# Natural Language Processing for Social Media

## Second 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, 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 Senrich, 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

Natural Language Processing for Social Media, Second Edition

Atefeh Farzindar and Diana Inkpen

www.morganclaypool.com

ISBN: 9781681736129 paperback
ISBN: 9781681736136 ebook
ISBN: 9781681736143 hardcover

DOI 10.2200/S00809ED2V01Y201710HLT038

A Publication in the Morgan & Claypool Publishers 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*

MORGAN & CLAYPOOL PUBLISHERS

# 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

CHAPTER 1

# Introduction to Social Media Analysis

## 1.1 INTRODUCTION

Social media is a phenomenon that has recently expanded throughout the world and quickly attracted billions of users. This form of electronic communication through social networking platforms allows users to generate its content and share it in various forms of information, personal words, pictures, audio, and videos. Therefore, social computing is formed as an emerging area of research and development that includes a wide range of topics such as Web semantics, artificial intelligence, natural language processing, network analysis, and Big Data analytics.

Over the past few years, online social networking sites (Facebook, Twitter, YouTube, Flickr, MySpace, LinkedIn, Metacafe, Vimeo, etc.) have revolutionized the way we communicate with individuals, groups, and communities, and have altered everyday practices [Boyd and Ellison, 2007].

The broad categories of social media platforms are: content-sharing sites, forums, blogs, and microblogs. On content sharing sites (such as Facebook, Instagram, Foursquare, Flickr, YouTube) people exchange information, messages, photos, videos, or other types of content. On Web user forums (such as StackOverflow, CNET forums, Apple Support) people post specialized information, questions, or answers. Blogs (such as Gizmodo, Mashable, Boing Boing, and many more) allow people to post messages and other content and to share information and opinions. Micro-blogs (such as Twitter, Sina Weibo, Tumblr) are limited to short texts for sharing information and opinions. The modalities of sharing content in order: posts; comments to posts; explicit or implicit connections to build social networks (friend connections, followers, etc.); cross-posts and user linking; social tagging; likes/favorites/starring/voting/rating/etc.; author information; and linking to user profile features.[1] In Table 1.1, we list more details about social media platforms and their characteristics and types of content shared [Barbier et al., 2013].

Social media statistics for January 2014 have shown that Facebook has grown to more than 1 billion active users, adding more than 200 million users in a single year. Statista,[2] the world's largest statistics portal, announced the ranking for social networks based on the number of active users. As presented in Figure 1.1, the ranking shows that Qzone took second place

[1] http://people.eng.unimelb.edu.au/tbaldwin/pubs/starsem2014.pdf
[2] http://www.statista.com/

Table 1.1: Social media platforms and their characteristics

| Type | Characteristics | Examples |
|---|---|---|
| Social Networks | A social networking website allows the user to build a web page and connect with a friend or other acquaintance in order to share user-generated content. | MySpace, Facebook, LinkedIn, Meetup, Google Plus+ |
| Blogs and Blog Comments | A blog is an online journal where the blogger can create the content and display it in reverse chronological order. Blogs are generally maintained by a person or a community. Blog comments are posts by users attached to blogs or online newspaper posts. | Huffington Post, Business Insider, Engadget, and online journals |
| Microblogs | A microblog is similar to a blog but has a limited content. | Twitter, Tumblr, Sina Weibo, Plurk |
| Forums | An online forum is a place for members to discuss a topic by posting messages. | Online Discussion Communities, phpBB Developer Forum, Raising Children Forum |
| Social Bookmarks | Services that allow users to save, organize, and search links to various websites, and to share their bookmarks of Web pages. | Delicious, Pinterest, Google Bookmarks |
| Wikis | These websites allow people to collaborate and add content or edit the information on a community-based database. | Wikipedia, Wikitravel, Wikihow |
| Social News | Social news encourage their community to submit news stories, or to vote on the content and share it. | Digg, Slashdot, Reddit |
| Media Sharing | A website that enables users to capture videos and pictures or upload and share with others. | YouTube, Flickr, Snapchat, Instagram, Vine |

with more than 600 million users. Google+, LinkedIn, and Twitter completed the top 5 with 300 million, 259 million, and 232 million active users, respectively.

Statista also provided the growth trend for both Facebook and LinkedIn, illustrated in Figure 1.2 and Figure 1.3, respectively. Figure 1.2 shows that Facebook, by reaching 845 million

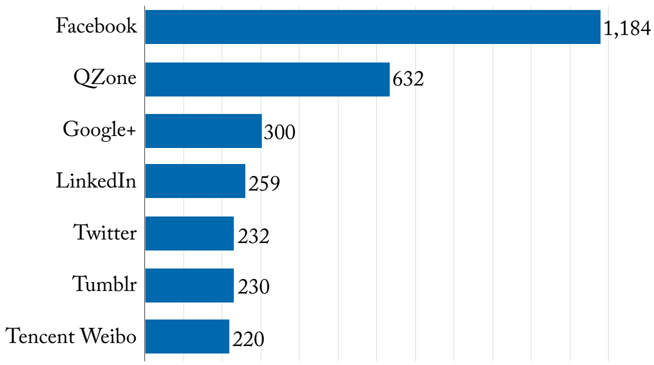


Figure 1.1: Social networks ranked by the number of active users as of January 2014 (in millions) provided by Statista.

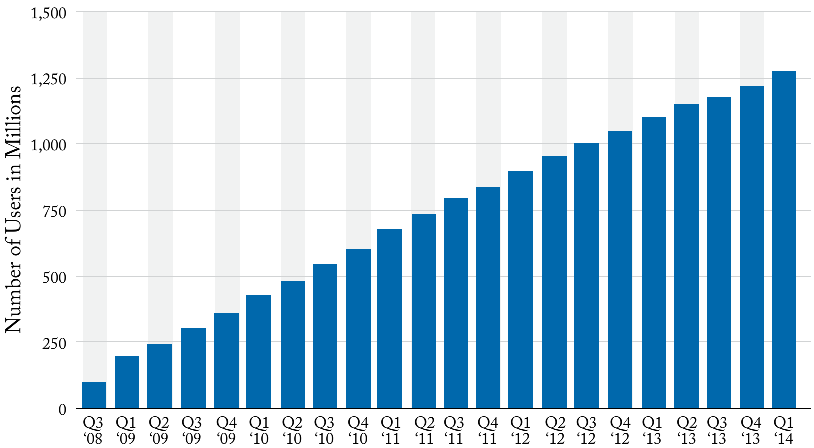


Figure 1.2: Number of monthly active Facebook users from the third quarter of 2008 to the first quarter of 2014 (in millions) provided by Statista.

users at the end of 2011, totaled 1,228 million users by the end of 2013. As depicted in Figure 1.3, LinkedIn also reached 277 million users by the end of 2013, whereas it only had 145 million users at the end of 2011. Statista also calculated the annual income for both Facebook and LinkedIn, which in 2013 totalled US$7,872 and US$1,528 million, respectively.

Social computing is an emerging field that focuses on modeling, analysis, and monitoring of social behavior on different media and platforms to produce intelligent applications. Social media is the use of electronic and Internet tools for the purpose of sharing and discussing information and experiences with other human beings in efficient ways [Moturu, 2009]. Various social media platforms such as social networks, forums, blogs, and micro-blogs have recently

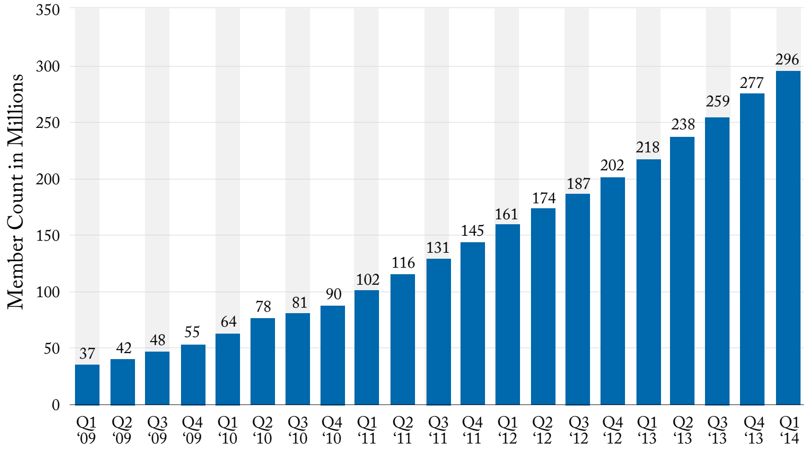


Figure 1.3: Number of LinkedIn members from the first quarter of 2009 to the first quarter of 2014 (in millions) provided by Statista.

evolved to ensure the connectivity, collaboration, and formation of virtual communities. While traditional media such as newspapers, television, and radio provide unidirectional communication from business to consumer, social media services have allowed interactions among users across various platforms. Social media have therefore become a primary source of information for business intelligence.

There are several means of interaction in social media platforms. One of the most important is via text posts. The natural language processing (NLP) of traditional media such as written news and articles has been a popular research topic over the past 25 years. NLP typically enables computers to derive meaning from natural language input using the knowledge from computer science, artificial intelligence, and linguistics.

NLP for social media text is a new research area, and it requires adapting the traditional NLP methods to these kinds of texts or developing new methods suitable for information extraction and other tasks in the context of social media.

There are many reasons why the "traditional" NLP are not good enough for social media texts, such as their informal nature, the new type of language, abbreviations, etc. Section 1.3 will discuss these aspects in more detail.

A social network is made up of a set of actors (such as individuals or organizations) and a set of binary relations between these actors (such as relationships, connections, or interactions). From a social network perspective, the goal is to model the structure of a social group to identify how this structure influences other variables and how structures change over time. Semantic analysis in social media (SASM) is the semantic processing of the text messages as well as of the meta-data, in order to build intelligent applications based on social media data.

SASM helps develop automated tools and algorithms to monitor, capture, and analyze the large amounts of data collected from social media in order to predict user behavior or extract other kinds of information. If the amount of data is very large, techniques for "big data" processing need to be used, such as online algorithms that do not need to store all the data in order to update the models based on the incoming data.

In this book, we focus on the analysis of the textual data from social media, via new NLP techniques and applications. Recently, workshops such as the EACL 2014 Workshop on Language Analysis in Social Media [Farzindar et al., 2014], the NAACL/HLT 2013 workshop on Language Analysis in Social Media [Farzindar et al., 2013], and the EACL 2012 Workshop for Semantic Analysis in Social Media [Farzindar and Inkpen, 2012] have been increasingly focusing on NLP techniques and applications that study the effect of social media messages on our daily lives, both personally and professionally.

Social media textual data is the collection of openly available texts that can be obtained publicly via blogs and micro-blogs, Internet forums, user-generated FAQs, chat, podcasts, online games, tags, ratings, and comments. Social media texts have several properties that make them different than traditional texts, because the nature of the social conversations, posted in realtime. Detecting groups of topically related conversations is important for applications, as well as detection emotions, rumors, and incentives. Determining the locations mentioned in the messages or the locations of the users can also add valuable information. The texts are unstructured and are presented in many formats and written by different people in many languages and styles. Also, the typographic errors and chat slang have become increasingly prevalent on social networking sites like Facebook and Twitter. The authors are not professional writers and their postings are spread in many places on the Web, on various social media platforms.

Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information, which would not have been available from traditional media outlets. Semantic analysis of social media has given rise to the emerging discipline of big data analytics, which draws from social network analysis, machine learning, data mining, information retrieval, and natural language processing [Melville et al., 2009].

Figure 1.4 shows a framework for semantic analysis in social media. The first step is to identify issues and opportunities for collecting data from social networks. The data can be in the form of stored textual information (the big datacould be stored in large and complex databases or text files), it could be dynamic online data collection processed in real time, or it could be retrospective data collection for particular needs. The next step is the SASM pipeline, which consists of specific NLP tools for the social media analysis and data processing. Social media data is made up of large, noisy, and unstructured datasets. SASM transforms social media data to meaningful and understandable messages through social information and knowledge. Then, SASM analyzes the social media information in order to produce social media intelligence. Social media intelligence can be shared with users or presented to decision-makers to improve awareness, communication, planning, or problem solving. The presentation of analyzed data by SASM could be completed by data visualization methods.

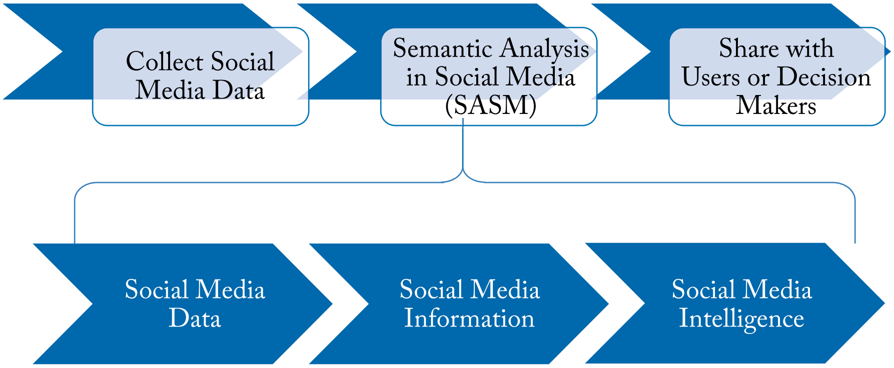


Figure 1.4: A framework for semantic analysis in social media, where NLP tools transform the data into intelligence.

## 1.2 SOCIAL MEDIA APPLICATIONS

The automatic processing of social media data needs to design appropriate research methods for applications such as information extraction,automatic categorization, clustering, indexing data for information retrieval, and statistical machine translation. The sheer volume of social media data and the incredible rate at which new content is created makes monitoring, or any other meaningful manual analysis, unfeasible. In many applications, the amount of data is too large for effective real-time human evaluation and analysis of the data for a decision maker.

Social media monitoring is one of the major applications in SASM. Traditionally, **media monitoring** is defined as the activity of monitoring and tracking the output of the hard copy, online, and broadcast media which can be performed for a variety of reasons, including political, commercial, and scientific. The huge volume of information provided via social media networks is an important source for open intelligence. Social media make the direct contact with the target public possible. Unlike traditional news, the opinion and sentiment of authors provide an additional dimension for the social media data. The different sizes of source documents—such as a combination of multiple tweets and blogs—and content variability also render the task of analyzing social media documents difficult.

In social media, the real-time event search or event detection The search queries consider multiple dimensions, including spatial and temporal. In this case, some NLP methods such as information retrieval and summarization of social data in the form of various documents from multiple sources become important in order to support the event search and the detection of relevant information.

The semantic analysis of the meaning of a day's or week's worth of conversations in social networks for a group of topically related discussions or about a specific event presents the challenges of cross-language NLP tasks. Social media—related NLP methods that can extract information of interest to the analyst for preferential inclusion also lead us to domain-based applications in computational linguistics.

### 1.2.1 CROSS-LANGUAGE DOCUMENT ANALYSIS IN SOCIAL MEDIA DATA

The application of existing NLP techniques to social media from different languages and multiple resources faces several additional challenges; the tools for text analysis are typically designed for specific languages. The main research issue therefore lies in assessing whether language-independence or language-specificity is to be preferred. Users publish content not only in English, but in a multitude of languages. This means that due to the language barrier, many users cannot access all available content. The use of machine translation technology can help bridge the language gap in such situations. The integration of machine translation and NLP tools opens opportunities for the semantic analysis of text via cross-language processing.

### 1.2.2 REAL-WORLD APPLICATIONS

The huge volume of publicly available information on social networks and on the Web can benefit different areas such as industry, media, healthcare, politics, public safety, and security. Here, we can name a few innovative integrations for social media monitoring, and some model scenarios of government-user applications in coordination and situational awareness. We will show how NLP tools can help governments interpret data in near real-time and provide enhanced command decision at the strategic and operational levels.

#### Industry

There is great interest on the part of industry in social media data monitoring. Social media data can dramatically improve business intelligence (BI). Businesses could achieve several goals by integrating social data into their corporate BI systems, such as branding and awareness, customer/prospect engagement, and improving customer service. Online marketing, stock market prediction, product recommendation, and reputation management are some examples of real-world applications for SASM.

#### Media and Journalism

The relationship between journalists and the public became closer thanks to social networking platforms. The recent statistics, published by a 2013 social journalism study, show that 25% of major information sources come from social media data.[3] The public relations professionals and journalists use the power of social media to gather the public opinion, perform sentiment

[3] http://www.cision.com/uk/files/2013/10/social-journalism-study-2013.pdf

analysis, implement crisis monitoring, perform issues- or program-based media analysis, and survey social media.

### Healthcare

Over time, social media became part of common healthcare. The healthcare industry uses social media tools for building community engagement and fostering better relationships with their clients. The use of Twitter to discuss recommendations for providers and consumers (patients, families, or caregivers), ailments, treatments, and medication is only one example of social media in healthcare. This was initially referred to as social health. Medical forums appeared due to the needs of the patients to discuss their feelings and experiences.

This book will discuss how NLP methods on social media data can help develop innovative tools and integrate appropriate linguistic information in order to allow better health monitoring (such as disease spread) or availability of information and support for patients.

### Politics

Online monitoring can help keep track of mentions made by citizens across the country and of international, national, or local opinion about political parties. For a political party, organizing an election campaign and gaining followers is crucial. Opinion mining, awareness of comments and public posts, and understanding statements made on discussion forums can give political parties a chance to get a better idea of the reality of a specific event, and to take the necessary steps to improve their positions.

### Defense and Security

Defense and security organizations are greatly interested in studying these sources of information and summaries to understand situations and perform sentiment analysis of a group of individuals with common interests, and also to be alerted against potential threats to defense and public safety. In this book, we will discuss the issue of information flow from social networks such as MySpace, Facebook, Skyblog, and Twitter. We will present methods for information extraction in Web 2.0 to find links between data entities, and to analyze the characteristics and dynamism of networks through which organizations and discussions evolve. Social data often contain significant information hidden in the texts and network structure. Aggregate social behavior can provide valuable information for the sake of national security.

## 1.3 CHALLENGES IN SOCIAL MEDIA DATA

The information presented in social media, such as online discussion forums, blogs, and Twitter posts, is highly dynamic and involves interaction among various participants. There is a huge amount of text continuously generated by users in informal environments.

Standard NLP methods applied to social media texts are therefore confronted with difficulties due to non-standard spelling, noise, and limited sets of features for automatic clustering and classification. Social media are important because the use of social networks has made everybody a potential author, so the language is now closer to the user than to any prescribed norms [Beverungen and Kalita, 2011, Zhou and Hovy, 2006]. Blogs, tweets, and status updates are written in an informal, conversational tone—often more of a "stream of consciousness" than the carefully thought out and meticulously edited work that might be expected in traditional print media. This informal nature of social media texts presents new challenges to all levels of automatic language processing.

At the surface level, several issues pose challenges to basic NLP tools developed for traditional data. Inconsistent (or absent) punctuation and capitalization can make detection of sentence boundaries quite difficult—sometimes even for human readers, as in the following tweet: "#qcpoli enjoyed a hearty laugh today with #plq debate audience for @jflisee #notrehome tune was that the intended reaction?" Emoticons, incorrect or non-standard spelling, and rampant abbreviations complicate tokenization and part-of-speech tagging, among other tasks. Traditional tools must be adapted to consider new variations such as letter repetition ("heyyyyyy"), which are different from common spelling errors. Grammaticality, or frequent lack thereof, is another concern for any syntactic analyses of social media texts, where fragments can be as commonplace as actual full sentences, and the choice between "there," "they are," "they're," and "their" can seem to be made at random.

Social media are also much noisier than traditional print media. Like much else on the Internet, social networks are plagued with spam, ads, and all manner of other unsolicited, irrelevant, or distracting content. Even by ignoring these forms of noise, much of the genuine, legitimate content on social media can be seen as irrelevant with respect to most information needs. André et al. [2012] demonstrate this in a study that assesses user-perceived value of tweets. They collected over 40,000 ratings of tweets from followers, in which only 36% of tweets were rated as "worth reading," while 25% were rated as "not worth reading." The least valued tweets were so-called presence maintenance posts (e.g., "Hullo twitter!"). Pre-processing to filter out spam and other irrelevant content, or models that are better capable of coping with noise are essential in any language-processing effort targeting social media.

Several characteristics of social media text are of particular concern to NLP approaches. The particularities of a given medium and the way in which that medium is used can have a profound effect on what constitutes a successful summarization approach. For example, the 140-character limit imposed on Twitter posts makes for individual tweets that are rather contextually impoverished compared to more traditional documents. However, redundancy can become a problem over multiple tweets, due in part to the practice of retweeting posts. Sharifi et al. [2010] note the redundancy of information as a major issue with microblog summarization in their experiments with data mining techniques to automatically create summary posts of Twitter trending topics.

A major challenge facing detection of events of interest from multiple Twitter streams is therefore to separate the mundane and polluted information from interesting real-world events.

In practice, highly scalable and efficient approaches are required for handling and processing the increasingly large amount of Twitter data (especially for real-time event detection). Other challenges are inherent to Twitter design and usage. These are mainly due to the shortness of the messages: the frequent use of (dynamically evolving) informal, irregular, and abbreviated words, the large number of spelling and grammatical errors, and the use of improper sentence structure and mixed languages. Such data sparseness, lack of context, and diversity of vocabulary make the traditional text analysis techniques less suitable for tweets [Metzler et al., 2007]. In addition, different events may enjoy different popularity among users, and can differ significantly in content, number of messages and participants, time periods, inherent structure, and causal relationships [Nallapati et al., 2004].

Across all forms of social media, subjectivity is an ever-present trait. While traditional news texts may strive to present an objective, neutral account of factual information, social media texts are much more subjective and opinion-laden. Whether or not the ultimate information need lies directly in opinion mining and sentiment analysis, subjective information plays a much greater role in semantic analysis of social texts.

Topic drift is much more prominent in social media than in other texts, both because of the conversational tone of social texts and the continuously streaming nature of social media. There are also entirely new dimensions to be explored, where new sources of information and types of features need to be assessed and exploited. While traditional texts can be seen as largely static and self-contained, the information presented in social media, such as online discussion forums, blogs, and Twitter posts, is highly dynamic and involves interaction among various participants. This can be seen as an additional source of complexity that may hamper traditional summarization approaches, but it is also an opportunity, making available additional context that can aid in summarization or making possible entirely new forms of summarization. For instance, Hu et al. [2007a] suggest summarizing a blog post by extracting representative sentences using information from user comments. Chua and Asur [2012] exploit temporal correlation in a stream of tweets to extract relevant tweets for event summarization. Lin et al. [2009] address summarization not of the content of posts or messages, but of the social network itself by extracting temporally representative users, actions, and concepts in Flickr data.

As we mentioned, standard NLP approaches applied to social media data are therefore confronted with difficulties due to non-standard spelling, noise, limited sets of features, and errors. Therefore some NLP techniques, including normalization, term expansion, improved feature selection, and noise reduction, have been proposed to improve clustering performance in Twitter news [Beverungen and Kalita, 2011]. Identifying proper names and language switch in a sentence would require rapid and accurate name entity recognition and language detection techniques. Recent research efforts focus on the analysis of language in social media for understanding social behavior and building socially aware systems. The goal is the analysis of language with implications for fields such as computational linguistics, sociolinguistics, and psycholinguistics. For example, Eisenstein [2013a] studied the phonological variation and factors when transcribed into social media text.

Several workshops organized by the Association for Computational Linguistics (ACL) and special issues in scientific journals dedicated to semantic analysis in social media show how active this research field is. We enumerate some of them here (we also mentioned them in the Preface):

- The EACL 2014 Workshop Language Analysis in Social Media (LASM 2014)[4]
- The NAACL/HLT 2013 Workshop on Language Analysis in Social Media (LASM 2013)[5]
- The EACL 2012 Workshop on Semantic Analysis in Social Media (SASM 2012)[6]
- The NAACL/HLT 2012 Workshop on Language in Social Media (LSM 2012)[7]
- The ACL/HLT 2011 Workshop on Language in Social Media (LSM 2011)[8]
- The WWW 2015 Workshop on Making Sense of Microposts[9]
- The WWW 2014 Workshop on Making Sense of Microposts[10]
- The WWW 2013 Workshop on Making Sense of Microposts[11]
- The WWW 2012 Workshop on Making Sense of Microposts[12]
- The ESWC 2011 Workshop on Making Sense of Microposts[13]
- The COLING 2014 Workshop on Natural Language Processing for Social Media (SocialNLP)[14]
- The IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)[15]

In this book, we will cite many papers from conferences such as ACL, WWW, etc.; many workshop papers from the above-mentioned workshops and more; several books; and many journal papers from various relevant journals.

[4] https://aclweb.org/anthology/W/W14/#1300
[5] https://aclweb.org/anthology/W/W13/#1100
[6] https://aclweb.org/anthology/W/W12/#2100
[7] https://aclweb.org/anthology/W/W12/#2100
[8] https://aclweb.org/anthology/W/W11/#0700
[9] http://www.scc.lancs.ac.uk/microposts2015/
[10] http://www.scc.lancs.ac.uk/microposts2014/
[11] http://oak.dcs.shef.ac.uk/msm2013/
[12] htpp://ceur-ws.org/Vol-838/
[13] htpp://ceur-ws.org/Vol-718/
[14] https://sites.google.com/site/socialnlp/2nd-socialnlp-workshop
[15] https://sites.google.com/site/socialnlp/1st-socialnlp-workshop

## 1.4 SEMANTIC ANALYSIS OF SOCIAL MEDIA

Our goal is to focus on innovative NLP applications (such as opinion mining, information extraction, summarization, and machine translation), tools, and methods that integrate appropriate linguistic information in various fields such as social media monitoring for healthcare, security and defense, business intelligence, and politics. The book contains four major chapters.

- **Chapter 1:** This chapter highlights the need for applications that use social media messages and meta-data. We also discuss the difficulty of processing social media data vs. traditional texts such as news articles and scientific papers.
- **Chapter 2:** This chapter discusses existing linguistic pre-processing tools such as tokenizers, part-of-speech taggers, parsers, and named entity recognizers, with a focus on their adaptation to social media data. We briefly discuss evaluation measures for these tools.
- **Chapter 3:** This chapter is the heart of the book. It presents the methods used in applications for semantic analysis of social network texts, in conjunction with social media analytics as well as methods for information extraction and text classification. We focus on tasks such as: geo-location detection, entity linking, opinion mining and sentiment analysis, emotion and mood analysis, event and topic detection, summarization, machine translation, and other tasks. They tend to pre-process the messages with some of the tools mentioned in Chapter 2 in order to extract the knowledge needed in the next processing levels. For each task, we discuss the evaluation metrics and any existing test datasets.
- **Chapter 4:** This chapter presents higher-level applications that use some of the methods from Chapter 3. We look at: healthcare applications, financial applications, predicting voting intentions, media monitoring, security and defense applications, NLP-based information visualization for social media, disaster response applications, NLP-based user modeling, and applications for entertainment.
- **Chapter 5:** This chapter discusses chapter complementary aspects such as data collection and annotation in social media, privacy issues in social media, spam detection in order to avoid spam in the collected datasets, and we describe some of the existing evaluation benchmarks that make available data collected and annotated for various tasks.
- **Chapter 6:** The last chapter summarizes the methods and applications described in the preceding chapters. We conclude with a discussion of the high potential for research, given the social media analysis needs of end-users.

As mentioned in the Preface, the **intended audience** of this book is researchers that 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. Nonetheless, we will try to define as many notions as we can, in order to facilitate the understanding for beginners in these two areas. We also assume basic knowledge of computer science in general.

## 1.5 SUMMARY

In this chapter, we reviewed the structure of social network and social media data as the collection of textual information on the Web. We presented semantic analysis in social media as a new opportunity for big data analytics and for intelligent applications. Social media monitoring and analyzing of the continuous flow of user-generated content can be used as an additional dimension which contains valuable information that would not have been available from traditional media and newspapers. In addition, we mentioned the challenges with social media data, which are due to their large size, and to their noisy, dynamic, and unstructured nature.

# CHAPTER 2

# Linguistic Pre-processing of Social Media Texts

## 2.1 INTRODUCTION

In this chapter, we discuss current Natural Language Processing (NLP) linguistic pre-processing methods and tools that were adapted for social media texts. We survey the methods used for adaptation to this kind of texts. We briefly define the evaluation measures used for each type of tool in order to be able to mention the state-of-the-art results.

In general, evaluation in NLP can be done in several ways:

- manually, by having humans judge the output of each tool;
- automatically, on test data that humans have annotated with the expected solution ahead of time; and
- task-based, by using the tools in a task and evaluating how much they contribute to the success in the task.

We primarily focus on the second approach here. It is the most convenient since it allows the automatic evaluation of the tools repeatedly after changing/improving their methods, and it allows comparing different tools on the same test data. Care should be taken when human judges annotate data. There should be at least two annotators that are given proper instructions on what and how to annotate (in an annotation manual). There needs to be a reasonable agreement rate between the two or more annotators, to ensure the quality of the obtained data. When there are disagreements, the expected solution will be obtained by resolving the disagreements by taking a vote (if there are three annotators or more, an odd number), or by having the annotators discuss until they reach an agreement (if there are only two annotators, or an even number). When reporting the inter-annotator agreement for a dataset, the kappa statistic also needs to be reported, in order to compensate the obtained agreement for possible agreements due to chance [Artstein and Poesio, 2008, Carletta, 1996].

NLP tools often use supervised machine learning, and the training data are usually annotated by human judges. In such cases, it is convenient to keep aside some of the annotated data for testing and to use the remaining data to train the models. Many of the methods discussed in this book use machine learning algorithms for automatic text classification. That is why we give a very brief introduction here. See, e.g., [Witten and Frank, 2005] for details of the algorithms and [Sebastiani, 2002] for how they can be applied to text data.

A supervised text classification model predicts the label $c$ of an input $x$, where $x$ is a vector of feature values extracted from document $d$. The class $c$ can take two or more possible values from a specified set (or even continuous numeric values, in which case the classifier is called a regression model). The training data contain document vectors for which the classes are provided. The classifier uses the training data to learn associations between features or combinations of features that are strongly associated with one of the classes but not with the other classes. In this way, the trained model can make predictions for unseen test data in the future. There are many classification algorithms. We name three classifiers most popular in NLP tasks.

Decision trees take one feature at a time, compute its power of discriminating between the classes and build a tree with the most discriminative features in the upper part of the tree; decision trees are useful because the models can be easily understood by humans. Naïve Bayes is a classifier that learns the probabilities of association between features and classes; these models are used because they are known to work well with text data (see a more detailed description in Section 2.8.1). Support Vector Machines (SVM) compute a hyper plane that separates two classes and they can efficiently perform nonlinear classification using what is called a kernel to map the data into a high-dimensional feature space where it become linearly separable [Cortes and Vapnik, 1995]; SVMs are probably the most often used classifiers due to their high performance on many tasks.

A sequence-tagging model can be seen as a classification model, but fundamentally differs from a conventional one, in the sense that instead of dealing with a single input $x$ and a single label $c$ each time, it predicts a sequence of labels $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ based on a sequence of inputs $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and the predictions from the previous steps. It was applied with success in natural language processing (for sequential data such as sequences of part-of-speech tags, discussed in the previous chapter) and in bioinformatics (for DNA sequences). There exist a number of sequence-tagging models, including Hidden Markov Model (HMM) [Baum and Petrie, 1966], Conditional Random Field (CRF) [Lafferty et al., 2001], and Maximum Entropy Markov Model (MEMM) [Berger et al., 1996].

The remainder of this chapter is structured as follows. Section 2.2 discusses generic methods of adapting NLP tools to social media texts. The next five sections discuss NLP tools of interest: tokenizers, part-of-speech taggers, chunkers, parsers, and named entity recognizers, as well as adaptation techniques for each. Section 2.7 enumerates the existing toolkits that were adapted to social media texts in English. Section 2.8 discusses multi-lingual aspects and language identification issues in social media. Section 2.9 summarizes this chapter.

## 2.2 GENERIC ADAPTATION TECHNIQUES FOR NLP TOOLS

NLP tools are important because they need to be used before we can build any applications that aim to understand texts or extract useful information from texts. Many NLP tools are now available, with acceptable levels of accuracy on texts that are similar to the types of texts used for training the models embedded in these tools. Most of the tools are trained on carefully edited texts, usually newspaper texts, due to the wide availability of these kinds of texts. For example, the Penn TreeBank corpus, consisting of 4.5 million words of American English [Marcus et al., 1993], was manually annotated with part-of-speech tags and parse trees, and it is often the main resource used to train part-of-speech taggers and parsers.

Current NLP tools tend to work poorly on social media texts, because these texts are informal, not carefully edited, and they contain grammatical errors, misspellings, new types of abbreviations, emoticons, etc. They are very different than the types of texts used for training the NLP tools. Therefore, the tools need to be adapted in order to achieve reasonable levels of performance on social media texts.

Table 2.1 shows three examples of Twitter messages, taken from Ritter et al. [2011], just to illustrate how noisy the texts can be.

Table 2.1: Three examples of Twitter texts

| No. | Example |
|---|---|
| 1 | The Hobbit has FINALLY started filming! I cannot wait! |
| 2 | @c@ Yess! Yess! It's official Nintendo announced today that theyWill release the Nintendo 3DS in north America march 27 for $250 |
| 3 | Government confirms blast n #nuclear plants n #japan...don't knw wht s gona happen nw... |

There are two ways to adapt NLP tools to social media texts. The first one is to perform text normalization so that the informal language becomes closer to the type of texts on which the tools were trained. The second one is to re-train the models inside the tool on annotated social media texts. Depending on the goal of the NLP application, a combination of the two techniques could be used, since both have their own limitations, as discussed below (see Eisenstein [2013b] for a more detailed discussion).

### 2.2.1 TEXT NORMALIZATION

Text normalization is a possible solution for overcoming or reducing linguistic noise. The task can be approached in two stages: first, the identification of orthographic errors in an input text, and second, the correction of these errors. Normalization approaches typically include a dictionary of known correctly spelled terms, and detects in-vocabulary and out-of-vocabulary (OOV) terms with respect to this dictionary. The normalization can be basic or more advanced. Basic normalization deals with the errors detected at the POS tagging stage, such as unknown words, misspelled words, etc. Advanced normalization is more flexible, taking a lightly supervised automatic approach trained on an external dataset (annotated with short forms vs. their equivalent long or corrected forms).

For social media texts, the normalization that can be done is rather shallow. Because of its informal and conversational nature, social media text cannot become carefully edited English. Similar issues appear in SMS text messages on phones, where short forms and phonetic abbreviations are often used to save the typing time. According to Derczynski et al. [2013b], text normalization in Twitter messages did not help too much in the named entity recognition task.

Twitter text normalization into traditional written English [Han and Baldwin, 2011] is not only difficult, but it can be viewed as a "lossy" translation task. For example, many of Twitter's unique linguistic phenomena are due not only to its informal nature, but also to a set of authors that is heavily skewed toward younger ages and minorities, with heavy usage of dialects that are different than standard English [Eisenstein, 2013a, Eisenstein et al., 2011].

Demir [2016] describes a method of context-tailored text normalization. The method considers contextual and lexical similarities between standard and non-standard words, in order to reduce noise. The non-standard words in the input context in a given sentence are tailored into a direct match, if there are possible shared contexts. A morphological parser is used to analyze all the words in each sentence. Turkish social media texts were used to evaluate the performance of the system. The dataset contains tweets (~11 GB) and clean Turkish texts (~ 6 GB). The system achieved state-of-the-art results on the 715 Turkish tweets.

Akhtar et al. [2015] proposed a hybrid approach for text normalization for tweets. Their methodology proceeds in two phases: the first one detects noisy text, and the second one uses various heuristic-based rules for normalization. The researchers trained a supervised learning model, using 3-fold cross validation to determine the best feature set. Figure 2.1 depicts a schematic diagram of the proposed approach. Their system yielded precision, recall, and F-measure values of 0.90, 0.72, and 0.80, respectively, for their test dataset.

Most practical applications leverage the simpler approach of replacing non-standard words with their standard counterparts as a "one size fits all" task. Baldwin and Li [2015] devised a method that uses a taxonomy of normalization edits. The researchers evaluated this method on three different downstream applications: dependency parsing, named entity recognition, and text-to-speech synthesis. The taxonomy of normalization edits is shown in Figure 2.2. The method categorizes edits at three levels of granularity and its results demonstrate that the targeted application of the taxonomy is an efficient approach to normalization.

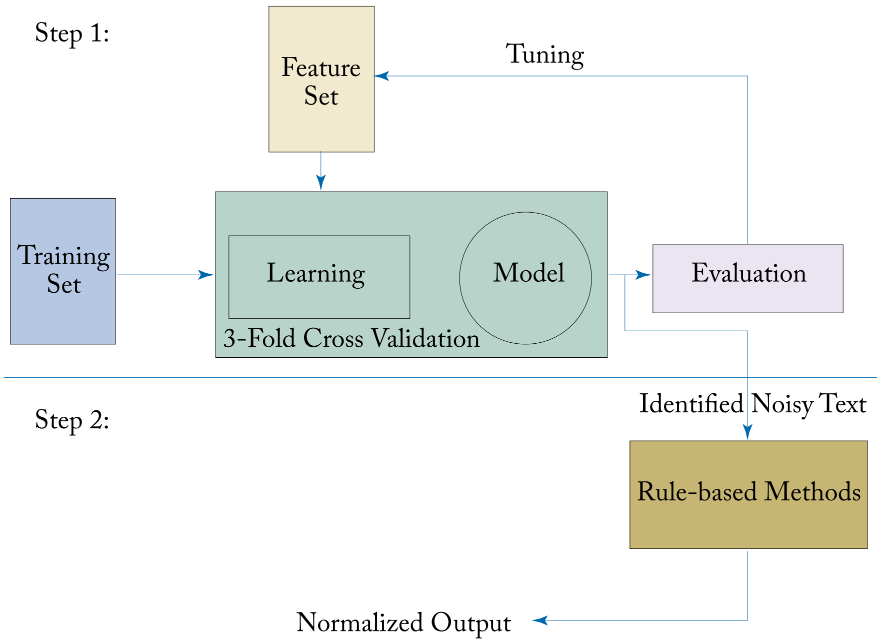


Figure 2.1: Methodology for tweet normalization. The dotted horizontal line separates the two steps (detecting the text to be normalized and applying normalization rules) [Akhtar et al., 2015].

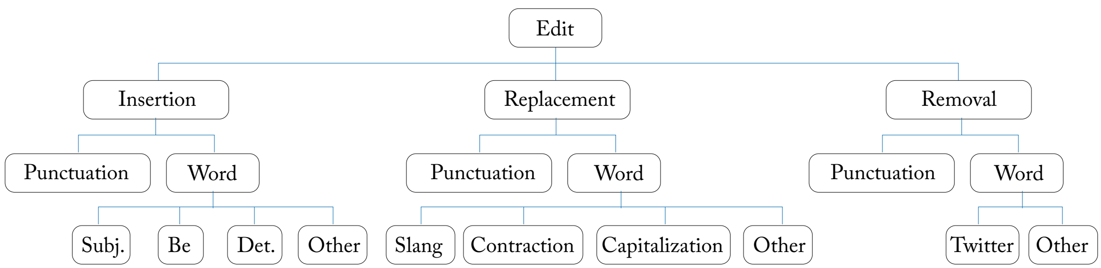


Figure 2.2: Taxonomy of normalization edits [Baldwin and Li, 2015].

### 2.2.2 RE-TRAINING NLP TOOLS FOR SOCIAL MEDIA TEXTS

Re-training NLP tools for social media texts is relatively easy if annotated training data are available. In general, adapting a tool to a specific domain or a specific type of text requires producing annotated training data for that kind of text. It is easy to collect text of the required kind, but to annotate it can be a difficult and time-consuming process.

Currently, some annotated social media data have become available, but the volume is not high enough. Several NLP tools have been re-trained on newly annotated data, sometimes by also keeping the original annotated training data for newspaper texts, in order to have a large enough training set. Another approach is to use some unannotated social media text in an unsupervised manner in addition to the small amounts of annotated social media text.

Another question is what kinds of social media texts to use for training. It seems that Twitter messages are more difficult to process than blog posts or messages from forums. Because of the limitation of Twitter messages to 140 characters, more abbreviations and shortened forms of words are used, and more simplified syntax. Therefore, training data should include several kinds of social media texts (unless somebody is building a tool designed for a particular kind of social media text).

We define the tasks accomplished by each kind of tool and we discuss techniques for adapting them to social media texts.

## 2.3 TOKENIZERS

The first step in processing a text is to separate the words from punctuation and other symbols. A tool that does this is called a tokenizer. White space is a good indicator of words separation (except in some languages, e.g., Chinese), but even white space is not sufficient. The question of what is a word is not trivial. When doing corpus analysis, there are strings of characters that are clearly words, but there are strings for which this is not clear. Most of the time, punctuation needs to be separated from words, but some abbreviations might contain punctuation characters as part of the word. Take, for example, the sentence: "We bought apples, oranges, etc." The commas clearly need to be separated from the word "apples" and from the word "oranges," but the dot is part of the abbreviation "etc." In this case, the dot also indicates the end of the sentence (two dots were reduced to one). Other examples among the many issues that appear are: how to treat numbers (if they contain commas or dots, these characters should not be separated), or what to do with contractions such as "don't" (perhaps to expand them into two words "do" and "not").

While tokenization usually consists of two subtasks (sentence boundary detection and token boundary detection), the EmpiriST shared task[1] provided sentence boundaries and the participating teams only had to detect token boundaries. Missing whitespace characters presents a major challenge to the task of tokenization. Table 2.2 shows a few examples with their correct tokenization.

### Methods for Tokenizers

Horsmann and Zesch [2016] evaluated a method for dealing with token boundaries consisting of three steps. First, the researchers split the text according to the white space characters. Then they employed regular expressions to refine the splitting of alpha-numerical text segments from

[1] https://sites.google.com/site/empirist2015/