When something goes wrong, it is hard to know the cause and how to fix it.

In presenting linguistic constructs and perspectives, I do not follow the linguistic tradition in writing because the knowledge and the way that the knowledge is presented in the traditional linguistics literature are mainly for people to understand rather than for computers to operationalize to solve real-life problems. Although the knowledge of human beings and instructions for computers can largely intersect, they also have major differences. As a case in point, when I was working on the problem of mining opinions from conditional sentences, I read several linguistics books about conditionals. However, to my surprise, I found almost no linguistic knowledge that can be operationalized computationally to help solve the problem. I believe this is partially because the current computation technologies are not mature enough to have the same understanding capability as people and partially because much of the linguistic knowledge is not meant for computers to use. Another feature of this book is that it is not just about studying the language for human understanding per se, as much of the traditional linguistic literature does; it is also about practical applications of mining sentiment and opinion expressed in natural language, for which we not only want to recognize sentiment or opinion expressions and their polarities (or orientations) but also to

extract several other pieces of important information associated with sentiment or opinion. For example, we want to identify the real-world entities or topics that a sentiment or opinion is about. These entities or topics are called *opinion* (or *sentiment*) *targets*. Extracting opinion targets is extremely important in practice. For example, in the sentence "*I am disgusted by tax increase for the poor*," if we only find that the sentence expresses a negative sentiment and/or an emotion of *disgust* from the sentence author, it is not that useful in practice. But if we also find that the negative sentiment is toward '*tax increase for the poor*,' which is the target of the negative sentiment or emotion, the information becomes much more valuable. I hope this book can serve to encourage linguists to develop a comprehensive theory about sentiment and opinion and their associated concepts.

I write this book as an introductory text to the field of sentiment analysis and as a research survey. In many places, it is one or the other, and in some other places, it is a mixture of both. The reason for this mixed or somewhat unusual presentational style is that there are few mature techniques or algorithms for sentiment analysis, although numerous researchers have attempted to solve each subproblem. In many cases, we can see from the accuracy of the results of the published papers that they are not yet ready for prime time. Another reason for the mixed presentational style of this book is that most existing research methods are direct applications of machine learning and data mining algorithms employing text features. Because many books on machine learning and data mining cover these algorithms extensively, these algorithms are thus not detailed in this book. This book also does not detail the basics of linguistics or natural language processing, such as part-of-speech tagging, syntactic parsing, shallow parsing, and grammar. Although these topics are very important to sentiment analysis too, again, they have been covered in numerous books on natural language processing. This book thus assumes that readers know the basics of machine leaning and natural language processing.

I tried to cover all major developments of the field in this book. It is thus quite comprehensive. Evidence of this is that the book cites more than six hundred publications from all major conferences and journals. I organize the book as follows. Chapter 1 introduces the book and gives the motivations for the study of sentiment analysis. We see that sentiment analysis is a fascinating and yet challenging problem with almost unlimited practical applications. Chapter 2 defines the sentiment analysis problem and discusses many of its related issues. Here we see that although sentiment analysis is a natural language processing problem, it can be defined structurally. Through the definition, we can transform unstructured text to structured data. This facilitates subsequent qualitative and quantitative analyses, which are critical for real-life applications. We also see that sentiment analysis is a multifaceted problem with many challenging and interrelated subproblems.

Chapter 3 studies the topic of document-level sentiment classification, which classifies an opinion document (e.g., a product review) as expressing a positive or negative sentiment. Chapter 4 studies the same classification problem but focuses on each individual sentence. Related problems of sentiment rating

prediction, transfer learning, and multilingual sentiment classification are also discussed in these two chapters.

Chapters 5 and 6 go to the fine-grained level to study the most important topic of aspect-based sentiment analysis, which not only classifies sentiment but also identifies the target of sentiment or opinion. Most practical sentiment analysis or opinion mining systems in industry are based on this fine-grained level of analysis. Chapter 5 focuses on aspect sentiment classification, and Chapter 6 focuses on aspect or target extraction.

Chapter 7 describes research that compiles sentiment lexicons. A sentiment lexicon is a list of words and phrases (e.g., *good*, *amazing*, *bad*, *horrible*) that people often use to express positive or negative opinions. Chapter 8 studies opinions expressed in comparative sentences. Chapter 9 focuses on opinion summarization and opinion search. Chapter 10 looks into a different type of sentiment (agreement and disagreement) expressed in online debates and discussions, which involve extensive interactive exchanges among participants. Chapter 11 investigates intention mining, which aims to discover intentions expressed in language.

Chapter 12 switches to a very different topic: detecting fake or deceptive online opinions. Chapter 13 studies the problem of ranking online reviews based on their usefulness so that users can view the most useful reviews first. Chapter 14 concludes the book and discusses some future research.

The book is suitable for students, researchers, and practitioners who are interested in social media analysis and natural language processing in general and sentiment analysis or opinion mining in particular. It is written not only for the computer science audience but also for researchers and practitioners in management sciences and social sciences. Consumer sentiments and public opinions are central to many management and social science areas such as marketing, economics, communication, and political science. Lecturers can readily use the book in class for courses on natural language processing, social media analysis, social computing, and text and data mining. Lecture slides are available online.

# Acknowledgments

Many researchers have assisted me technically in writing this book. Without their help, the book might never have become reality. First of all, I would like to thank my former and current students – Junsheng Chen, Zhiyuan Chen, Xiaowen Ding, Geli Fei, Murthy Ganapathibhotla, Minqing Hu, Nitin Jindal, Abhinav Kumar, Huayi Li, Arjun Mukherjee, Ramanathan Narayanan (Northwestern University), Federico Alberto Pozzi (visiting student from University of Milano-Bicocca), Guang Qiu (visiting student from Zhejiang University), Sathish Ramadoss, Jianfeng Si (visiting student from City University of Hong Kong), William Underwood, Andrea Vaccari, Vivek Venkataraman, Zhongwu Zhai (visiting student from Tsinghua University), and Lei Zhang – for contributing numerous research ideas over the years. Discussions with many researchers also helped shape the book: Shuanhu Bai, Jim Blomo, Erik Cambria, Malu G. Castellanos, Dennis Chong, Umesh Dayal, Eduard Dragut, Boi Faltings, Ronen Feldman, Christiane D. Fellbaum, Zhiqiang Gao, Alexander Gelbukh, Riddhiman Ghosh, Natalie Glance, Meichun Hsu, Joshua Huang, Minglie Huang, Jing Jiang, Birgit König, Xiao-li Li, Qian Liu, Boia Marina, Sharon Meraz, Enza Messina, Tieyun Qian, Jidong Shao, Mehrbod Sharifi, Hao Wang, Jan Wiebe, Qiang Yang, Clement Yu, Philip S. Yu, ChengXiang Zhai, Fangwei Zhang, Yuanlin Zhang, Jun Zhao, Xiaoyan Zhu, and Chengqing Zong. I am also in debt to the two anonymous reviewers; despite their busy schedules, they read the book very carefully and gave me numerous excellent suggestions. I have taken each and every one of them into consideration while improving the book. The web and the social media have made writing this book so much easier. I have found much valuable information from them that has helped in writing the book tremendously.

I would also like to express my gratitude to the National Science Foundation, Google, HP Labs, and the Microsoft Corporation for their generous support of my research over the years. The Department of Computer Science at the University of Illinois at Chicago provided computing resources and a supportive environment for this project.

On the publication side, it was a pleasure working with the helpful staff of Cambridge University Press. I thank my editor, Lauren Cowles, who asked me in August 2011 whether I was interested in writing a book about sentiment analysis for Cambridge University Press. It has been a wonderful experience working with her since. I also thank my copy editor, Holly T. Monteith, for helping me improve the presentation and my production editor, Sonika Rai, for guiding me through the final production process.

Finally, I thank my parents, brother, and sister for their constant support and encouragement. My deepest gratitude goes to my own family: Yue, Shelley, and Kate. They have helped me in so many ways. My wife has taken care of almost everything at home and put up with me and the long hours that I have spent on this book. I dedicate this book to them.

Bing Liu
Chicago, USA
July 10, 2014