

Natural Language Processing for Social Media

Third Edition

Synthesis Lectures on Human Language Technologies

Editor

Graeme Hirst, *University of Toronto*

Synthesis Lectures on Human Language Technologies is edited by Graeme Hirst of the University of Toronto. The series consists of 50- to 150-page monographs on topics relating to natural language processing, computational linguistics, information retrieval, and spoken language understanding. Emphasis is on important new techniques, on new applications, and on topics that combine two or more HLT subfields.

Natural Language Processing for Social Media, Third Edition

Anna Atefeh Farzindar and Diana Inkpen
2020

Statistical Significance Testing for Natural Language Processing

Rotem Dror, Lotem Peled, Segev Shlomov, and Roi Reichart
2020

Deep Learning Approaches to Text Production

Shashi Narayan and Claire Gardent
2020

Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics

Emily M. Bender and Alex Lascarides
2019

Cross-Lingual Word Embeddings

Anders Søgaard, Ivan Vulić, Sebastian Ruder, Manaal Faruqui
2019

Bayesian Analysis in Natural Language Processing, Second Edition

Shay Cohen
2019

Argumentation Mining

Manfred Stede and Jodi Schneider
2018

Quality Estimation for Machine Translation

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold
2018

Natural Language Processing for Social Media, Second Edition

Atefeh Farzindar and Diana Inkpen
2017

Automatic Text Simplification

Horacio Saggion
2017

Neural Network Methods for Natural Language Processing

Yoav Goldberg
2017

Syntax-based Statistical Machine Translation

Philip Williams, Rico Sennrich, Matt Post, and Philipp Koehn
2016

Domain-Sensitive Temporal Tagging

Jannik Strötgen and Michael Gertz
2016

Linked Lexical Knowledge Bases: Foundations and Applications

Iryna Gurevych, Judith Eckle-Kohler, and Michael Matuschek
2016

Bayesian Analysis in Natural Language Processing

Shay Cohen
2016

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

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2020

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

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

ISBN: 978-3-031-01047-7	paperback
ISBN: 978-3-031-02175-6	ebook
ISBN: 978-3-031-00186-4	hardcover

DOI 10.1007/978-3-031-02175-6

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

Lecture #46
Series Editor: Grame Hirst, *University of Toronto*
Series ISSN
Print 1947-4040 Electronic 1947-4059

Cover art illustration by Anna Atefeh Farzindar.

Natural Language Processing for Social Media

Third Edition

Anna Atefeh Farzindar
University of Southern California

Diana Inkpen
University of Ottawa

SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES #46

ABSTRACT

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 that extract relevant information from a large volume of data coming from multiple sources and languages in various formats or in free form. This book will 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, and it shows how innovative NLP approaches can integrate appropriate linguistic information in various fields such as social media monitoring, health care, and business intelligence. The book further covers 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), the National Institute of Standards and Technology via the Text REtrieval Conference (TREC) and the Text Analysis Conference (TAC), or the Conference and Labs of the Evaluation Forum (CLEF).

In this third edition of the book, the authors added information about recent progress in NLP for social media applications, including more about the modern techniques provided by deep neural networks (DNNs) for modeling language and analyzing social media data.

KEYWORDS

social media, social networking, natural language processing, social computing, big data, semantic analysis, artificial intelligence, deep learning

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

– Anna Atefeh Farzindar

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

– Diana Inkpen

Contents

List of Figures	xvii
List of Tables	xix
Preface	xxi
Acknowledgments	xxv
1 Introduction to Social Media Analysis	1
1.1 Introduction	1
1.2 Social Media Applications	6
1.2.1 Cross-language Document Analysis in Social Media Data	7
1.2.2 Deep Learning techniques for Social Media Data	7
1.2.3 Real-world Applications	8
1.3 Challenges in Social Media Data	9
1.4 Semantic Analysis of Social Media	12
1.5 Summary	13
2 Linguistic Pre-processing of Social Media Texts	15
2.1 Introduction	15
2.2 Generic Adaptation Techniques for NLP Tools	17
2.2.1 Text Normalization	18
2.2.2 Re-training NLP Tools for Social Media Texts	20
2.3 Tokenizers	21
2.4 Part-of-speech Taggers	22
2.5 Chunkers and Parsers	25
2.6 Named Entity Recognizers	29
2.7 Existing NLP Toolkits for English and Their Adaptation	31
2.8 Multi-linguality and Adaptation to Social Media Texts	32
2.8.1 Language Identification	32
2.8.2 Dialect Identification	35
2.9 Summary	41

3	Semantic Analysis of Social Media Texts	43
3.1	Introduction	43
3.2	Geo-location Detection	43
3.2.1	Mapping Social Media Information on Maps	44
3.2.2	Readily Available Geo-location Information	44
3.2.3	Geo-location based on Network Infrastructure	44
3.2.4	Geo-location based on the Social Network Structure	45
3.2.5	Content-based Location Detection	45
3.2.6	Evaluation Measures for Geo-location Detection	49
3.3	Entity Linking and Disambiguation	51
3.3.1	Detecting Entities and Linked Data	52
3.3.2	Evaluation Measures for Entity Linking	55
3.4	Opinion Mining and Emotion Analysis	55
3.4.1	Sentiment Analysis	55
3.4.2	Emotion Analysis	58
3.4.3	Sarcasm Detection	60
3.4.4	Evaluation Measures for Opinion and Emotion Classification	61
3.5	Event and Topic Detection	62
3.5.1	Specified vs. Unspecified Event Detection	62
3.5.2	New vs. Retrospective Events	69
3.5.3	Emergency Situation Awareness	70
3.5.4	Evaluation Measures for Event Detection	71
3.6	Automatic Summarization	71
3.6.1	Update Summarization	73
3.6.2	Network Activity Summarization	73
3.6.3	Event Summarization	74
3.6.4	Opinion Summarization	75
3.6.5	Keyphrase Generation	76
3.6.6	Evaluation Measures for Summarization	76
3.7	Machine Translation	77
3.7.1	Neural Machine Translation	78
3.7.2	Adapting Phrase-based Machine Translation to Normalize Medical Terms	79
3.7.3	Translating Government Agencies' Tweet Feeds	79
3.7.4	Hashtag Occurrence, Layout, and Translation	81
3.7.5	Machine Translation for Arabic Social Media	84
3.7.6	Evaluation Measures for Machine Translation	85

3.8	Summary	86
4	Applications of Social Media Text Analysis	87
4.1	Introduction	87
4.2	Healthcare Applications	87
4.3	Financial Applications	96
4.4	Predicting Voting Intentions	99
4.5	Media Monitoring	101
4.6	Security and Defense Applications	104
4.7	Disaster Response Applications	107
4.8	NLP-based User Modeling	109
4.9	Applications for Entertainment	115
4.10	NLP-based Information Visualization for Social Media	117
4.11	Government Communication	117
4.12	Rumor Detection	118
4.13	Recommender systems	119
4.14	Preventing Sexual Harassment	120
4.15	Summary	120
5	Data Collection, Annotation, and Evaluation	121
5.1	Introduction	121
5.2	Discussion on Data Collection and Annotation	121
5.3	Spam and Noise Detection	122
5.4	Privacy and Democracy in Social Media	125
5.5	Evaluation Benchmarks	126
5.6	Summary	128
6	Conclusion and Perspectives	129
6.1	Conclusion	129
6.2	Perspectives	129
A	TRANSLI: a Case Study for Social Media Analytics and Monitoring	133
A.1	TRANSLI architecture	133
A.2	User Interface	134

Glossary	139
Bibliography	141
Authors' Biographies	191
Index	193

List of Figures

1.1	Social networks ranked by the number of active users as of January 2014 (in millions) provided by Statista.	3
1.2	Number of monthly active Facebook users from the third quarter of 2008 to the first quarter of 2014 (in millions) provided by Statista.	3
1.3	Number of LinkedIn members from the first quarter of 2009 to the first quarter of 2014 (in millions) provided by Statista.	4
1.4	A framework for semantic analysis in social media, where NLP tools transform the data into intelligence.	6
2.1	Methodology for tweet normalization. The dotted horizontal line separates the two steps (detecting the text to be normalized and applying normalization rules) [Akhtar et al., 2015].	19
2.2	Taxonomy of normalization edits [Baldwin and Li, 2015].	20
2.3	Arabic dialects distribution and variation across Asia and Africa [Sadat et al., 2014a].	36
2.4	Division of Arabic dialects in six groups/divisions [Sadat et al., 2014a].	36
2.5	Accuracies on the character-based n -gram Markov language models for 18 countries [Sadat et al., 2014a].	38
2.6	Accuracies on the character-based n -gram Markov language models for the six divisions/groups [Sadat et al., 2014a].	39
2.7	Accuracies on the character-based n -gram Naïve Bayes classifiers for 18 countries [Sadat et al., 2014a].	40
2.8	Accuracies on the character-based n -gram Naïve Bayes classifiers for the six divisions/groups [Sadat et al., 2014a].	41
3.1	Example of a pair of tweets extracted from the bilingual feed pair Health Canada/Santé Canada, after tokenization.	81
3.2	An original tweet with hashtags in its three possible regions.	82
4.1	Examples of annotated social media posts discussing ADRs [Nikfarjam et al., 2015].	89

4.2	The DeepHealthMiner neural net architecture [Nikfarjam, 2016].	90
4.3	SVM-based text mining procedure for impact management [Schniederjans et al., 2013].	99
A.1	TRANSLI Social Media Analytics and monitoring module architecture. . . .	134
A.2	TRANSLI user interface for event creation module.	135
A.3	TRANSLI user interface for event browsing module.	135
A.4	TRANSLI user interface to present an event. Components are identified with their IDs.	137

List of Tables

1.1	Social media platforms and their characteristics	2
2.1	Three examples of Twitter texts	18
2.2	Examples of tokenization	21
2.3	Penn TreeBank tagset	23
2.4	POS tagset from Gimpel et al. [2011]	26
2.5	Example of tweet parsed with the TweepoParser	28
3.1	An example of annotation with the true location [Inkpen et al., 2015]	49
3.2	Classification accuracies for user location detection on the Eisenstein dataset [Liu and Inkpen, 2015]	50
3.3	Mean error distance of predictions on the Eisenstein dataset [Liu and Inkpen, 2015]	50
3.4	Results for user location prediction on the Roller dataset [Liu and Inkpen, 2015]	50
3.5	Performance of the classifiers trained on different features for cities [Inkpen et al., 2015]	51
3.6	Classification results for emotion classes and non-emotion by Ghazi et al. [2014]	63
3.7	Accuracy of the mood classification by Keshtkar and Inkpen [2012]	63
3.8	Statistics on hashtag use in the aligned bilingual corpus [Gotti et al., 2014]	82
3.9	Distribution of hashtags in epilogues and prologues [Gotti et al., 2014]	82
3.10	Percentage of unknown hashtags to English and French vocabularies of the Hansard corpus [Gotti et al., 2014]	83
3.11	Percentage of unknown hashtags to “standard” English and French vocabularies, after automatic segmentation of multiword hashtags into simple words [Gotti et al., 2014]	83
3.12	Translation performance obtained by Gotti et al. [2014]	86

Preface

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 third edition adds information about recently proposed methods and their results for the tasks and applications that we covered in the first and second editions.

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 research work 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, 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 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 three tutorials on *Applications of Social Media Text Analysis* at EMNLP 2015¹, on *Natural Language Processing for Social Media* at the 29th Canadian Con-

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

ference on Artificial Intelligence (AI 2016)², and on *How Natural Language Processing Helps Uncover Social Media Insights* at the 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS 2020). Also on this topic, we organized several workshops (Semantic Analysis in Social Networks (SASM 2012)³, Language Analysis in Social Media (LASM 2013⁴, and LASM 2014⁵) in conjunction with conferences organized by the Association for Computational Linguistics⁶ (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 to 2016. These workshops focused in particular on short informal texts that are published without much effort (such as tweets, Facebook shares, Instagram-like shares, Google+ messages). There has been another series of Workshops on Natural Language Processing for Social Media (SocialNLP) since 2013. For example, SocialNLP 2017 was in conjunction with EACL 2017⁸ and IEEE BigData 2017⁹, and SocialNLP 2020 had two editions, one in conjunction with TheWebConf 2020 and one in conjunction with ACL 2020¹⁰.

The **intended audience** of this book is researchers that are interested in developing tools and applications for automatic analysis social of 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 analysis, in particular text mining techniques and NLP applications (such as summarization,

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://eac2017.org/>

⁹<http://cci.drexel.edu/bigdata/bigdata2017/>

¹⁰<https://sites.google.com/site/socialnlp2020/>

localization detection, sentiment and emotion analysis, topic detection and machine translation) designed specifically for social media texts.

Besides updating each section in this third edition, we added a new section on keyphrase generation from social media messages and one on neural machine translation in Chapter 3 and three new applications in Chapter 4: rumor detection, recommender systems for social media, and preventing sexual harassment. We discuss the new methods and their results. The number of research projects and publications that use social media data is constantly increasing. Finally, we added more than 50 new references to the approximately 400 references from the second edition.

Anna Atefeh Farzindar and Diana Inkpen
March 2020

Acknowledgments

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

Anna Atefeh Farzindar and Diana Inkpen
March 2020