

Series on Language Processing, Pattern Recognition,
and Intelligent Systems — Vol. 3

Social Media Content Analysis

Natural Language Processing and Beyond

Edited by

**Kam-Fai Wong, Wei Gao, Ruifeng Xu
& Wenjie Li**



Social Media Content Analysis

Natural Language Processing and Beyond

Social media platforms have been ubiquitously used in our daily lives and are steadily transforming the ways people communicate, socialize and conduct business. However, the growing popularity of social media adversely leads to wild spread of unreliable information. This in turn inevitably creates serious pollution problem of the global social media environment, which is harmful against humanity. For example, President Donald Trump used social media strategically to win in the 2016 USA Presidential Election. But it was found that many messages he delivered over social media were unproven, if not untrue. This problem must be prevented at all cost and as soon as possible. Thus, analysis of social media content is a pressing issue. It is a timely and important research subject worldwide. However, the short and informal nature of social media messages renders conventional content analysis, which is based on natural language processing (NLP), ineffective. This book presents the latest advances in NLP technologies for social media content analysis, especially content on microblogging platforms such as Twitter and Weibo.

This volume consists of a collection of highly relevant scientific articles published by the authors in different international conferences and journals, and is divided into three distinct parts: (I) Search and Filtering; (II) Opinion and Sentiment Analysis; and (III) Event Detection and Summarization.

Wong
Gao
Xu
Li

Social Media Content Analysis Natural Language Processing and Beyond



**Series on Language Processing, Pattern Recognition,
and Intelligent Systems — Vol. 3**

Social Media Content Analysis

Natural Language Processing and Beyond

Edited by

Kam-Fai Wong

The Chinese University of Hong Kong and
Key Laboratory of High Confidence in Software Technologies,
Ministry of Education, China

Wei Gao

Victoria University of Wellington, New Zealand

Ruifeng Xu

Harbin Institute of Technology, China

Wenjie Li

The Hong Kong Polytechnic University, Hong Kong

 **World Scientific**

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Series on Language Processing, Pattern Recognition, and Intelligent Systems — Vol. 3

SOCIAL MEDIA CONTENT ANALYSIS

Natural Language Processing and Beyond

Copyright © 2018 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 978-981-3223-60-8

Printed in Singapore

Social Media Content Analysis

Natural Language Processing and Beyond

Series on Language Processing, Pattern Recognition, and Intelligent Systems

Editors

Ching Y. Suen

Concordia University, Canada

parmidir@enes.concordia.ca

Lu Qin

The Hong Kong Polytechnic University, Hong Kong

csluqin@comp.polyu.edu.hk

Published

Vol. 1 Digital Fonts and Reading

edited by Mary C. Dyson and Ching Y. Suen

Vol. 2 Advances in Chinese Document and Text Processing

edited by Cheng-Lin Liu and Yue Lu

Vol. 3 Social Media Content Analysis:

Natural Language Processing and Beyond

edited by Kam-Fai Wong, Wei Gao, Wenjie Li and Ruifeng Xu

Forthcoming

Vol. 4 Computational Linguistics, Speech and Image Processing for Arabic Language

edited by Neamat El Gayar and Ching Y. Suen

Preface

Due to the growing popularity of social media platforms, online services like Facebook, Twitter, Weibo, etc., are now widely used in our daily lives for applications such as content sharing in social communities, information dissemination in e-commerce, content mining for business intelligence and so on. Social media provides a wealth of unfettered natural conversations and interactions with large volumes of information in free-form nature. For this reason, information processing over content through these social media channels faces many new opportunities and challenges; and becomes one of the important research topics in computer science.

This book is unique from the existing literature [1, 2, 3]. It covers more advanced topics reflecting the latest research outcomes on techniques and algorithms developed by a team of scientists across many countries and regions including China, Hong Kong, Qatar, UK, and USA. Although the techniques presented in this book are practically language-independent, their application to English and Chinese over social medial platforms, such as Twitter and Weibo are presented. Microblogging in English and Chinese over Twitter and Weibo, respectively, is the latest trend in social media worldwide. This book, therefore, is very timely and is significant to the advancement of global digital economy. It consists of a collection of highly relevant scientific articles published by the authors and their collaborators in different international conferences and journals and is divided into three distinct parts: (I) search and filtering; (II) opinion and sentiment analysis; and (III) event detection and summarization.

In the gigantic sea of information in the world of social media, searching a piece of relevant information is worse than finding a needle in a haystack. This situation gets much worse when the sea is flooded with uncertain information such as rumors, “fake news”, etc. For example, the news about “the Pope sponsored Donald Trump” and “arm sales between the Islamic State and Hillary Clinton” wildy flew across the social media during the 2016 USA Presidential Election. Despite being unjustified and highly skept-

tical, they swamped the public media and badly influenced the Election. How can one effectively **search and filter** similar information before they become main stream becomes a big challenge and forms the core subject of Part I of the book.

Normal text applications, e.g. news, mostly represent objective information (i.e. facts) using proper natural languages, e.g. English, Chinese, etc. But information exchange on social media is achieved differently. For example, microblogging in Twitter or Weibo represents information using non-formal text mixed with emoticons and hashtags, and purpose-wise this channel is widely used by netizens to share their opinions or sentiments which are highly subjective information. How to differentiate between subjective (fact) and objective (opinion) information? How to identify the association between a sentiment word and its target object? How to determine the sense and degree of a sentiment? etc. These are typical challenges in **opinion and sentiment analysis** addressed in Part II.

Part III looks into techniques for **extracting** important messages from microblogging sessions (i.e. re-post trees) and selecting the relevant ones for **summarization**. The non-formal nature of individual microblog messages and weak contextual information between them render existing summarization methods ineffective. These form the challenges addressed in this part of the book.

References

- [1] Huang, He-Yan, Jun, Lu, Zhang, Hua-Ping (eds.). *Social Media Processing (Proceedings of 3rd National Conference, SMP2014)*, Beijing, China, Nov. 1–2, 2014). Communications in Computer and Information Sciences, Springer-Verlag, 2014.
- [2] Zhang, X., Sun, M., Wang, Z., Hunag, X. (eds.). *Social Media Processing (Proceedings of 4th National Conference, SMP2015)*, Guangzhou, China, Nov. 16–17, 2015). Communications in Computer and Information Sciences, Springer-Verlag, 2015.
- [3] Ateh Farzindar and Diana Inkpen. *Natural Language Processing for Social Media*. Morgan & Claypool, August 2015.

About the Editors



K. F. Wong obtained his Ph.D. from Edinburgh University, Scotland, in 1987. After his Ph.D., he was researcher in Heriot-Watt University (Scotland), UniSys (Scotland) and ECRC (Germany). At present he is Associate Dean (External Affairs) of the Faculty of Engineering, Professor in the Department of Systems Engineering and Engineering Management, Director, Centre for Innovation and Technology (CINTEC), and Associate Director, Centre for Entrepreneurship, The Chinese University of Hong Kong (CUHK). He is also Honorary Professor, Harbin Institute of Technology (Shenzhen Graduate School), Adjunct Professor, School of Computer Technology, Peking University and Adjunct Professor, Northeastern University, Shenyang China. His research interest focuses on Chinese computing, database and information retrieval. He has published over 250 technical papers in these areas in different international journals and conferences and books. He is a senior member of IEEE, fellows of BCS (UK), IEE (UK) and HKIE. He is the founding Editor-In-Chief of ACM Transactions on Asian Language Processing (TALIP), and International Journal on Computational Linguistics and Chinese Language Processing. He is the General Chair of BigComp2016, NLPCC2015, IJCNLP2011, AIRS2008 and ICCPOL2006; the Finance Chair SIGMOD2007; and the PC Co-chair of IJCNLP2006. Also he has served as PC member of many international conferences.



Wei Gao is currently a Senior Lecturer of Information Systems in the School of Information Management, Victoria University of Wellington in New Zealand. Previously, he was a Scientist in Qatar Computing Research Institute, Hamad Bin Khalifa University in Qatar in 2011–2017, a Research Assistant Professor in the Chinese University of Hong Kong in 2010–2011, and a Research Fellow in the Institute for Infocomm Research in Singapore in 2010. His research interests lie in information retrieval, natural language processing, social media analytics, and artificial intelligence. He published over 60 papers and articles in the major international conferences and journals including ACL, EMNLP, SIGIR, CIKM, WSDM, IJCAI, ACM TOIS, ACM TIST, etc. He served in the program committees of a good number of international conferences and worked

as the reviewer for top-tier journals in the relevant research areas. He also served as the workshop co-chair of BigComp 2016, the session chair in ASONAM 2015, the area co-chair in NLPCC2015, and the tutorial co-chair in IJCNLP 2011. He received his Ph.D. and M.Phil. degrees of Information Systems from The Chinese University of Hong Kong.



Wenjie Li is currently an associate professor of the Department of Computing at The Hong Kong Polytechnic University. She received the B.Sc. and M.Sc. degrees from Tianjin University, China, and the Ph.D. degree from the Department of Systems Engineering and Engineering Management at the Chinese University of Hong Kong, Hong Kong. Her research interests include natural language processing, text mining, social media analysis, information retrieval, extraction and summarization. She has directed and participated in quite a number of research projects and published over 200 papers in major international journals and conference proceedings, including IEEE TKDE, IEEE TNNLS, ACM TOIS, ACM TALIP, CL, AAAI, ACL, WWW, SIGIR and CIKM. She has served as the information officer of SIGHAN and the associate editor of IJCPOL.



Ruifeng Xu received his B.Eng. degree in computer science from Harbin Institute of Technology, China in 1995, and M. Phil. and Ph.D. degrees in computer science from The Hong Kong Polytechnic University in 2002 and 2006, respectively. He worked in The Chinese University of Hong Kong, The Hong Kong Polytechnic University and City University of Hong Kong from 2006 to 2009, as Postdoctoral Fellow, Research Associate and Research Fellow, respectively. In 2009, he joined the Harbin Institute of Technology, Shenzhen Graduate School as Associate Professor and later promoted to Full Professor and Ph.D. Advisor. Prof. Xu serves as the Associate Editor of International Journal of Machine Learning and Cybernetics (Springer), Secretary General of China Association of Artificial Intelligence (CAAI) committee of Youth Work and Deputy Secretary General of China Computer Federation (CCF) committee of Chinese Information Technology. His research areas are in natural language processing, emotion computing, text mining, Bioinformatics and human-machine interface. In these areas, he has published more than 100 academic papers in international journals such as IEEE Computational Intelligence Magazine, Knowledge-based System, Scientific Reports, Bioinformatics, BMC bioinformatics and conferences such as IJCAI, ACL, EMNLP and CIKM. He is the PIs of more than 20 projects funded by National Natural Science Foundation of China, National 863 Program of China and others.

Contents

Preface	v
---------	---

About the Editors	xxi
-------------------	-----

Part I: Search and Filtering

Chapter 1. Ranking Model Selection and Fusion for Effective Microblog Search	3
---	---

Z. Wei, W. Gao, T. El-Ganainy, W. Magdy and K.-F. Wong

1. Introduction	3
2. Related Work	5
3. Ranker Selection and Fusion	5
3.1. Ranking model selection	7
3.2. Rank aggregation	8
3.3. Base rankers	9
3.4. Candidate rankers	10
3.4.1. Feature design	11
4. Evaluation	12
4.1. Experimental setting	12
4.2. Results and discussions	13
5. Conclusion and Future Work	15

Chapter 2. Microblog Search and Filtering with Real-Time Dynamics Based on BM25	19
--	----

W. Gao, Z. Wei and K.-F. Wong

1. Introduction	19
2. Systems Overview	20
2.1. Real-time ad hoc search system	20
2.2. Real-time filtering system	21

3. Dataset and Statistics	22
4. Technical Details	23
4.1. Peak-find-based blind feedback	23
4.2. Merge of different result lists	24
4.3. Greedy algorithm for online filtering	26
5. Experiments and Results	27
5.1. Topics and statistics	27
5.2. Results	28
5.2.1. Result of real-time ad hoc task	28
5.2.2. Result of real-time filtering task	29
6. Conclusions	30

Chapter 3. Exploring Tweets Normalization and Query Time

Sensitivity for Twitter Search	31
--	----

Z. Wei, W. Gao, L. Zhou, B. Li and K.-F. Wong

1. Introduction	31
2. Our Framework	32
3. Dataset and Preprocessing	34
4. Tweets Normalization	35
4.1. OOV word detection	35
4.2. Slang word translation	36
4.3. Candidate set generation	36
4.4. Candidate selection	36
5. Time-Sensitive Query Detection	38
6. Experiments and Results	38
6.1. Test dataset and setup	39
6.2. Results and discussions	39
6.2.1. Normalization	40
6.2.2. Temporal property of queries	41
6.2.3. Number of tweets returned	42
6.2.4. Performance on individual queries	43
7. Conclusion and Future Work	44

Chapter 4. A Hierarchical Knowledge Representation for Expert Finding on Social Media

.	45
-----------	----

Y. Li, W. Li and S. Li

1. Introduction	45
2. Knowledge Representation with Hierarchical Tree	47

3. Expert Finding with Approximate Tree Matching	49
4. Experiments	50
5. Conclusion	52
 Chapter 5. Twitter Hyperlink Recommendation with User-Tweet-Hyperlink Three-Way Clustering	 55
<i>D. Gao, R. Zhang, W. Li and Y. Hou</i>	
1. Introduction	56
2. Hyperlink Analysis	58
3. Tensor-based Hyperlink Recommendation	59
3.1. Collaborative tensor construction	59
3.2. Tensor decomposition-based clustering	60
3.2.1. Tensor decomposition and HOSVD	60
3.2.2. Tensor-based spectrum clustering	61
3.3. Tensor decomposition-based clustering	62
3.4. Tensor-based recommendation	63
4. Experiments	64
5. Conclusion and Future Work	65
 Chapter 6. Detect Rumors Using Time Series of Social Context Information on Microblogging	 67
<i>J. Ma, W. Gao, Z. Wei, Y. Liu and K.-F. Wong</i>	
1. Introduction	68
2. Time Series of Microblog Event	69
2.1. Time stamps generation	70
2.2. Dynamic series-time structure (DSTS)	70
3. Feature Engineering	71
4. Experimental Evaluation	73
4.1. Datasets and setup	73
4.2. Experimental results	74
4.3. Rumor early detection	76
5. Conclusion and Future Work	76
 Chapter 7. An Empirical Study on Uncertainty Identification in Social Media Context	 79
<i>Z. Wei, J. Chen, W. Gao, B. Li, L. Zhou, Y. He and K.-F. Wong</i>	
1. Introduction	79
2. Related Work	81

2.1. Uncertainty corpus	81
2.2. Uncertainty identification	81
3. Uncertainty Corpus for Microblogs	82
3.1. Types of uncertainty in microblogs	82
3.2. Annotation result	83
4. Experiment and Evaluation	84
4.1. Overall performance	85
4.2. Error analysis	86
5. Conclusion and Future Work	86

Chapter 8. Detecting Semantic Uncertainty by Learning Hedge Cues in Sentences Using an HMM 89

X. Li, W. Gao and J. W. Shavlik

1. Introduction	90
2. Related Work	91
3. Our Approach	93
3.1. Dataset	93
3.2. Design	93
3.2.1. Model definition	96
3.2.2. Parameter estimation	96
4. Experiments and Analysis	97
4.1. Results	98
4.1.1. Baseline results using Naive Bayes classifier	98
4.1.2. Results of HMM-based sentence-level classification	100
4.1.3. Results of HMM-based cue annotation	102
5. Discussion	103
6. Conclusion and Future Work	105

Part II: Opinion and Sentiment Analysis

Chapter 9. A Unified Graph Model for Sentence-Based Opinion Retrieval 111

B. Li, L. Zhou, S. Feng and K.-F. Wong

1. Introduction	111
2. Motivation	113
2.1. Formal description of problem	113
2.2. Motivation of our approach	113

3. Graph-Based Model	115
3.1. Basic idea	115
3.2. HITS model	117
4. Experiment	119
4.1. Experiment setup	119
4.1.1. Benchmark datasets	119
4.1.2. Sentiment lexicon	120
4.1.3. Topic term collection	120
4.2. Performance evaluation	121
4.2.1. Parameter tuning	121
4.2.2. Opinion retrieval model comparison	121
5. Related Work	124
5.1. Lexicon-based opinion identification	124
5.2. Unified opinion retrieval model	125
6. Conclusion and Future Work	126

Chapter 10. Intersubjectivity and Sentiment: From Language to Knowledge 129

L. Gui, R. Xu, Y. He, Q. Lu and Z. Wei

1. Introduction	129
2. Related Works	131
3. Our Approach	132
3.1. Intersubjective network	132
3.2. Network embedding and author representation learning	135
3.3. Incorporating author representations into CNN for sentiment classification	136
4. Evaluations and Discussions	138
4.1. Experimental setup	138
4.2. Comparison to other methods	138
4.3. Further analysis on author modeling	139
5. Conclusion	142

Chapter 11. Event-Driven Emotion Cause Extraction with Corpus Construction 145

L. Gui, R. Xu, D. Wu, Q. Lu and Y. Zhou

1. Introduction	145
2. Related Works	147

3. Construction of Corpus	148
3.1. Linguistic phenomenon of emotion causes	148
3.2. Collection and annotation	149
3.3. Details of dataset and its annotations	150
4. Event-Driven Emotion Cause Extraction	151
4.1. Event tree construction	152
4.2. Emotion cause extraction	153
5. Performance Evaluations	155
5.1. Experimental setup	155
5.2. Emotion cause extraction	155
Chapter 12. Learning Task Specific Distributed Paragraph Representations Using a 2-tier Convolutional Neural Network	161
<i>T. Chen, R. Xu, Y. He and X. Wang</i>	
1. Introduction	161
2. Our Approach	163
2.1. Distributed word representation model	164
2.2. Distributed sentence representation tier	164
2.3. Distributed paragraph representation tier	166
3. Evaluation and Discussion	166
3.1. Experiment settings	166
3.2. DBpedia ontology classification	167
3.3. Amazon review sentiment analysis	168
3.4. Discussion	168
4. Conclusion and Future Directions	169
Chapter 13. Build Emotion Lexicon from Microblogs by Combining Effects of Seed Words and Emoticons in a Heterogeneous Graph	171
<i>K. Song, S. Feng, W. Gao, D. Wang, L. Chen and C. Zhang</i>	
1. Introduction	172
2. Related Work	173
3. Emotion Distribution of Emoticons	175
3.1. Building emoticon dataset	175
3.2. Inferring emotion distribution	176
4. Seed Emotion Words	177
5. Building Emotion Lexicon	177
5.1. Symbols and notations	178
5.2. Building heterogeneous graph	178