10

SOCIAL MEDIA HARVESTING

Man-pui Sally Chan, Alex Morales, Mohsen Farhadloo, Ryan Joseph Palmer, and Dolores Albarracín

The use of social media and social networking sites is currently widespread and is only expected to increase in the coming years. In a recent survey (Greenwood, Perrin, & Duggan, 2016), over 70% of U.S. adults aged 18 to 29 and 55% of adults aged 30 to 49 reported having a Facebook profile. Further, a solid 33% of adults aged 50 to 64 reported using Facebook, and 24% of the 86% of Americans who use the Internet interact on Twitter (Greenwood et al., 2016). Social media comprise websites and applications that facilitate the creation, expression, and sharing of information and ideas among users who (a) maintain a personal profile within the system, (b) privately or publicly interact with other users within the social networks, (c) expand their connections by searching for other users or accepting connections suggested by the platforms, and (d) may also leave the social networks and remove their connections (Boyd & Ellison, 2007).

Social media is an enjoyable outlet for people to express themselves and interact with other network members, and the increasing number of users (Mangukiya, 2016; Statista, 2010), along with the growing usage (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015), position such outlets as powerful sources of information to be used in research (e.g., with the goal of identifying discussion topics on a Facebook page; see "Topics as Important Semantic Features"). In recent decades, social media analysis has received considerable attention in various areas of research, from examining the associations between the use of social media and mental health (Jelenchick, Eickhoff, & Moreno, 2013; Lin et al., 2016) to analyzing the effects of social media on interpersonal relationships (Finkel, Eastwick, Karney, Reis, & Sprecher, 2012; Ward, 2016). Furthermore, the number of psychology articles that utilize social media as a tool or treat it as the subject of scrutiny has risen rapidly in the last decade. According to the Psychology Article Database *Psychinfo*, there were over seven times more studies involving social

media since 2010 than in the entire previous decade. The increase of research publications is expected to rise because social media are growing in popularity and becoming ever more influential in our everyday lives. Researchers can now use social media platforms to harvest a wide range of information about a population, such as the demographics of personal profiles (i.e., non-semantic features) as well as likes, favorites, follow, and text posts/messages (i.e., semantic features).

The harnessing of social media data has allowed researchers to uncover numerous aspects about its users at the individual, community, and national levels. In fact, an emerging group of scholars has analyzed social media data to understand a wide range of behaviors and attitudes, including but not limited to consumer decisions (Bennett & Lanning, 2007; De Souza & Ferris, 2015; Farhadloo, Winneg, Chan, Jamieson, & Albarracín, 2018), influenza infections (Signorini, Segre, & Polgreen, 2011), and political orientation/opinions (Schwartz & Ungar, 2015; Wu, Kosinski, & Stillwell, 2015). In the following sections, we provide a detailed overview of some sample platforms ("Social Media Platforms") and describe different harvesting methods to collect social media data ("Harvesting Social Media Data: Approaches and Sources of Data Collection") as well as a range of harnessing techniques to analyze non-semantic and semantic features ("Harnessing Social Media Data: Analytical Techniques for Non-Semantic and Semantic Features"). We then provide a discussion of important semantic features, including topics and the use of sentiment analysis and opinion spam detection. In the last section, we present an example to illustrate how social media data can be utilized for predictive and explanatory models. Finally, we end this chapter by describing ongoing challenges and future directions of measuring social media data in psychological research.

Social Media Platforms

At first glance, social media might appear to generate data streams that are far too shallow to advance knowledge in any meaningful way because most platforms impose constraints on how users express themselves. For instance, Twitter has a limit of about 280-character on each post/reply (i.e., tweet), Facebook has a 63,206-character allotment for a status update, and Weibo has an approximately 2,000-character restriction for every message, augmented with additional space given for photos, videos, polls, GIFs, and quotes. Given that, by design, these messages are limited, it might seem reasonable to conclude that there is little to learn from the seemingly shallow communications these sites typically generate. However, this is not what we found in a review of the relevant literature. Table 10.1 presents sample studies that have analyzed data from social media and differ in key functions, including networking, microblogging, messaging, commenting and discussion, media sharing, and news and classified advertisements. In the coming sections, we provide an overview of harvesting methods and analytical techniques in relation to the key functions of the social media used in previous studies.

TABLE 10.1 An Overview of Key Functions and Top Social Media With Sample Studies

Key Functions and Top Social Media	*Sample Studies in Social Science*	*Methods of Data Collection (see "Harvesting Social Media Data" and Table 10.2 for approaches and sources)*	*Methods of Data Analyses (see "Harnessing Social Media Data" and Table 10.3 for the uses of semantic features)*
Networking: Users manage a personal profile, which can be used to connect with people with similar interests and background and to create groups for interactions.			
• *LinkedIn—www.linkedin.com*	(Zide, Elman, & Shahani-Denning, 2014)	Three hundred user profiles were collected by groups of human resources and sales/marketing professionals and industrial/organizational psychologists	Carried out human coding of profiles and conducted chi-square tests
• *Facebook—www.facebook.com*	(Kosinski, Stillwell, & Graepel, 2013)	Over 58,000 Facebook-user profiles and a list of their Likes were collected via myPersonality Project	Created a user-Like matrix, reduced the dimensions using singular-value decomposition, and performed linear/logistic regressions
Microblogging: Microblogging focuses on short updates, and users can push updates out to anyone who is subscribed to/following the corresponding account. Followers can pass along updates by reposting.			
• *Twitter—https://twitter.com*	(Ireland, Schwartz, Chen, Ungar, & Albarracín, 2015)	Over 150 million tweets were obtained via Twitter APIs	Mapped tweets to U.S. counties and conducted text analyses, including the use of the Linguistic Inquiry and Word Count (LIWIC) and natural language processing (NLP) techniques
• *Weibo—www.weibo.com*	(Yuan, Feng, & Danowski, 2013)	About 18,000 Weibo messages were collected by searching keyword on weibo.com	Identified a semantic network of messages using WORDij 3.0 and performed a cluster analysis with software package NodeXL.
Messaging: Users can send and receive multimedia messages instantly from family, friends, and other publishers. Messages are presented in various formats, including text, audio, photo, video, and emoticons.			
• *Snapchat—www.snapchat.com/l/en-gb/*	No study available	–	–
• *WhatsApp—https://web.whatsapp.com*	(Cheung et al., 2015)	Messages of 40 WhatsApp users who participated in a smoking cessation intervention were obtained	Carried out human coding of messages and performed Mann-Whitney *U* tests

Commenting and Discussion: Online forums and blog comments allow users to make interactive conversations by posting messages. However, discussion of blog comments usually centers around the topic of the blog post, such as a particular restaurant.			
• *TripAdvisor—www.tripadvisor.com*	(Lawrence & Perrigot, 2015)	Over 6,000 hotel reviews from 134 hotels were collected automatically	Carried out human coding of reviews and performed regression analyses
• *Yelp—www.yelp.com*	(Gui, Zhou, Xu, He, & Lu, 2017)	More than 1 million reviews and user data were obtained from Yelp 2013 and 2014 Data Challenge Dataset	Conducted sentiment classification on reviews
Media Sharing: These platforms allow users to upload and share various media such as pictures and video. Additional functions include creating profiles, creating/subscribing channels, commenting, etc.			
• *YouTube—www.youtube.com*	(J. Huang, Kornfield, & Emery, 2016)	Over 28,000 videos related to e-cigarette were recorded via a YouTube crawling program, ContextMiner	Performed human coding of content and carried out descriptive analyses
• *Instagram—www.instagram.com*	(Moreno, Ton, Selkie, & Evans, 2016)	Over 1 million Instagram posts were obtained by searching for nonsuicidal self-injury (NSSI) hashtags www.instagram.com	Performed human coding to assess the NSSI hashtags meaning
News and Classified Advertisements: Users can post various items, such as classified advertisements, news, and links to third-party articles, pictures, or videos. Users are then allowed to interact with the items. For example, the order of display of the items on Reddit is subject to time or to number of votes, which is the core social aspect in these communities. The Reddit community jointly decides which news items get seen by more people.			
• *Reddit—www.reddit.com*	(Zhan, Liu, Li, Leischow, & Zeng, 2017)	Over 27,000 posts were collected by keyword searches and analyses of metadata via Reddit API	Performed topic modeling using natural language processing (NLP) techniques
• *Craigslist—www.craigslist.org*	No study available	–	–

As revealed in Table 10.1, social media functions related to networking are more appropriate to address research questions about social networks. For example, researchers may use profile information on LinkedIn to explore how users of different professions present themselves on LinkedIn (Zide et al., 2014). Similarly, social media designed for commenting and discussion, such as Yelp, may allow researchers to examine the use of positive versus negative words in reviews of restaurants and shops (Gui et al., 2017). Additionally, some platforms, such as Facebook and Twitter, combine networking, microblogging, and commenting functions, which offers ample opportunity for research. The use of data from these multi-function social media is thus less restrictive than that of data generated from social media with a single function. Previous studies collected Twitter data to examine the relation between the usage of pre-identified vocabularies and health outcomes (Ireland, Chen, Schwartz, Ungar, & Albarracin, 2015; Ireland, et al., 2015) and harvested Facebook Likes data to predict dispositional characteristics (Kosinski et al., 2013; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones et al., 2013). As different harvesting methods yield distinct data characteristics, we next discuss the available harvesting methods.

Harvesting Social Media Data: Approaches and Sources of Data Collection

Social media data are available from a variety of sources, and the data can be collected via different approaches varying in cost, programming techniques required, and data completeness (see Figure 10.1). Figure 10.1 illustrates how six harvesting approaches relate to the data costs, the representativeness of the sample, and the subsequent cleaning and processing procedures. Researchers must carefully select which sources of data best suits their research needs. For example, researchers with limited resources or those who want to pilot test research ideas may use existing free datasets, whereas researchers who have sufficient IT resources and want to monitor the influence of policies in the public for a period, may use the services of monitoring vendors and/or set up application program interfaces (APIs).

Given the availability of multiple sources for each approach, we present the most common and up-to-date sources with their main features in Table 10.2. In the following section, we discuss harvesting social media data from the least expensive approach to the most expensive ones. However, the data access policy of social media platforms can change and thus requires researchers to check the social media regulations at the time of conducting their research.

The Least Expensive Approach: Downloading Free Datasets

The fastest and least expensive approach to accessing social media data is through sample data libraries and directly from the social media sites, such as Yelp, Wikipedia, and YouTube (see Table 10.2). These free datasets have several limitations

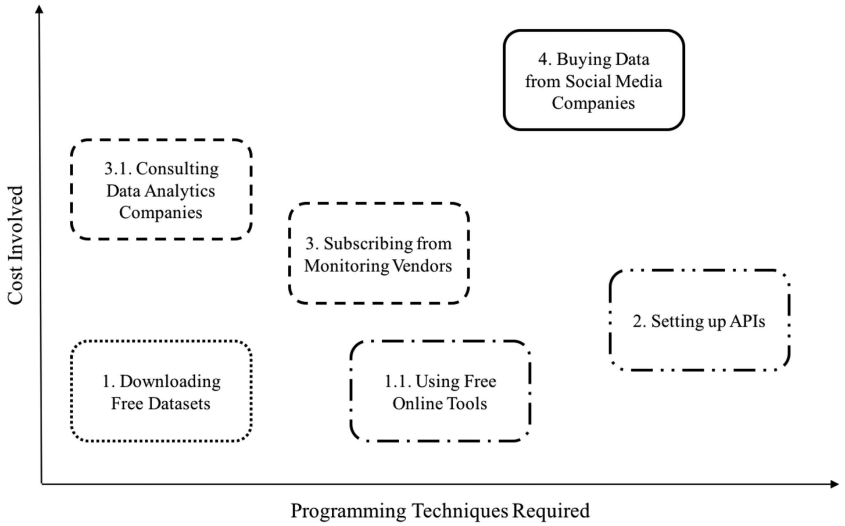

FIGURE 10.1 Various Approaches to Social Media Data Harvesting

Note: By cost, programming techniques and data completeness (solid line refers to the maximum level of completeness, long-dash-dot-dot lines refer to the moderate level of completeness, dash lines refer to the moderate-to-minimum level, and dot line refers to the minimum level of completeness).

because the data collection is already completed, limited user metadata are included, and the researchers have no control over the cleaning/preprocessing processes. Despite these limitations, researchers can use such data sets for validating techniques and pilot testing of their hypotheses.

Apart from these dataset repositories, researchers can also access free social media data (e.g., the Observatory on Social Media (OSoMe), developed by academic initiatives at Indiana University to promote public access to social media data). Moreover, Twitter Search, the search function available on the Twitter website, can also provide a small sample of Twitter data (i.e., retrieving up to seven days of historical data or 1,500 tweets). This method requires researchers to manually copy and paste search results into a database, which is cumbersome for a project examining an extended period.

The Less Expensive Approach: Setting Up Application Program Interfaces (APIs)

Social media platforms publish different APIs, which are sets of protocols and tools to enhance the functionalities of software applications developed by researchers. Researchers are required to register as a developer and obtain consumer and access token credentials to set up the API. Although using the APIs are free, there are tangible costs in setting up/monitoring the API and storing the data.

TABLE 10.2 A Summary of Approaches and Sample Sources of Social Media Data Collections

Approaches	*Sample Sources*	*Sample Social Media*	*Major Characteristics*	*Websites*
Download from Data Repositories	Stanford Large Network Dataset Collection	Facebook, Google+, Friendster, etc.	• Datasets have different time periods • Each dataset has different attributes	https://snap.stanford.edu/data/#onlinecoms
	Social Computing Data Libraries	Flickr, Twitter, YouTube, etc.	• Datasets have different time periods • Each dataset has different attributes	http://socialcomputing.asu.edu/pages/datasets
	Network Repository	Facebook, Twitter, YouTube, etc.	• Repositories donated by other users • Built-in interactive graph analytic tools for visualizing social networks	http://networkrepository.com/
	myPersonality Project	Facebook	• Match participants' self-reported questionnaires with Facebook data • Access to participants' Facebook profile and social network data via a Facebook application	https://www.psychometrics.cam.ac.uk/productsservices/mypersonality
	Yelp Dataset Challenge	Yelp	• Release annually by Yelp Co. • Datasets include information for a small number of cities and countries	https://www.yelp.com/dataset_challenge
	Wikipedia Database	Wikipedia	• Download database directly from Wikipedia • Databases are released on a regular basis	https://en.wikipedia.org/wiki/Wikipedia:Database_download
	YouTube-8M Dataset	YouTube videos	• Videos are pre-processed and selected from a list of popular contents • Each video was tagged with labels	https://research.google.com/youtube8m/
Use Free Tools	Observatory on Social Media (OSoMe)	Twitter	• Allow access to about 1% of total public tweets since 2010 • Offer a set of web-tools to study how information/ideas spread online	https://osome.iuni.iu.edu/tools/

Set-up APIs	Twitter APIs	Twitter	• Use Streaming and REST APIs to collect tweets and metadata (in JSON format) • Able to stream up to 1% of total public tweets • Rate limits of the REST APIs are applied.	https://dev.twitter.com/overview/api https://dev.twitter.com/rest/public/rate-limiting
	Facebook Graph API	Facebook	• Search a user, page, event, group, place, and topic (in JSON format) • Rate limits of the API are imposed on each page/group	https://developers.facebook.com/ https://developers.facebook.com/docs/graph-api/advanced/rate-limiting
	Instagram APIs	Instagram	• Use Real-Time and REST APIs to collect contents and metadata (in JSON format) • Different rate limits are imposed to the modes of applications	https://www.instagram.com/developer/ https://www.instagram.com/developer/limits/
	Reddit API	Reddit	• Specify API to collect data at various levels e.g., listings, live threats, and forums (subreddit) etc. (in JSON format) • Rate limit of 60 requests per minute is imposed	https://www.reddit.com/dev/api/ https://github.com/reddit/reddit/wiki/API#rules
	Weibo API	Weibo	• Use Weibo API to collect messages and metadata (in JSON format) • Rate limit of 150 requests per hour is applied	http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3_V2/en http://open.weibo.com/wiki/Account/rate_limit_status/en
	LinkedIn API	LinkedIn	• Use REST API to collect data of users and companies (in XML or JSON formats) • Rate limits are imposed to each application.	https://developer.linkedin.com/ https://developer.linkedin.com/legal/api-terms-of-use

(*Continued*)

TABLE 10.2 (Continued)

Approaches	*Sample Sources*	*Sample Social Media*	*Major Characteristics*	*Websites*
	YouTube API	YouTube	• Use data v3 API to collect data of video, channel, or playlist (in JSON format) • Rate limit of 1 million units per day is imposed	https://developers.google.com/youtube/v3/getting-started
Subscribe Monitoring Services	Crimson Hexagon	Twitter, Facebook, Instagram, Blogs, Forums, News, Comments, Reviews, and YouTube, etc.	• Create monitors by selecting sources of data and setting parameters/filters • Access to tweets since 2008 and export data in CSV format • Rate limits of 10,000 tweets per export and 50,000 tweets per day are imposed	https://www.crimsonhexagon.com/
	DataSift	bitly, Blogs, DailyMotion, Instagram, Facebook, Tumblr, and YouTube, etc.	• Use built-in algorithms for sentiments, topics, and content analyses • Access to historical data and export data in JSON and CSV formats • Rate limit of 500,000 per day is imposed	http://datasift.com/
	Watson Analytics for Social Media	Twitter, Facebook pages, Forums, Blogs, Reviews, and Videos, etc.	• Use built-in algorithms for identifying important keywords • Use built-in algorithms for sentiments, topics, and content analyses • Rate limits by types of subscription are imposed	https://www.ibm.com/us-en/marketplace/social-media-data-analysis/purchase
Buy Data	Gnip, Inc.	Twitter	• Use APIs to collect real-time and historical tweets • Use APIs to obtain aggregate interest and demographic information for a collection of Twitter users	https://gnip.com/sources/

Furthermore, basic familiarity with programming techniques, as well as server side programming languages, are necessary for the use of APIs. For instance, researchers have to be familiar with Python, a programming language, to use Tweepy, an open-sourced python program, to communicate with the Twitter API python package (see http://docs.tweepy.org/en/v3.5.0/getting_started.html for details). Other intangible costs include the absence of retrospective data (because data are crawled prospectively) and the time required for data cleaning (because of missing fields and inconsistent information).

Apart from the accredited APIs, free web scraping programs available online supply tools to scrape information on designated websites and save into a JSON and XML format. These automated software programs (also referred to as bots) can also utilize fake user accounts to harvest data on social media.

Despite the availability, researchers should be cautious about the legal constraints of such tools. In 2016, LinkedIn filed a lawsuit against 100 unnamed individuals using bots to harvest user profiles from its website (Conger, 2016; LinkedIn, 2016). Web scraping tools are also subject to regulations (Bilton, 2012). For example, in 2015, the airline company Ryanair sued other travel agencies for screen-scraping price information. The Court of Justice of the European Union (ECJ) ruled that websites can set restrictions to limit scraping (Consonni & Anselmi, 2015). As the legitimacy of web scraping tools is bounded by the laws of respective countries, researchers should consult their institutions' legal services before scraping social media data.

The More Expensive Approach: Subscribing Services From Monitoring Vendors

Although the use of computer programs for harvesting involves concerns about technical and legal issues, subscribing services from monitoring vendors can make data retrieval, preparation, and basic analysis potentially easier (see Table 10.2). Monitoring vendors, such as Crimson Hexagon and DataSift, pre-process social media data, such as from Facebook, Weibo, Twitter, and provide information through automatic dashboards, real-time social listening and influencer identification tools, as well as built-in visualization tools (e.g., word cloud, and figures). However, a major drawback of such vendor services is the subscription cost, which may be very prohibitive depending on the retrieval volume and types of data. Furthermore, users can neither customize the algorithms of the built-in analyses nor modify any parameters of the machine learning model for analyses.

The Most Expensive Approach: Buying Data

The most expensive option regarding harvesting social media is buying data directly from the reseller. Gnip is a Twitter data reseller that provides the full raw Twitter data and sells the data to match the researchers' needs by customizing the

programming infrastructure and computational algorithms. Interested researchers can contact the reseller and purchase a dataset that meets specific needs, and there are occasional promotions and grants for academic work. Apart from Twitter data, to our knowledge, there are no other official resellers of raw social media data currently available. This approach certainly gives investigators complete control of how to retrieve, store, and analyze the full sample of Twitter data, but is infrequently used given the extensive cost and resources needed to build such a system.

Harnessing Social Media Data: Analytical Techniques for Non-Semantic and Semantic Features

Apart from the collection of social media data, a major challenge of using social media data for research is the selection of an appropriate analytical technique to measure the variables of interest. Social media data can be included in the analysis as measured variables or used to extract latent variables, depending on the characteristics. Figure 10.2 presents an overview of social media data, including two main features, i.e., non-semantic and semantic, and the corresponding analytical techniques. Non-semantic features include attributes of non-lexical items, such as age, gender, and location, which are usually specified on user profiles. Semantic features include lexical items with different levels of information, ranging from less detailed contents, such as Facebook Likes, to more detailed ones, such as text messages. As social media data vary in characteristics, the analytical techniques vary. For example, user attributes can be entered directly into a regression analysis whereas text messages require content analysis or natural language processing, followed by regression analyses. In the following sections, we reference published studies to illustrate how various analytical techniques can be used to research non-semantic and semantic features.

Non-Semantic Features

The first and most obvious application of social media analysis is to measure the demographic characteristics of populations. Social media data can reveal characteristics of the populations, especially those that are difficult to reach or less likely to participate in a survey. The majority of participants that are studied using traditional research methods are mainly white, female, Western undergraduate students, a.k.a. the "WEIRD" demographic described in Henrich, Heine, and Norenzayan (2010). Given that sampling directly from other regions of the globe and collecting responses from a national representative samples can be extremely expensive and time-consuming (Teitler, Reichman, & Sprachman, 2003), the analysis of social media data is likely to allow researchers to measure global populations on a larger scale, with a lower investment of money and time. The relatively low costs related to social media has enabled their use in everything from

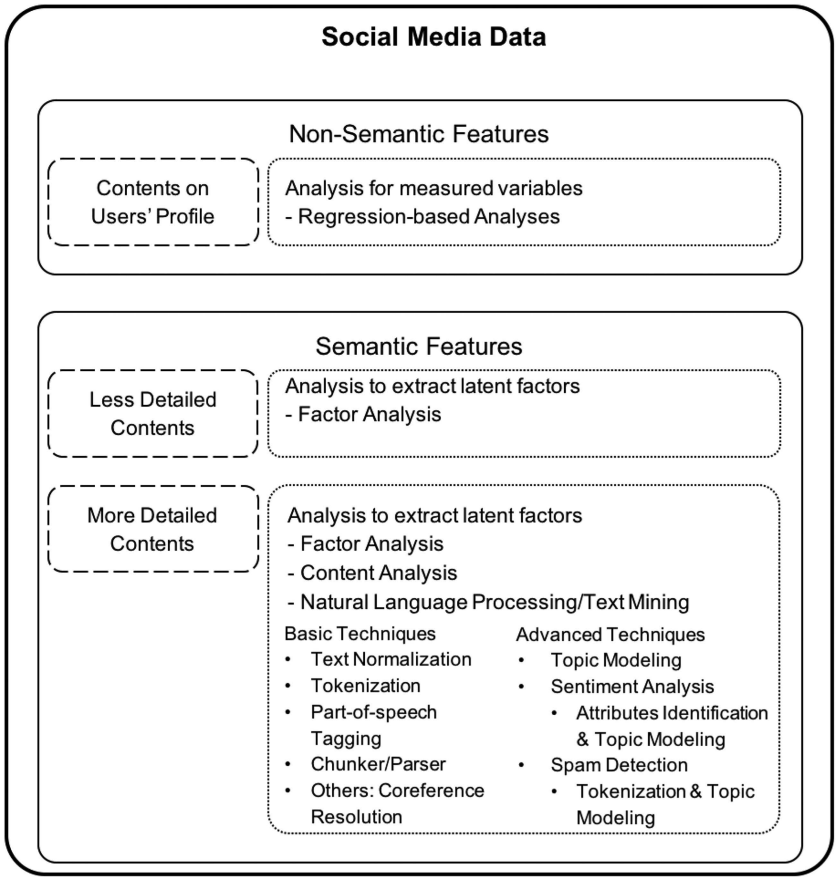

FIGURE 10.2 Data Features and Analytical Techniques of Social Media Analyses

the analysis of the effects of advertisements on consumer behavior, to those of government-led health campaigns on public opinion (Smyser, 2013).

Furthermore, the penetration of social media gives researchers the opportunity to examine broader research questions in which demographics (and other individual differences) can be systematically studied. Researchers can investigate the spread of emerging risk behaviors such as electronic cigarettes (e-cigarette) and vaping (Chu et al., 2015) in particular regions and age groups by analyzing social media non-semantic features such as location information specified on user profiles. Researchers can also examine the effectiveness of tobacco control campaigns in social media because lesbian, gay, and bisexual (LGB) populations are more likely to be smokers and social media users, compared to heterosexual ones

(Kostygina, Tran, & Emery, 2016; Seidenberg et al., 2017; Stevens, Carlson, & Hinman, 2004). Ultimately, researchers can sample diverse users from social media and use their responses to test theories that hypothesize variability on race, gender, education, culture, etc.—the variability unlikely to be found in "WEIRD" college samples.

In addition to the demographic information, researchers can quantify the size of users' social networks by measuring the number of friends and followers on social media. Lönnqvist and Itkonen (2014) examined the mediating role of social network size on the link between personality traits and life satisfaction. Instead of asking individuals to report how many friends they have, Facebook friend counts can serve as a proxy for their social network size. Likewise, Johnston, Tanner, Lalla, and Kawalski (2013) used Facebook friend counts to gauge the levels of social capital and examined its impact on subjective well-being. Other researchers recorded changes in friendship ties on Facebook and MySpace as a measure of friendship selection, which was then linked to smoking and drinking behavior (G. C. Huang, Soto, Fujimoto, & Valente, 2014).

Although non-semantic features, including users' demographic information and their social networks, are important for analyses, not everybody is willing to provide complete information on their social media profile. Only 20% of users provide demographic information and meaningful locations in their profile (Cheng, Caverlee, & Lee, 2010). Due to the sparseness of attributes in social media data, researchers have begun to use the available profile information to predict other missing user attributes including age, gender, ethnicity/race, location, language, and other demographic characteristics (S. Chang et al., 2014; Rao et al., 2011; Schwartz, Eichstaedt, Kern, Dziurzynski, Agrawal et al., 2013; Zamal, Liu, & Ruths, 2011). Previous studies have demonstrated satisfactory performance of these predictors and classifiers, even though attribute identification tasks are still resource-intensive due to the use of manual annotation procedures. For example, previous studies relied on users' first name on their account profiles to infer gender (Burger, Henderson, Kim, & Zarrella, 2011), even though the accuracy of this method is not well validated.

Semantic Features

Less Detailed Contents

Other studies have examined semantic data (e.g., Facebook Likes) to predict personal attributes, personality traits, and psychological outcomes (Kosinski et al., 2013; Wu et al., 2015). Table 10.3 summarizes common semantic data with less detailed contents that are available in the top three social media (see the left side of the table). Favorite/like, follow, and share/retweet are the examples of semantic data that work similarly as web browsing cookies: Clicking Favorite/Like for a message indicates users' positive evaluations of that post, clicking Share/Retweet

TABLE 10.3 Examples of Less-Detailed and More-Detailed Semantic Features on Facebook, Twitter, and Instagram and Possible Research Questions

Sample Media	*Less-Detailed*		*More-Detailed*	
	Features	*Possible Research Questions*	*Features*	*Possible Research Questions*
Facebook	Post reactions	How do post-reactions (i.e., clicking Like, Love, Haha, Wow, Sad, and Angry) towards messages link to voting preferences?	Individual posts	What are the levels of satisfaction with a product?
	Follow/like	How do *Likes*/*Followings* (i.e., clicking Like or Follow) of pages/groups relate to dispositional characteristics?	Post conversations	What are the sentiments of a specific event?
	Share	How do *Sharing* (i.e., clicking Share) of messages on his/her Facebook timeline relate to attitudes, beliefs, and behaviors?		
Twitter	Favorite/like	How do *Likes* (i.e., clicking Like) of tweets link to related donation campaigns?	Individual tweets	What are the topics discussed in the community?
	Follow	How do *Followings* (i.e., clicking Follow) of other user accounts relate to mental and physical well-being?	Tweet conversations	What are the important factors in discussion of HIV prevention?
	Retweet	How do *Retweets* (i.e., forwarding messages) relate to the perceptions of public health crisis?		
Instagram	Like	How do *Likes* (i.e., clicking Like) of posts link to music/movies preferences?	Photo captions	What are the factors that suggest positive dyadic relationships?
	Follow	How do *Followings* (i.e., clicking Follow) of user accounts relate to social norms?		

involves forwarding a message posted by other users, and clicking Follow shows users' choice of receiving all updates from that page/group. Even though these semantic features are minimal or condensed, they are useful for examining a wide range of research questions (see Table 10.3). For example, Kalampokis, Tambouris, and Tarabanis (2013) used Facebook Likes data to develop machine learning models to predict personal attributes, going from sexual orientation to intelligence. Wu et al. (2015) further validated the predicted personality scores and revealed that computer-based personality predictions, rather than the estimates made by the participants' Facebook friends, were more highly correlated with participants' self-report scores. Semantic features such as Likes and the related analytical techniques are likely to have a major influence on psychological research in the next decade. Apart from Likes/Follow, researchers can collect users' Share/Retweet as clear expressions of particular events and apply machine learning techniques to predict psychological variables without asking participants to complete self-report questionnaires. The collection of semantic data and the corresponding analyses tend not to be limited to particular social media, with a caveat that Facebook frequently changes APIs for public access to their contents, which creates uncertainties about Facebook as a stable data source.

More Detailed Contents

An expanding body of research has concentrated on the content analysis of semantic features with more detailed contents, such as posts and messages on social media (see Table 10.3 for examples; Curini, Iacus, & Canova, 2015; De Souza & Ferris, 2015). Such messages and posts may include emoticons, which are the use of keyboard characters to represent a facial expression, such as a smile ":-)", and text content that can be used to directly reveal a user's emotion. Researchers can either use tweets originally codified by Twitter as happy versus unhappy for valence analyses (Curini et al., 2015) or analyze the message content to obtain verbal information. As described in Table 10.3, the analyses of individual messages posted on social media allow marketing campaigners to understand the level of satisfaction with a product (De Souza & Ferris, 2015). Using social media data to measure customer satisfaction resembles the collection of product comments in focus groups, except that the online customers can participate in the product review meeting whenever and wherever they want. Additionally, the general usage of certain words can also reflect an individual's emotions, thoughts, and behaviors. Therefore, an emerging field uses social media data to infer users' behaviors, attitudes, and health status. For example, a study conducted by Asur and Huberman (2010) showed that Twitter data could predict how many tickets would be purchased for the upcoming release of a movie. These findings indicate that the analysis of social media to derive semantic features is likely to provide valuable insights.

Scientists have been studying how to convert raw text and its representations into manageable inputs for computers since the early 1960s. Natural language processing (NLP), which aims to understand human communications using computers, has allowed researchers to extract meaningful representations from text messages (i.e., words, phrases, and sentences) and use them as inputs for machine learning models (see Figure 10.2). The major use of NLP research once concentrated on deriving representations from structured text passages in formal written language, such as news articles, academic journal articles, records, and archives. However, as social media data have become increasingly available, NLP techniques have evolved to analyze the short and unstructured user-generated message contents that characterize posts and messages on social media. The most well-established basic NLP techniques include text normalization, tokenization, part-of-speech tagging, chunkers and parsers, as well as named-entity recognition. Other basic NLP methods that have not yet received much attention in social media analysis include coreference resolution. These new technologies are attempts to respond to the challenges of understanding user-generated content on social media, such as identifying HIV risk among users (Thangarajan, Green, Gupta, Little, & Weibel, 2015) or predicting crime rates using Twitter data (Gerber, 2014).

The first step in applying natural language processing is text normalization, which is an abstraction used to convert raw text into a standardized representation. This step involves some knowledge of the data available and how it will be utilized. For example, Harrison et al. (2014) have collected restaurant reviews in which customers have described various aspects of each restaurant such as location, food quality, atmosphere, and price. If a researcher interested in analyzing price information may find some customers using "$" to describe the monetary price while others might use the word "dollars," which requires consolidating different representations into one norm. Researchers can, of course, substitute numerical characters for respective words. Similarly, there are methods for word stemming (e.g., maps the texts *car*, *cars*, *car's*, and *cars'* to *car*), stop words removal (e.g., removes words like *a*, *an*, and *are*, etc.), and lower-case conversions (e.g., converts *Health* to *health*). However, the adoption of these methods in social media analysis requires consideration of informal language use, idiosyncratic writing styles, and vernacular orthography (e.g., *that* as *dat*; Beckley, 2015). Tweets may signal emphasis with capitalization, which is traditionally used for the starting boundary of a sentence or some named entity. Furthermore, tweets contain punctuations that are used not just to end a sentence, but also as a part of emoticons (Kaufmann, 2010).

The second step in text preprocessing is text tokenization, which reduces raw texts to a number of basic units, typically in the form of words, phrases, sentences, and/or paragraphs (see Figure 10.2). For instance, an n-grams tokenizer breaks the text down into a contiguous sequence of n items such as words; an n-gram of size one refers to as a unigram, and an n-gram of size two refers to as a bigram,

etc. The tokenizers also segment sentences into valid partitions. For example, the punctuation period "." usually indicates the end of a sentence, although applying such a rule to a sentence with a term "U.S.A." may lead to incorrect segmentation, resulting in meaningless text fragments. In this situation, a text tokenizer would decode the word set correctly into "the United States of America." This example suggests the need for more analytic tools to tackle the challenges of informal language (Gimpel et al., 2011; Ritter, Clark, Mausam, & Etzioni, 2011). Another basic form of syntactic analysis can be derived from identifying the part-of-speech (POS) components of a sentence (i.e. nouns, verbs, adjectives, etc.). Although many POS taggers and tokenizers are trained using a standard corpus (the Wall Street Journal corpus), Gimpel et al. (2011) developed a Twitter POS tagger and tokenizer tool, which creates an appropriate annotation corpus for the training of the text preprocessing tool. The importance of Gimpel and colleagues (2011) tool lies in the invention of phonetic normalization, which derives a common representation of a word that receives many alternate spellings on Twitter.

The third NLP step is to identify some structure in texts, that is, parsing grammatical components of sentences, such as noun, prepositional, or verb phrases (see Figure 10.2). This goal is achieved by parsers, an umbrella term for fully grammatical parsers and shallow parsers/chunkers. The challenge of identifying structures in texts is that very few structures exist, not to mention the presence of large amounts of noise. Hence, parsers developed for Twitter typically perform less accurately than tools developed for news articles or journals (Kong et al., 2014). Named entity recognition (NER) is another process of identifying and categorizing tokens that refer to people, locations, organizations, etc. NER may be useful when a researcher tries to identify tweets about the World Health Organization (WHO), a case in which the keyword search "WHO" is likely to return noisy results. In that case, tweets can be further processed with NER to identify correct tweets, but currently, this process only works for tweets with sufficient textual content, i.e., the larger the number of characters the better the performance (Ritter et al., 2011).

In addition to the well-established NLP techniques, we present recent NLP techniques that have not yet been widely applied to analyze social media data but might improve analysis in the future (see Figure 10.2). Coreference resolution is a basic NLP technique that involves identifying noun phrases and clustering those that refer to the same named entity (K. Chang, Samdani, & Dan Roth, 2013). Despite the availability of various techniques, their performance at correctly identifying referents depends on the presence of structure or context, both of which are limited in Twitter and other social media. To improve coreference resolution methods, scientists have begun research in cross-document coreference resolution to identify if two mentions refer to the same concept (Upadhyay, Gupta, Christodoulopoulos, & Roth, 2016). Alvarez-Melis and Saveski (2016) have proposed an interesting approach to overcome the limited content issue by keeping track of the conversation on Twitter and aggregating the tweets replying

to the original tweets. Such an approach is likely to gather more tweets that meet the needs of NLP methods and generate more accurate results (Alvarez-Melis & Saveski, 2016).

Given the unique writing style and sentence structure of posts and messages on social media, scientists are actively developing new techniques to address such challenges, leading to steady progress in the advancement of NLP research on social media data. There are many collective efforts and conferences, such as the Workshop on Semantic Evaluation (SemEval), the Text Retrieval Conference (TREC) and the Workshop on Noisy User-generated Text (WNUT), which are dedicated to advance the state-of-the-art (to improve the performance) in text normalization, tokenization, named entity recognition, and other methods for Twitter and other media (Baldwin et al., 2015). In the realm of measuring social media data, NLP techniques can serve both the purpose of language identification and the less attended problem of improving data quality. In the next section, we present other advanced NLP techniques (i.e., topic modeling for text mining, sentiment analyses, and spam detection), which can be incorporated to identify meaningful semantic features and improve data quality for further language identification and analysis.

Topics as Important Semantic Features

Topic modeling is widely used to cluster semantically similar words that frequently co-occur in a collection, and each cluster refers to a topic, which corresponds to a different distribution of words. Among the most popular methods for discovering topic models are Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003; Hofmann, 1999). These topic models adopt hierarchical Bayesian networks that do not require labeled training data and are able to identify topics (i.e., latent semantic features) in an unsupervised fashion. LDA assumes that the documents contain a mixture of topics and that each topic includes a list of words based on their probability distribution. LDA attempts to figure out what topics emerge in a particular set of documents. It is a matrix factorization technique, and the corpus (a collection of documents) can be represented as a document-term matrix. The corpus has N documents D1, D2, . . . Dn and a vocabulary size of M words W1, W2 . . . Wm. We can apply the LDA model for converting such a document-term matrix into two lower dimensional matrices: M1 is a document-topics matrix with dimensions (N, K) and M2 is a topic-terms matrix with dimensions (K, M), where N is the number of documents, K is the number of topics, and M is the vocabulary size. Although these two matrices provide the distributions of topic word and document topic, such distributions need further improvement by making use of sampling techniques. Therefore, LDA, for example with Gibbs sampling, iterates through each word for each document and tries to adjust the current topic-word assignment with a new assignment (Gilks, Richardson, & Spiegelhalter, 1996).

A steady state, or convergence point, is achieved with satisfactory distributions of the document topics and topic words after multiple iterations. The identified topic model captures topic proportions and assignments as well as the weights of each word in a specific topic in each document (i.e., the measurement unit).

Topic models can help to organize and offer insights about large collections of unstructured text messages. Consider an analysis of Facebook to identify popular topics. In this example, we used the Python package scikit-learn (other packages are also available, see https://pypi.python.org/pypi/lda for details). Furthermore, the topic modeling analysis can be performed in R and other computer languages based on available resources and familiarity with the programming environments. We first collected data and prepared the documents, that is using Facebook API to collect posts on the Society for Personality and Social Psychology (SPSP) Facebook page from November 10 to December 11 in 2016. As the SPSP page is a public page where subscribers can freely post messages, the sources of messages varied from mainstream news media sites to specific research-oriented outlets. The top five sources of messages include "The Wall Street Journal," "The New York Times," "The Atlantic," "VOX," and "Washington Post," all traditional rather than academic media. Second, we used the Python package scikit-learn to remove all stop words (e.g., and, the, is, etc.) and then tokenized the corpus into bigrams (i.e., a sequence of two adjacent words). Next, we converted the bigrams into a document-term matrix using the built-in function of the package, created an object for the LDA model, and trained it on a document-term matrix. We set a few parameters as required in the training (see Appendix 10.1 for the sample codes). Finally, from the training corpus, we identified an LDA model that could be used to discover topic distributions of posts on other Facebook pages (i.e., new and unseen documents). Figure 10.3 shows first five topics with top-20 words (due to limited of space) that were identified in this example.

Identification of Sentiments

Social media has become a unique platform for individuals to express their opinions and is a valuable source for researchers to examine attitudes in diverse areas. However, the size and the complexity of the social media data require the development of automatic methods for organizing, analyzing, and extracting attitudes. The main objective of sentiment analysis is to identify attitudes (either positive or negative) in a corpus (i.e., a collection of documents, and each document is a unit of measurement). Sentiment analysis varies in scope, ranging from the document- and sentence-level to the aspect-levels. In the following paragraphs, the discussion focuses on the aspect-level analysis, which first extracts attributes (aspects) of the object and then identifies the sentiments of those attributes (Farhadloo & Rolland, 2013; Hu & Liu, 2004; Popescu & Etzioni, 2005; Su et al., 2008).

In recent years, different text mining techniques have flourished to extract attributes (i.e., attitudes) of the object. A group of researchers has proposed automatic

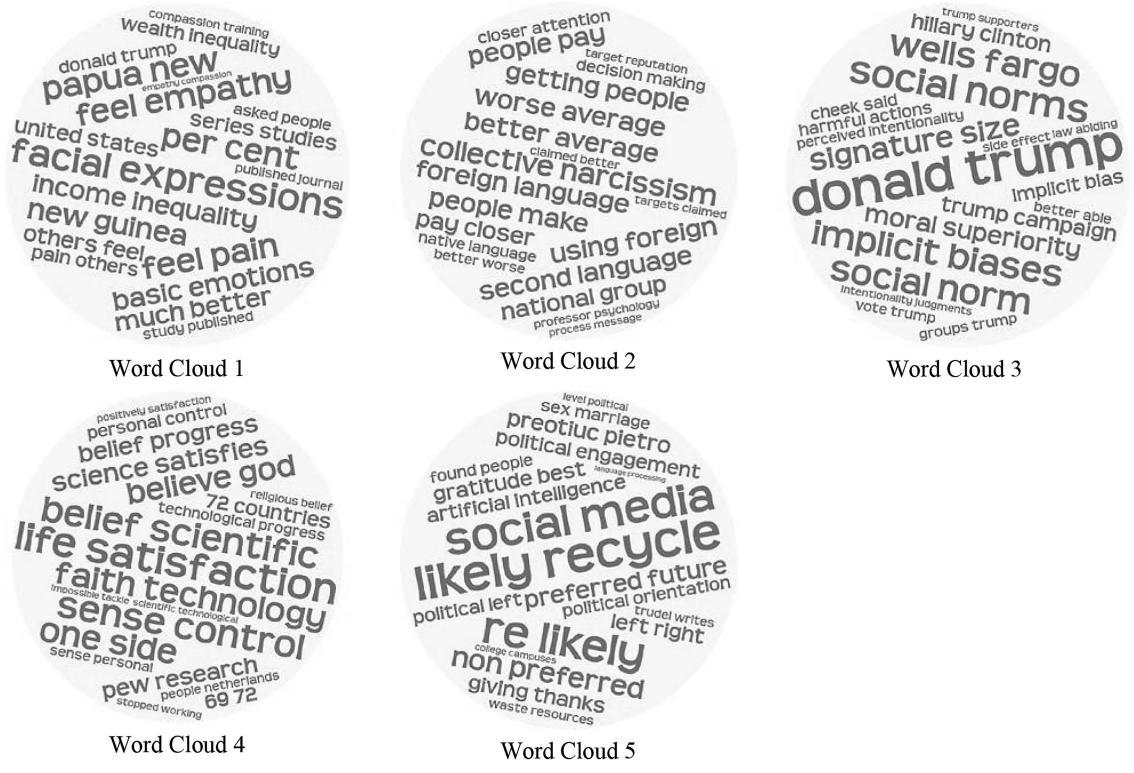

FIGURE 10.3 Top Five Topics Identified Based on Posts of the SPSP Page on Facebook

Note: The size of each word does not represent the relative weight in each topic.

methods, such as an aspect-based summarization model (Blair-goldensohn et al., 2008) to discover attributes, whereas others have used (semi) automatic methods with the same goal. For example, Hu and Liu (2004) used association mining in a combination of pre-identified adjectives with known positive/negative orientations to identify frequent (vs. infrequent) attributes: i.e., how likely are people to talk about those aspects? Other researchers have proposed the use of clustering to extract attributes in a hierarchical manner (Gamon et al., 2005) and the use of nouns to improve the clustering results for attributes identification (Farhadloo & Rolland, 2013).

In the process of identifying sentiments, researchers have mainly used a close-vocabulary approach to reveal the polarity of opinions of text fragments (Andreevskaia & Bergler, 2006; Esuli & Sebastiani, 2006; Hu & Liu, 2004; Subasic & Huettner, 2001; Wiebe, 2000). The close-vocabulary approach involves the use of a list of words (pre-identified terms) as a priori to examine the sentiment, and the presence of such words determines the sentiment polarity. The use of dictionaries words/terms is not limited to supervised learning but is also found in unsupervised learning. Turney (2002) has introduced an unsupervised technique that examines the number of occurrence and co-occurrences between two pre-identified terms and words found via the web search engine. For example, a term that frequently appears with the term "excellent" (a pre-identified positive term) is considered as positive whereas another term that often appears with the term "poor" (a pre-identified negative term) is considered as negative. Whereas a group of researchers identifies the sentiments by measuring the frequencies of specific words/terms (i.e., a regression problem), other researchers consider the sentiment identification as a classification problem (i.e., the presence/absence of features). Different classification techniques have been introduced to identify sentiments (Gamon et al., 2005; Lakkaraju, Bhattacharyya, Bhattacharya, & Merugu, 2011; Moghaddam & Ester, 2012; Pang, Lee, & Vaithyanathan, 2002), and the reliability of these techniques depends on the quality of the features revealed in the process. Hence, recent work has attempted to develop new computing techniques and algorithms, such as a score representation of positivity, negativity, and neutrality as new features (Farhadloo & Rolland, 2013), and a hierarchical deep learning framework (Lakkaraju, Socher, & Manning, 2014).

In addition to the close-vocabulary approach, a topic modeling, which attempts to identify attributes and sentiments simultaneously, is also frequently adopted for the analysis of sentiments. Topic modeling uses probabilistic methods to discover aspects and their associated sentiments at the same time. Topic modeling algorithms can distinguish between attribute-topics and sentiment-topics and determine the probability distribution of each term within a particular topic. One of the main advantages of such topic models as hierarchical Bayesian networks is that they do not require labeled training data and find the topics by analysis of the original collection of documents. As explained in the previous section, Latent Dirichlet Allocation (LDA) assumes the presence of a mixture of topics in each

document (Blei et al., 2003). In the case of sentiment analysis, when individuals talk about an attribute of an object, they are likely to use different terms. Likewise, individuals tend to use various terms to indicate a particular sentiment of that attribute. For instance, "excellent," "fabulous," and "extraordinary" are used to suggest a higher level of positive sentiments among individuals. Therefore, each sentiment-level can be considered a topic in topic modeling (see Brody, 2010; Farhadloo, Patterson, & Rolland, 2016 for further details).

Detection of Spam

Detecting spam within social media is a classic problem and is useful in many areas, including consumer, health, political, and social psychology. Although email spam is relatively easy to identify using unigrams or bigrams as input features for machine learning models, spam in social networks can take different forms (e.g., advertising spam, opinion spam, and deceptive opinion spam) and is therefore challenging. In the context of reviews, messages that do not include any opinions, but instead market products/services, are considered as advertising spam or duplicate spam. The detection of this type of spam is relatively easy (Jindal & Liu, 2008). Deceptive opinion spam is defined as "fictitious opinions that have been deliberately written to sound authentic" (Ott, Choi, Cardie, & Hancock, 2011). Some companies hire large numbers of users to post fake, and sometimes malicious, reviews or posts (Wang, Wang, Zhai, & Han, 2011). The detection of this type of spam is more challenging and requires data-driven models to pinpoint anomalous user behaviors (Lim, Nguyen, Jindal, Liu, & Lauw, 2010).

To detect opinion spam in the text, available methods include obtaining basic semantics features (e.g., n-grams) and identifying advanced semantic features (e.g., topics model). Character-level n-grams can be developed to effectively deal with the mistakes, typos, and errors in spelling that are quite common in social media but difficult to detect. Previous research has shown that using these character-level n-grams as features can improve the classification of news articles (Cavnar, Trenkle, & Mi, 1994). Others have demonstrated an 80% accuracy by using unigrams as simple features to identify individuals' race and ethnicity (Mohammady & Culotta, 2014).

In the area of detecting opinion spam, Ott et al. (2011) found that n-gram based text categorization best identified the opinion spam whereas a combined classifier with both n-grams and psycholinguistic deception features, i.e., terms obtained from the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) performed only slightly better than the former method. Furthermore, n-grams are used within language models for spam detection. Language models are probability distributions over units, where units can be anything from words, phrases, n-grams, or characters. In categorization tasks including text classification, a common method is to develop a language model for each category. For example, messages may be labeled as spam or ham (not spam) by

developing a spam language model and a ham language model. A product review may then be analyzed to determine the probability that it was generated from the spam or the ham language model (Sun, Morales, & Yan, 2013). This approach for using language models assumes that the text in the different categories uses the same words or phrases ("click," "here," "online," "cheap," etc.) or shares features that can be classified with appropriate models. This line of work has achieved successful spam detection, with a nearly 90% accuracy in spam detection (Ott et al., 2011). Nonetheless, recent research has found that a devious adversary can synthesize faked reviews by using similar data-driven methods (Sun et al., 2013; Tran, Hornbeck, Ha-Thuc, Cremer, & Srinivasan, 2011).

More sophisticated methods such as topic modeling may be more successful in detecting content because no specific words are predetermined in the process. As explained above, a topic model includes a number of topics in which each topic corresponds to a different distribution of words. Therefore, it is widely used to infer latent variables of words that frequently co-occur in a collection. Topic models have been applied to a host of problems, including TopicSpam, a topic modeling approach to identify deceptive opinion spam (Li, Cardie, & Li, 2013). However, topic modeling assumes a bag-of-word (BOW) representation that disregards the word order in the text and requires a sizable corpus to discover meaningful and interpretable topics (for alternate methods, see J. Chang, Gerrish, Wang, & Blei, 2009).

Using Social Media to Obtain Inferences

Predictive and Explanatory Models

An emerging field has used social media data to investigate public health challenges such as influenza infections and sexually transmitted infections, including HIV, chlamydia, and gonorrhea (Chan et al., 2018; Ireland et al., 2015; Young, Rivers, & Lewis, 2014). These studies have either developed a predictive model or an explanatory model. A predictive model is often bottom-up, open vocabulary without predetermined features, whereas an explanatory model is often top-down or closed-vocabulary with pre-established dictionaries of terms/phrases (Pennebaker, Mehl, & Niederhoffer, 2003). However, previous research has demonstrated the use of a closed-vocabulary approach for predicting influenza outbreaks (Santos & Matos, 2014; Signorini et al., 2011) and investigating links with HIV prevalence (Ireland, Schwartz et al., 2015; Young, Rivers, & Lewis, 2014). Potential challenges of closed-vocabulary methods include people's reluctance to discuss stigmatized conditions (e.g., HIV) or behaviors (e.g., drug use) online. Another limitation is that social media communications are informal and constantly evolving as a function of users' needs, culture, and idiosyncrasies (Gouws, Metzler, Cai, Hovy, & Rey, 2011).

An open-vocabulary approach can be used for prediction and explanation. Two major available methods differ in the degree to which they use predetermined terms to limit the collection of tweets: (a) a partial method, that is, including only tweets with pre-established dictionaries of terms/words, and (b) a full method, that is, including tweets without filtering by dictionaries. The partial method is likely to obtain more interpretable (explanatory) latent factors, whereas the later one can identify predictive factors relevant to users' needs, culture, and idiosyncrasy. As each method has its strengths and weaknesses, a mixed method is optimal for maximizing the predictability while improving the interpretability of latent factors that are identified on Twitter. For example, we used a Twitter application program interface (API; Garden Hose) to obtain about 10% random sample of all tweets in 2009–2010 and a Twitter streaming API to obtain approximately 1% of its publicly available stream in 2011–2012. We used the time metadata to exclude tweets not originating from U.S. time zones, and combined users' profile location with each tweet's precise coordinates to map tweets to U.S. counties. At the same time, we obtained the available county-level data on HIV prevalence and new diagnoses from the Centers for Disease Control and Prevention and AIDSVu (http://aidsvu.org/).

We first carried out an extensive search of research articles, news reports, as well as public health and slang dictionaries to identify a list of relevant terms and phrases of HIV and sexually transmitted infections (STIs). We identified 15 sources from various research teams in psychology and language processing, and together with the public health experts from the Health and Social Media Group at the University of Illinois at Urbana-Champaign, we devised nine categories that are related to HIV/STIs, including (a) HIV including treatment, (b) HIV and STI prevention, (c) drugs and alcohol, (d) other STIs, (e) sex, (f) men who have sex with men, (g) full-service sex work, (h) sexual violence and abuse, and (i) runaway youth. We collected words and phrases for these categories by incorporating prior dictionaries about sex and risky behaviors (Ireland et al., 2015), by using topic-specific glossaries (e.g., Drugs.com, 2013: HIV prevention measures), and by referring to slang databases (e.g., Urban Dictionary). The dictionaries contain 510 words.

Analytical Procedures and Results

Three methods of the open-vocabulary approach were assessed, and the major difference among these methods is the way for which tweets are prepared. For the partial method, we used the pre-established HIV/STIs dictionaries to filter out tweets that did not include one of the terms/words. For the full method, we included all tweets into the analyses. For the mixed method, we used the word embedding techniques to develop a lexicon of HIV and then included the lexicon as a prior in the machine learning model. Altogether, we had three sets of tweets,

and each was converted into a matrix of the token count. We then used the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to identify a model and to automatically discover topics (i.e., latent factors) in a collection of documents. We examined the distributions over words in each document and identified two hundred topics, then using the extremely randomized tree regressor method (Geurts, Ernst, & Wehenkel, 2006) to rank topics that associated with the new HIV diagnoses rates. We learned three topic models and evaluated the performance of each model by obtaining the topics probabilities of the 2012 tweets based on the word distributions and using the topic coefficients to compute the predicted 2012 new HIV diagnoses rate for each county. We then correlated the predicted 2012 HIV rates with the observed ones reported by the CDC to compare the performance.

Table 10.4 presents two model-fit indicators of models for three methods. By definition, a model with more predictive latent factors should yield a higher correlation and a lower mean squared error with the observed outcomes than the other models. As shown in Table 10.4, the proposed mixed method had the highest correlation coefficient and the lowest mean squared error among three methods. The results of different ethnicity representations were consistent with each other, indicating that the mixed method is likely to identify factors that explain the largest amount of variance in HIV prevalence rates. Apart from the numeric indicators, we also compared topics that were identified by different methods. The top three topics were selected from areas with higher and lower ethnic-minority representation. In general, the partial method identified latent factors with more words/terms about sex and drugs compared to other methods and revealed both norms about specific risk behaviors and general risk-taking notions. The partial method is likely more appropriate for detecting the presence of specific risk behaviors whereas the full and mixed methods are likely important for researchers to identify broader norms or perceptions linking to HIV risks in the communities.

TABLE 10.4 Results of Model-Fit Analyses Among Three Methods of the Open-Vocabulary Approach

Models	*df*	*Partial Method*[c]		*Full Method*[d]		*Mixed Method*[e]	
		r	*MSE*	*r*	*MSE*	*r*	*MSE*
Ethnicity representation 1[a]	2,596	.37***	0.95	.41***	0.83	.47***	0.76
Ethnicity representation 2[b]	2,768	.29***	0.94	.46***	0.78	.51***	0.72

Note: a = percentages of black population; b = percentages of white population; c = model-fit analyses based on tweets with filtering; d = model-fit analyses based on tweets without filtering; e = model-fit analyses based on including the HIV lexicon as prior; df = degree of freedom; *r* = correlation coefficients; MSE = mean squared errors.

*** < .001.

Ongoing Challenges and Concluding Notes

As a whole, social media data characterize of high spatial resolution (i.e., with an extensive coverage of geographical areas), the location information of individual users and of their contents becomes an important non-semantic feature for researchers to address questions of differences in areas (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Adomavicius & Tuzhilin, 2015; Kalampokis, Tambouris, & Tarabanis, 2013; Mohammady & Culotta, 2014). For example, Mohammady and Culotta (2014) developed a model to predict each Twitter user's ethnicity/race based on the ethnicity/race makeup of tweets that clustered by county. Recent work has also combined geo-mapping techniques with the analysis of social media data to detect terrorism and predict presidential elections (Cody, Reagan, Dodds, & Danforth, 2016; Cohen, Johansson, Kaati, & Mork, 2014). Although location identification is a key, the geo-mapping/geo-tagging of social media data at the user- and message-levels is far from simple (Eisenstein, O'Connor, Smith, & Xing, 2010; Han, Cook, & Baldwin, 2014). Given the growing concern with online privacy and cyberstalking (A. L. Young & Quan-Haase, 2009), less than 2% of social media users enable the GPS functionality (Ireland, Schwartz et al., 2015), and about 26% American teenagers fake their online information, including name, age, or location (Madden et al., 2013).

The limitation and sparseness of location information on social media have become a driving force in geo-mapping research, and different methods have been proposed to identify users locations (Cheng et al., 2010; Eisenstein et al., 2010; Schwartz, Eichstaedt, Kern, Dziurzynski, Agrawal et al., 2013). Schwartz et al. (2013) have proposed a rule-based mapping method, which uses information about the location and coordinates available in the metadata to map each post/message to a county. This method relies on either the coordinates information attached to a tweet/Facebook post (latitude, longitude) or the free-response location information in the users' profile on social media. As reported in recent studies, about 15%–20% of tweets could be mapped to U.S. counties. The percentage depends on the selection/inclusion criteria of tweets (Chan et al., 2018; Eichstaed et al., 2015; Ireland et al., 2016) Another group of scientists has suggested text-based geo-mapping that uses users' time zones, the number of followers/friends, and/or text messages for location prediction (Cheng et al., 2010; Eisenstein et al., 2010; Roller, Speriosu, Rallapalli, Wing, & Baldridge, 2012). Previous studies have demonstrated that text messages alone with neural network models can predict users' locations, from fine-grained coordinates to regions such as states (Cha, Gwon, & Kung, 2015; Han et al., 2014; Liu & Inkpen, 2015). The state-of-the-art performance is about 42% accuracy for the states prediction (Cha et al., 2015) and 50% for coordinate prediction, with a tolerance of about 161 km (Wing & Baldridge, 2014). The performance of such text-based geo-mapping techniques is subject to several factors, including the choice of activation functions, the number

of neurons per layer, initialization and regularization affects performance on predicting the actual geographical user coordinates, and classifying users per state or region (Morales et al., n.d.). Further work is required until this line of research can be used for "neural geotagging."

Another challenge with language identification of social media data is code-switching, which is the interchanging of different words in different languages in text messages. Recent work has used neural networks models, a popular classifier which automatically creates higher order representations of the input features for language identification (J. C. Chang & Lin, 2014). The code-switching makes it particularly challenging for tasks such as sentiment analysis, which typically assumes a single language and narrative (Vilares, Alonso, & Gómez-Rodríguez, 2016).

Social media is a unique data source that is worth exploring. Researchers can analyze a wide range of social media data, from demographic information, personal attributes, and location information, to various forms of messages, to determine characteristics of populations, investigate beliefs and attitudes, and ultimately understand behaviors. The widespread use of social media renders social media analysis more generalizable than results produced through conventional self-report methods with convenience samples. Furthermore, individuals and populations that are inherently difficult to reach due to lack of representation in academic settings may be more easily studied through social media analysis. The power and reach of social media analysis makes it a staunch ally to the contemporary researchers in social psychology and its allied sciences.

APPENDIX 10.1

Sample python codes of the topic modeling analysis:

```
### Import packages
from glob import glob
from sklearn.feature_extraction.text import
CountVectorizer
from sklearn.decomposition import
LatentDirichletAllocation
from nltk.corpus import stopwords
### Get the social media data file
text_data = glob('facebook_data/*.txt')
### Convert the data into bigrams and remove stopwords
cv = CountVectorizer(input='filename', ngram_range=(2,
2), stop_words=stopwords.words('english'))
### Transform the vocabularies into a matrix
X = cv.fit_transform(text_data)
### Performe the LDA topic modeling
lda = LatentDirichletAllocation(n_topics=15, max_
iter=100, random_state=42)
model = lda.fit_transform(X)
### Create a function to print out the outputs
def print_top_words(model, feature_names, n_top_words):
for topic_idx, topic in enumerate(model.components_):
print("Topic #%d:" % topic_idx, "
".join([feature_names[i]
```

```
for i in topic.argsort()[:-n_top_words–1:-1]]))
print()
### Print the topics with the first 20 words
feature_names = cv.get_feature_names()
print_top_words(lda, feature_names, 20)
```

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011). *Predicting flu trends using Twitter data*. 2011 IEEE conference on computer communications workshops. Shangai, China: IEEE, pp. 702–707. Retrieved from http://doi.org/10.1109/INFCOMW.2011.5928903

Adomavicius, G., & Tuzhilin, A. (2015). Context-aware recommender systems. *Recommender Systems Handbook* (2nd ed., pp. 191–226). Retrieved from http://doi.org/10.1007/978-1-4899-7637-6_6

Alvarez-Melis, D., & Saveski, M. (2016). *Topic modeling in Twitter: Aggregating tweets by conversations*. The Tenth International AAAI Conference on Web and Social Media, (Icwsm), pp. 519–522. Cologne, Germany: Association for the Advancement of Artificial Intelligence. Retrieved from evernote:///view/779439927/s24/8594e3b8-85b2-4f4c-8fc8-1f4b55e818cb/8594e3b8-85b2-4f4c-8fc8-1f4b55e818cb/

Anderson, L. (2013). HIV prevention. Retrieved June 11, 2018, from https://www.drugs.com/aids-preventative.html

Andreevskaia, A., & Bergler, S. (2006). *Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*. Annual Meeting of The European Chapter of The Association of Computational Linguistics (T. 6, pp. 209–216). Association for Computational Linguistics: Trento, Italy. Retrieved from http://doi.org/10.1.1.60.8316

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (T. abs/1003.5, pp. 492–499). Toronto, Canada: IEEE. Retrieved from http://doi.org/10.1109/WI-IAT.2010.63

Baldwin, T., Kim, Y-B., de Marneffe, M. C., Ritter, A., Han, B., & Xu, W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *The ACL 2015 Workshop on Noisy User-generated Text*, (pp. 126–135). Beijing, China: Association for Computational Linguistics. Retrieved from https://noisy-text.github.io/

Beckley, R. (2015). *Bekli: A simple approach to Twitter text normalization*. The 53rd Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2015) (p. 82). Beijing, China: Association for Computational Linguistics

Bennett, J., & Lanning, S. (2007). The Netflix prize. *KDD Cup and Workshop*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.6998

Bilton, N. (2012). *Disruptions: Innovations snuffed out by Craiglist*. 2017 m. sausio 30 d. Retrieved from https://bits.blogs.nytimes.com/2012/07/29/when-craigslist-blocks-innovations-disruptions/?_r=0

Blair-goldensohn, S., Neylon, T., Hannan, K., Reis, G. A., Mcdonald, R., & Reynar, J. (2008). *Building a sentiment summarizer for local service reviews*. Proceedings of the WWW2008 Workshop: NLP in the Information Explosion Era (NLPIX). Beijing, China: Association for Computational Machinery. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.182.4520

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. Retrieved from http://doi.org/10.1162/jmlr.2003.3.4-5.993

Boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. Retrieved from http://doi.org/10.1111/j.1083-6101.2007.00393.x

Brody, S. (2010, June). An unsupervised aspect-sentiment model for online reviews. *Computational Linguistics*, 804–812. Retrieved from www.aclweb.org/anthology/N10-1122

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating gender on Twitter*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1301–1309). Edinburg: Association for Computational Linguistics.

Cavnar, W. B., Trenkle, J. M., & Mi, A. A. (1994). *N-gram-based text categorization*. The 3rd Annual Symposium on Document Analysis and Information Retrieval (pp. 161–175). Las Vegas: Information Science Research Institute. Retrieved from http://doi.org/10.1.1.53.9367

Cha, M., Gwon, Y., & Kung, H. T. (2015). *Twitter geolocation and regional classification via sparse coding*. The 9th International Conference on Web and Social Media (ICWSM) (pp. 1–4). Oxford: Association for the Advancement of Artificial.

Chan, M. S., Lohmann, S., Morale, A., Zhai, C., Ungar, L. H., Holtgrave, D. R., & Albarracín, D. (2018). An Online Risk Index for the cross-sectional prediction of new HIV, chlamydia, and gonorrhea diagnoses across U.S. counties and across years. *AIDS and Behavior*. http://doi.org/https://doi.org/10.1007/s10461-018-2046-0

Chang, J. C., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296. Retrieved from http://doi.org/10.1.1.100.1089

Chang, J. C., & Lin, C. C. (2014). Recurrent-neural-network for language detection on Twitter code-switching corpus. *CoRR*, *1412.4314*.

Chang, K., Samdani, R., & Dan Roth. (2013). *A constrained latent variable model for coreference resolution*. The 2013 Conference on Empirical Methods on Natural Language Processing (EMNLP). Seattle: Association for Computational Linguistics.

Chang, S., Chen, Y., Yip, P., Lee, W., Hagihara, A., & Gunnell, D. (2014). Regional changes in charcoal-burning suicide rates in East/Southeast Asia from 1995 to 2011: A time trend analysis. *PLoS Medicine*, *11*(4), e1001622. Retrieved from http://dx.doi.org/10.1371/journal.pmed.1001622

Cheng, Z., Caverlee, J., & Lee, K. (2010). *You are where you tweet : A content-based approach to geo-locating Twitter users*. The 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada: Association for Computing Machinery, pp. 759–768. Retrieved from http://doi.org/10.1145/1871437.1871535

Cheung, Y. T. D., Chan, C. H. H., Lai, C-K. J., Chan, W. F. V., Wang, M. P., Li, H. C. W., . . . Lam, T-H. (2015). Using WhatsApp and Facebook online social groups for smoking relapse prevention for recent quitters:A pilot pragmatic cluster randomized controlled trial. *Journal of Medical Internet Research*, *17*(10). Retrieved from http://doi.org/10.2196/jmir.4829

Chu, K-H., Unger, J. B., Allem, J-P., Pattarroyo, M., Soto, D., Cruz, T. B., . . . Yang, C. C. (2015). Diffusion of messages from an electronic cigarette brand to potential users through Twitter. *PLoS ONE*, *10*(12), e0145387. Retrieved from http://doi.org/10.1371/journal.pone.0145387

Cody, E. M., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2016). Public opinion polling with Twitter. *Physics and Society*. Retrieved from https://arxiv.org/abs/1608.02024

Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246–256. Retrieved from http://doi.org/10.1080/09546553.2014.849948

Conger, K. (2016). *LinkedIn sues anonymous data scrapers*. 2016 m. gruodžio 20 d. Retrieved from https://techcrunch.com/2016/08/15/linkedin-sues-scrapers/

Consonni, M., & Anselmi, L. (2015). ECJ rules on screen-scraping of Ryanair's database. *E-Commerce Law and Policy*, *17*(2). Retrieved from www.orsingher.com/pdf/ECLP-15-02.pdf

Curini, L., Iacus, S., & Canova, L. (2015). Measuring idiosyncratic happiness through the analysis of Twitter: An application to the Italian case. *Social Indicators Research*, *121*(2), 525–542. Retrieved from http://dx.doi.org/10.1007/s11205-014-0646-2

De Souza, I. M., & Ferris, S. P. (2015). Social media marketing in luxury retail. *International Journal of Online Marketing*, *5*(2), 18–36. Retrieved from http://doi.org/10.4018/IJOM.2015040102

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). *Social media update 2014*. 2015 m. kovo 31 d. Retrieved from www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf

Eichstaed, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. Retrieved from http://doi.org/10.1177/0956797614557867

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). *A latent variable model for geographic lexical variation*. The 2010 Conference on Empirical Methods in Natural Language Processing. Massachusetts: Association for Computational Linguistics, pp. 1277–1287.

Esuli, A., & Sebastiani, F. (2006). *SENTIWORDNET: A publicly available lexical resource for opinion mining*. The 5th Conference on Language Resources and Evaluation. Genoa, Italy: European Language Resources Association, pp. 417–422. Retrieved from http://doi.org/10.1.1.61.7217

Farhadloo, M., Winneg, K., Chan, M. S., Jamieson, K. H., & Albarracín, D. (2018). Associations of topics of discussion on Twitter with survey measures of attitudes, knowledge, and behaviors related to Zika: Probabilistic study in the United States. *JMIR Public Health and Surveillance*, *4*(1), e16. Retrieved from http://doi.org/10.2196/publichealth.8186

Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems*, *90*, 1–11. Retrieved from http://doi.org/10.1016/j.dss.2016.06.010

Farhadloo, M., & Rolland, E. (2013). *Multi-class sentiment analysis with clustering and score representation*. 2013 IEEE 13th International Conference on Data Mining Workshops. Dallas: IEEE, pp. 904–912. Retrieved from http://doi.org/10.1109/ICDMW.2013.63

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public Interest*, *13*(1), 3–66. Retrieved from http://doi.org/10.1177/1529100612436522

Gamon, M., Gamon, M., Aue, A., Corston-Oliver, S., Corston-Oliver, S., . . . Ringger, E. (2005). Pulse: Mining customer opinions from free text. *Lecture Notes in Computer Science*, *3646*, 121–132. Retrieved from http://doi.org/10.1007/11552253_12

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*(1), 115–125. Retrieved from http://doi.org/10.1016/j.dss.2014.02.003

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.7485&rep=rep1&type=pdf

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. *Technometrics*. Retrieved from http://doi.org/10.2307/1271145

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2011). *Part-of-speech tagging for Twitter: Annotation, features, and experiments*. The 49th Annual Meeting of the Association for Computational Linguistics: Shortpapers. Portland: Association for Computational Linguistics, pp. 42–47. Retrieved from http://doi.org/10.1.1.206.3224

Gouws, S., Metzler, D., Cai, C., Hovy, E., & Rey, M. (2011). *Contextual bearing on linguistic variation in social media*. The Workshop of Language in Social Media (LSM 2011). Oregon: Association for Computational Linguistics, pp. 20–29.

Greenwood, S., Perrin, A., & Duggan, M. (2016). *Social media update 2016*. 2016 m. lapkričio 12 d. Retrieved from www.pewinternet.org/2016/11/11/social-media-update-2016/#fn-17239-1

Gui, L., Zhou, Y., Xu, R., He, Y., & Lu, Q. (2017). Learning representations from heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems*. Retrieved from http://dx.doi.org/10.1016/j.knosys.2017.02.030

Han, B., Cook, P., & Baldwin, T. (2014). Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, *49*, 451–500. Retrieved from http://doi.org/10.1613/jair.4200

Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., . . . Balter, S. (2014). *Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013*. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6320a1.htm

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. Retrieved from http://doi.org/10.1017/S0140525X0999152X

Hofmann, T. (1999). *Probabilistic latent semantic indexing*. The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999). Berkeley: Association for Computational Linguistics, pp. 50–57. Retrieved from http://doi.org/10.1145/312624.312649

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. The 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 04, T. 4. Seattle: Association for Computing Machinery, p. 168. Retrieved from http://doi.org/10.1145/1014052.1014073

Huang, G. C., Soto, D., Fujimoto, K., & Valente, T. W. (2014). The interplay of friendship networks and social networking sites: Longitudinal analysis of selection and influence effects on adolescent smoking and alcohol use. *American Journal of Public Health*, *104*(8), e51–e59. Retrieved from http://search.proquest.com/docview/1549549180?accountid=9851

Huang, J., Kornfield, R., & Emery, S. L. (2016). 100 million views of electronic cigarette YouTube videos and counting: Quantification, content evaluation, and engagement levels of videos. *Journal of Medical Internet Research*, *18*(3). Retrieved from http://dx.doi.org/10.2196/jmir.4265

Ireland, M. E., & Iserman, M. (2018). *Lusi lab development dictionaries*. Retrieved June 11, 2018, from https://www.depts.ttu.edu/psy/lusi/resources.php

Ireland, M. E., Chen, Q., Schwartz, H. A., Ungar, L. H., & Albarracin, D. (2015). Action tweets linked to reduced county-level HIV prevalence in the United States: Online messages and structural determinants. *AIDS and Behavior*. Retrieved from http://doi.org/10.1007/s10461-015-1252-2

Ireland, M. E., Chen, Q., Schwartz, H. A., Ungar, L. H., & Albarracín, D. (2016). Action tweets linked to reduced county-level HIV prevalence in the united states: Online messages and structural determinants. *AIDS and Behavior*, *20*(6), 1256–1264. Retrieved from http://doi.org/10.1007/s10461-015-1252-2

Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L. H., & Albarracín, D. (2015). Future-oriented tweets predict lower county-level HIV prevalence in the United States. *Health Psychology*, *34*(Supplement), 1252–1260. Retrieved from http://doi.org/10.1037/hea0000279

Jelenchick, L. A., Eickhoff, J. C., & Moreno, M. A. (2013). "Facebook depression?" Social networking site use and depression in older adolescents. *Journal of Adolescent Health*, T. 52. Retrieved from www.sciencedirect.com/science/article/pii/S1054139X12002091

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. *The International Conference on Web Search and Web Data Mining 2008*, pp. 219–230. Retrieved from http://doi.org/10.1145/1341531.1341560

Johnston, K., Tanner, M., Lalla, N., & Kawalski, D. (2013). Social capital: The benefit of Facebook 'friends'. *Behaviour & Information Technology*, *32*(1), 24–36. Retrieved from http://dx.doi.org/10.1080/0144929X.2010.550063

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, *23*(39), 544–559. Retrieved from http://doi.org/10.1108/IntR-06-2012-0114

Kaufmann, M. (2010). *Syntactic normalization of Twitter messages*. International Conference on Natural Language Processing, T. 2, pp. 1–7. Kharagpur, India: Macmillan Publishers. Retrieved from www.cs.uccs.edu/%7B~%7Dkalita/work/reu/REUFinalPapers2010/Kaufmann.pdf

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., & Smith, N. A. (2014). *A dependency parser for Tweets*. The Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: Association for Computational Linguistics, pp. 1001–1012. Retrieved from http://doi.org/10.3115/v1/D14-1108

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. Retrieved from http://doi.org/10.1073/pnas.1218772110

Kostygina, G., Tran, H., & Emery, S. (2016). *Follow even if you don't smoke: The amount and themes of cigarillo and marijuana co-use content on Instagram*. The APHA 2016 Annual Meeting & Expo. Denver: American Public Health Association

Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., & Merugu, S. (2011). *Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments*. The 2011 SIAM International Conference on Data Mining. Arizona: American Statistical Association, pp. 498–509.

Lakkaraju, H., Socher, R., & Manning, C. D. (2014). *Aspect specific sentiment analysis using hierarchical deep learning*. NIPS 2014 Workshop on Deep Neural Networks and Representation Learning. Montreal, Canada: Neural Information Processing System Foundation, pp. 1–9.

Lawrence, B., & Perrigot, R. (2015). Influence of organizational form and customer type on online customer satisfaction ratings. *Journal of Small Business Management*, *53*(Supplement 1), 58–74. Retrieved from http://dx.doi.org/10.1111/jsbm.12184

Li, J., Cardie, C., & Li, S. (2013). *TopicSpam: A topic-model-based approach for spam detection*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, pp. 217–221.

Lim, E-P., Nguyen, V-A., Jindal, N., Liu, B., & Lauw, H. W. (2010, April 2016). *Detecting product review spammers using rating behaviors*. The 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada: Association for Computing Machinery, 939–948. Retrieved from http://doi.org/10.1145/1871437.1871557

Lin, L. yi, Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., . . . Primack, B. A. (2016). Association between social media use and depression among U.S. young adults. *Depression and Anxiety*, *33*(4), 323–331. Retrieved from http://doi.org/10.1002/da.22466

LinkedIn. (2016). *Prohibition of scraping software*. 2016 m. gruodžio 15 d. Retrieved from www.linkedin.com/help/linkedin/answer/56347/prohibition-of-scraping-software?lang=en

Liu, J., & Inkpen, D. (2015). *Estimating user location in social media with stacked denoising auto-encoders*. Proceedings of NAACL-HLT 2015. Denver: Association for Computational Linguistics, pp. 201–210.

Lönnqvist, J-E., & Itkonen, J. V. A. (2014). It's all about Extraversion: Why Facebook friend count doesn't count towards well-being. *Journal of Research in Personality*, *53*, 64–67. Retrieved from http://dx.doi.org/10.1016/j.jrp.2014.08.009

Madden, M., Lenhart, A., Cortesi, S., Gasser, U., Duggan, M., Smith, A., & Beaton, M. (2013). *Teens, social media, and privacy*. Retrieved from www.lateledipenelope.it/public/52dff2e35b812.pdf

Mangukiya, P. (2016, gegužės 26). Social media by the numbers. *The Huffington Post*. Retrieved from www.huffingtonpost.com/piyush-mangukiya/social-media-by-the-numbe_b_9757926.html

Moghaddam, S., & Ester, M. (2012). *On the design of IDA models for aspect-based opinion mining*. The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012). Association for Computing Machinery, p. 803. Maui, Hawaii. Retrieved from http://doi.org/10.1145/2396761.2396863

Mohammady, E., & Culotta, A. (2014). *Using county demographics to infer attributes of Twitter users*. The Joint Workshop on Social Dynamics and Personal Attributes in Social Media. Baltimore: Association for Computational Linguistics, pp. 7–16. Retrieved from http://acl2014.org/acl2014/W14-27/W14-27-2014.pdf%7B#%7Dpage=19%7B%25%7D5Cn; www.aclweb.org/anthology/W/W14/W14-2702.pdf

Moreno, M. A., Ton, A., Selkie, E., & Evans, Y. (2016). Secret society 123: Understanding the language of self-harm on Instagram. *Journal of Adolescent Health*, *58*(1), 78–84. Retrieved from http://dx.doi.org/10.1016/j.jadohealth.2015.09.015

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). *Finding deceptive opinion spam by any stretch of the imagination*. The 49th Annua Meeting of the Association for Computational Linguistics. Portland: Association for Computational Linguistics, p. 11. Retrieved from http://arxiv.org/abs/1107.4557

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques*. The 2002 Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), T. 10. Philadelphia: Association for Computational Linguistics, pp. 79–86. Retrieved from http://doi.org/10.3115/1118693.1118704

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC. Austin, TX. Retrieved from liwc.net.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. Retrieved from http://doi.org/10.1146/annurev.psych.54.101601.145041

Popescu, A-M., & Etzioni, O. (2005). *Extracting product features and opinions from reviews*. The conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005), pp. 339–346. Vancouver, British Columbia, Canada: Association for Computational Linguistics. Retrieved from http://doi.org/10.1007/978-1-84628-754-1_2

Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). *Hierarchical Bayesian models for latent attribute detection in social media*. The Fifth International AAAI Conference on Weblogs and Social Media, T. 11. Barcelona, Spain: Association for the Advancement of Artificial Intelligence, pp. 598–601. Retrieved from www.cs.jhu.edu/%7B~%7Dmpaul/files/2011.icwsm.nigeria.pdf%7B%25%7D5Cn; www.cs.jhu.edu/%7B~%7Ddelip/icwsm.pdf

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011*). Named entity recognition in Tweets: An experimental study*. The 2011 Conference on Empirical Methods in Natural Language Processing. Edinburg, United Kingdom: Association for Computational Linguistics, pp. 1524–1534. Retrieved from http://doi.org/10.1075/li.30.1.03nad

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., & Baldridge, J. (2012). *Supervised text-based geolocation using language models on an adaptive grid*. The 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1500–1510. Jeju Island, Korea: Association for Computational Linguistics.

Santos, J. C., & Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology & Medical Modelling*, *11*(Supplement 1), S6. Retrieved from http://doi.org/10.1186/1742-4682-11-S1-S6

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G. J., . . . Ungar, L. (2013). *Characterizing geographic variation in well-being using tweets*. The Seventh International AAAI Conference on Weblogs and Social Media (ICWSM-13). Cambridge: Association for the Advancement of Artificial Intelligence, pp. 583–591. Retrieved from http://doi.org/papers3://publication/uuid/43E3E88F-EFDC-4F9C-85AC-C60B4B8C8BCA

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), e73791. Retrieved from http://doi.org/10.1371/journal.pone.0073791

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 78–94. Retrieved from http://doi.org/10.1177/0002716215569197

Seidenberg, A., Jo, C., Ribisl, K., Lee, J., Butchting, F., Kim, Y., & Emery, S. (2017). A national study of social media, television, radio, and internet usage of adults by sexual orientation and smoking status: Implications for campaign design. *International Journal of Environmental Research and Public Health*, *14*(4), 450. Retrieved from http://doi.org/10.3390/ijerph14040450

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, *6*(5), e19467. Retrieved from http://doi.org/10.1371/journal.pone.0019467

Smyser, J. D. (2013). *Health communication and social media: A case study of the California Tobacco Control Program's "Toxic Butts" campaign. ProQuest Dissertations and Theses*. Retrieved from http://sfx.scholarsportal.info/guelph/docview/1494825145?accountid=

11233%7B%25%7D5Cn; http://sfx.scholarsportal.info/guelph?url%7B_%7Dver=Z39.88-2004%7B&%7Drft%7B_%7Dval%7B_%7Dfmt=info:ofi/fmt:kev:mtx:dissertation%7B&%7Dgenre=dissertations+%7B%25%7D26+these

Statista. (2010). Number of social media users worldwide from 2010 to 2020 (in billions). 2017 m. gegužės 21 d. Retrieved from www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Stevens, P., Carlson, L. M., & Hinman, J. M. (2004). An analysis of tobacco industry marketing to lesbian, gay, bisexual, and transgender (LGBT) populations: Strategies for mainstream tobacco control and preventio. *Health Promotion Practice*, *5*(3), 129–134. Retrieved from http://doi.org/10.1177/1524839904264617

Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., . . . Su, Z. (2008). *Hidden sentiment association in Chinese web opinion mining*. The 17th International Conference on World Wide Web. Beijing, China: Association for Computing Machinery, pp. 959–968. Retrieved from http://doi.org/10.1145/1367497.1367627

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, *9*(4), 483–496. Retrieved from http://doi.org/10.1109/91.940962

Sun, H., Morales, A., & Yan, X. (2013). Synthetic review spamming and defense. *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2013)*, p. 1088. New York, NY: Association for Computing Machinery. Retrieved from http://doi.org/10.1145/2487575.2487688

Teitler, J. O., Reichman, N. E., & Sprachman, S. (2003). Costs and benefits of improving response rates for a hard-to-reach population. *Public Opinion Quarterly*, *67*(1), 126–138. Retrieved from http://doi.org/10.1086/346011

Thangarajan, N., Green, N., Gupta, A., Little, S., & Weibel, N. (2015). *Analyzing social media to characterize local HIV at-risk populations*. The Conference on Wireless Health (WH 2015), pp. 1–). New York, NY: Association for Computing Machinery. Retrieved from http://doi.org/10.1145/2811780.2811923

Tran, H., Hornbeck, T., Ha-Thuc, V., Cremer, J., & Srinivasan, P. (2011). *Spam detection in online classified advertisements*. The 2011 Joint International Workshop on Information Credibility on the Web and Adversarial Information Retrieval on the Web Workshop on Web Quality, pp. 35–41. Hyderabad, India: Association for Computing Machinery. Retrieved from http://doi.org/10.1145/1964114.1964122

Turney, P. D. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. The 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 417–424). New York: Association for Computational Linguistics. Retrieved from http://doi.org/10.3115/1073083.1073153

Upadhyay, S., Gupta, N., Christodoulopoulos, C., & Roth, D. (2016). *Revisiting the evaluation for cross document event coreference*. The 26th International Conference on Computational Linguistics. Osaka, Japan: International Committee on Computational Linguistics.

Vilares, D., Alonso, M. A., & Gómez-Rodrıguez, C. (2016). *En-es-cs: An English-Spanish code-switching Twitter corpus for multilingual sentiment analysis*. The Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association, pp. 4149–4153.

Wang, H., Wang, C., Zhai, C., & Han, J. (2011). *Learning online discussion structures by conditional random fields*. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China: Association for Computing Machinery, pp. 435–444. Retrieved from http://doi.org/10.1145/2009916.2009976

Ward, J. (2016). What are you doing on Tinder? Impression management on a matchmaking mobile app. *Information, Communication & Society*, 1–16. Retrieved from http://doi.org/10.1080/1369118X.2016.1252412

Wiebe, J. M. (2000). *Learning subjective adjectives from corpora*. The National Conference on Artificial Intelligence (pp. 735–741). Austin: Association for the Advancement of Artificial Intelligence. Retrieved from http://doi.org/http://portal.acm.org/citation.cfm?id=721121&dl=ACM&coll=&CFID=15151515&CFTOKEN=6184618

Wing, B., & Baldridge, J. (2014). *Hierarchical discriminative classification for text-based geolocation*. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 336–348). Doha, Qatar: Association for Computational Linguistics. Retrieved from http://anthology.aclweb.org/D/D14/D14-1039.pdf

Wu, Y., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. Retrieved from http://doi.org/10.1073/pnas.1418680112

Young, A. L., & Quan-Haase, A. (2009). *Information revelation and internet privacy concerns on social network sites*. The Fourth International Conference on Communities and Technologies (C&T 2009) (p. 265). New York: Association for Computing Machinery. Retrieved from http://doi.org/10.1145/1556460.1556499

Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine*, *63*, 112–115. Retrieved from http://doi.org/10.1016/j.ypmed.2014.01.024

Yuan, E. J., Feng, M., & Danowski, J. A. (2013). "Privacy" in semantic networks on Chinese social media: The case of Sina Weibo. *Journal of Communication*, *63*(6), 1011–1031. Retrieved from http://dx.doi.org/10.1111/jcom.12058

Zamal, F. Al, Liu, W., & Ruths, D. (2011). *Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors*. The Sixth International AAAI Conference on Weblogs and Social Media (pp. 387–390). Dublin, Ireland: Association for the Advancement of Artificial Intelligence.

Zhan, Y., Liu, R., Li, Q., Leischow, S. J., & Zeng, D. D. (2017). Identifying topics for e-cigarette user-generated contents: A case study from multiple social media platforms. *Journal of Medical Internet Research*, *19*(1). Retrieved from http://doi.org/10.2196/jmir.5780

Zide, J., Elman, B., & Shahani-Denning, C. (2014). LinkedIn and recruitment: How profiles differ across occupations. *Employee Relations*, *36*(5), 583–604. Retrieved from http://doi.org/10.1108/ER-07-2013-0086