Natural Language Processing for Social Media

Second Edition

Synthesis Lectures on Human Language Technologies

Metaphor: A Computational Perspective
Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov
2016

Grammatical Inference for Computational Linguistics
Jeffrey Heinz, Colin de la Higuera, and Menno van Zaanen
2015

Automatic Detection of Verbal Deception
Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari
2015

Natural Language Processing for Social Media
Atefeh Farzindar and Diana Inkpen
2015

Semantic Similarity from Natural Language and Ontology Analysis
Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain
2015

Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition
Hang Li
2014

Ontology-Based Interpretation of Natural Language
Philipp Cimiano, Christina Unger, and John McCrae
2014

Automated Grammatical Error Detection for Language Learners, Second Edition
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2014

Web Corpus Construction
Roland Schäfer and Felix Bildhauer
2013

Recognizing Textual Entailment: Models and Applications
Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto
2013

Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
Emily M. Bender
2013

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing
Anders Søgaard
2013

Semantic Relations Between Nominals
Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz
2013

Computational Modeling of Narrative
Inderjeet Mani
2012

Natural Language Processing for Historical Texts
Michael Piotrowski
2012

Sentiment Analysis and Opinion Mining
Bing Liu
2012

Discourse Processing
Manfred Stede
2011

Bitext Alignment
Jörg Tiedemann
2011

Linguistic Structure Prediction
Noah A. Smith
2011

Learning to Rank for Information Retrieval and Natural Language Processing
Hang Li
2011

Computational Modeling of Human Language Acquisition
Afra Alishahi
2010

Introduction to Arabic Natural Language Processing
Nizar Y. Habash
2010

Cross-Language Information Retrieval
Jian-Yun Nie
2010

Automated Grammatical Error Detection for Language Learners
Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault
2010

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer
2010

Semantic Role Labeling
Martha Palmer, Daniel Gildea, and Nianwen Xue
2010

Spoken Dialogue Systems
Kristiina Jokinen and Michael McTear
2009

Introduction to Chinese Natural Language Processing
Kam-Fai Wong, Wenjie Li, Ruifeng Xu, and Zheng-sheng Zhang
2009

Introduction to Linguistic Annotation and Text Analytics
Graham Wilcock
2009

Dependency Parsing
Sandra Kübler, Ryan McDonald, and Joakim Nivre
2009

Statistical Language Models for Information Retrieval
ChengXiang Zhai
2008

Natural Language Processing for Social Media, Second Edition
Atefeh Farzindar and Diana Inkpen

ISBN: 978-3-031-01039-2 paperback
ISBN: 978-3-031-02167-1 ebook
ISBN: 978-3-031-00178-9 hardcover

DOI 10.1007/978-3-031-02167-1

A Publication in the Springer series
SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES

Lecture #38
Series Editor: Graeme Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Natural Language Processing for Social Media

Second Edition

Atefeh Farzindar
University of Southern California

Diana Inkpen
University of Ottawa

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #38

ABSTRACT TO THE SECOND EDITION

In recent years, online social networking has revolutionized interpersonal communication. The newer research on language analysis in social media has been increasingly focusing on the latter's impact on our daily lives, both on a personal and a professional level. Natural language processing (NLP) is one of the most promising avenues for social media data processing. It is a scientific challenge to develop powerful methods and algorithms which extract relevant information from a large volume of data coming from multiple sources and languages in various formats or in free form. We discuss the challenges in analyzing social media texts in contrast with traditional documents.

Research methods in information extraction, automatic categorization and clustering, automatic summarization and indexing, and statistical machine translation need to be adapted to a new kind of data. This book reviews the current research on NLP tools and methods for processing the non-traditional information from social media data that is available in large amounts (big data), and shows how innovative NLP approaches can integrate appropriate linguistic information in various fields such as social media monitoring, healthcare, business intelligence, industry, marketing, and security and defence.

We review the existing evaluation metrics for NLP and social media applications, and the new efforts in evaluation campaigns or shared tasks on new datasets collected from social media. Such tasks are organized by the Association for Computational Linguistics (such as SemEval tasks) or by the National Institute of Standards and Technology via the Text REtrieval Conference (TREC) and the Text Analysis Conference (TAC). In the concluding chapter, we discuss the importance of this dynamic discipline and its great potential for NLP in the coming decade, in the context of changes in mobile technology, cloud computing, virtual reality, and social networking.

In this second edition, we have added information about recent progress in the tasks and applications presented in the first edition. We discuss new methods and their results. The number of research projects and publications that use social media data is constantly increasing due to continuously growing amounts of social media data and the need to automatically process them. We have added 85 new references to the more than 300 references from the first edition. Besides updating each section, we have added a new application (digital marketing) to the section on media monitoring and we have augmented the section on healthcare applications with an extended discussion of recent research on detecting signs of mental illness from social media.

KEYWORDS

social media, social networking, natural language processing, social computing, big data, semantic analysis

This effort is dedicated to my husband, Massoud, and to my daughters, Tina and Amanda, who are just about the best children a mom could hope for: happy, loving, and fun to be with.

– Atefeh Farzindar

To my wonderful husband, Nicu, with whom I can climb any mountain, and to our sweet baby daughter Nicoleta.

– Diana Inkpen

Contents

Preface to the Second Edition

This book presents the state-of-the-art in research and empirical studies in the field of Natural Language Processing (NLP) for the semantic analysis of social media data. Because the field is continuously growing, this second edition adds information about recently proposed methods and their results for the tasks and applications that we covered in the first edition.

Over the past few years, online social networking sites have revolutionized the way we communicate with individuals, groups, and communities, and altered everyday practices. The unprecedented volume and variety of user-generated content and the user interaction network constitute new opportunities for understanding social behavior and building socially intelligent systems.

Much of the research on social networks and the mining of the social web is based on graph theory. That is apt because a social structure is made up of a set of social actors and a set of the dyadic ties between these actors. We believe that the graph mining methods for structure and information diffusion or influence spread in social networks need to be combined with the content analysis of social media. This provides the opportunity for new applications that use the information publicly available as a result of social interactions. Adapted classic NLP methods can partially solve the problem of social media content analysis focusing on the posted messages. When we receive a text of less than 10 characters, including an emoticon and a heart, we understand it and even respond to it! It is impossible to use NLP methods to process this type of document, but there is a logical message in social media data based on which two people can communicate. The same logic dominates worldwide, and people from all over the world use it to share and communicate with each other. There is a new and challenging language for NLP.

We believe that we need new theories and algorithms for semantic analysis of social media data, as well as a new way of approaching the big data processing. By semantic analysis, in this book, we mean the linguistic processing of the social media messages enhanced with semantics, and possibly also combining this with the structure of the social networks. We actually use the term in a more general sense to refer to applications that do intelligent processing of social media texts and meta-data. Some applications could access very large amounts of data; therefore, the algorithms need to be adapted to be able process data (big data) in an online fashion and without necessarily storing all the data.

This motivated us to give two tutorials: *Applications of Social Media Text Analysis* at EMNLP 2015[1] and *Natural Language Processing for Social Media* at the 29th Canadian Confer-

[1] http://www.emnlp2015.org/tutorials/3/3_OptionalAttachment.pdf
https://www.cs.cmu.edu/~ark/EMNLP-2015/proceedings/EMNLP-Tutorials/pdf/EMNLP-Tutorials06.pdf

ence on Artificial Intelligence (AI 2016).[2] We also organized several workshops on this topic, Semantic Analysis in Social Networks (SASM 2012)[3] and Language Analysis in Social Media (LASM 2013[4] and LASM 2014[5]), in conjunction with conferences organized by the Association for Computational Linguistics[6] (ACL, EACL, and NAACL-HLT).

Our goal was to reflect a wide range of research and results in the analysis of language with implications for fields such as NLP, computational linguistics, sociolinguistics, and psycholinguistics. Our workshops invited original research on all topics related to the analysis of language in social media, including the following topics.

- What do people talk about on social media?
- How do they express themselves?
- Why do they post on social media?
- How do language and social network properties interact?
- Natural language processing techniques for social media analysis.
- Semantic Web/ontologies/domain models to aid in understanding social data.
- Characterizing participants via linguistic analysis.
- Language, social media, and human behavior.

There were several other workshops on similar topics, for example, the *Making Sense of Microposts* (#Microposts)[7] workshop series in conjunction with the World Wide Web Conference 2012–2016. These workshops focused in particular on short informal texts that are published without much effort (such as tweets, Facebook shares, Instagram-like shares, and Google+ messages). There has been another series of workshops on Natural Language Processing for Social Media (SocialNLP) since 2013, with SocialNLP 2017 offered in conjunction with EACL 2017[8] and IEEE BigData 2017.[9]

The **intended audience** of this book is researchers who are interested in developing tools and applications for automatic analysis of social media texts. We assume that the readers have basic knowledge in the area of natural language processing and machine learning. We hope that this book will help the readers better understand computational linguistics and social media

[2] http://aigicrv.org/2016/
[3] https://aclweb.org/anthology/W/W12/#2100
[4] https://aclweb.org/anthology/W/W13/#1100
[5] https://aclweb.org/anthology/W/W14/#1300
[6] http://www.aclweb.org/
[7] http://microposts2016.seas.upenn.edu/
[8] http://eacl2017.org/
[9] http://cci.drexel.edu/bigdata/bigdata2017/

analysis, in particular text mining techniques and NLP applications (such as summarization, localization detection, sentiment and emotion analysis, topic detection, and machine translation) designed specifically for social media texts.

Atefeh Farzindar and Diana Inkpen
December 2017

Acknowledgments

This book would not have been possible without the hard work of many people. We would like to thank our colleagues at NLP Technologies Inc., the NLP research group at the University of Ottawa, and our students James Webb and Ruining Liu from the University of Southern California. We would like to thank in particular Prof. Stan Szpakowicz from the University of Ottawa for his comments on the draft of the book, and two anonymous reviewers for their useful suggestions for revisions and additions. We thank Prof. Graeme Hirst at the University of Toronto and Michael Morgan from Morgan & Claypool Publishers for their continuous encouragement.

Atefeh Farzindar and Diana Inkpen
December 2017