2017

Deception Detection and Rumor Debunking for Social Media

Victoria L. Rubin
Western University, vrubin@uwo.ca

DECEPTION DETECTION AND RUMOR DEBUNKING FOR SOCIAL MEDIA

Citation: Rubin, V. L. (2017). Deception Detection and Rumor Debunking for Social Media. In Sloan, L. & Quan-Haase, A. (Eds.) (2017) *The SAGE Handbook of Social Media Research Methods*, London: SAGE. https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-social-media-research-methods/book245370

Abstract

The main premise of this chapter is that the time is ripe for more extensive research and development of social media tools that filter out intentionally deceptive information such as deceptive memes, rumors and hoaxes, fake news or other fake posts, tweets and fraudulent profiles. Social media users' awareness of intentional manipulation of online content appears to be relatively low, while the reliance on unverified information (often obtained from strangers) is at an all-time high. I argue there is need for content verification, systematic fact-checking and filtering of social media streams. This literature survey provides a background for understanding current automated deception detection research, rumor debunking, and broader content verification methodologies, suggests a path towards hybrid technologies, and explains why the development and adoption of such tools might still be a significant challenge.

Biographical Details

Victoria L. Rubin is an Associate Professor at the Faculty of Information and Media Studies and the Director of the Language and Information Technologies Research Lab (LiT.RL) at the University of Western Ontario. She specializes in information retrieval and natural language processing techniques that enable analyses of texts to identify, extract, and organize structured knowledge. She studies complex human information behaviors that are, at least partly, expressed through language such as deception, uncertainty, credibility, and emotions. Her research on Deception Detection has been published in recent core workshops on the topic and prominent information science conferences, as well as the Journal of the Association for Information Science and Technology. Her 2015-2018 project entitled *Digital Deception Detection: Identifying Deliberate Misinformation in Online News* is funded by the Government of Canada Social Sciences and Humanities Research Council (SSHRC) Insight Grant. For further information, see http://victoriarubin.fims.uwo.ca/.

Key words: deception, fake news, fraudulent profiles, rumors, hoaxes; deception detection, social media information manipulation, content verification, fact-checking, rumor debunking, credibility assessment, trust.

INTRODUCTION

The goal of this chapter is to introduce readers to automated deception detection research, with a cursory look at the roots of the field in pre-social media data types. My intent is to draw attention to existing analytical methodologies and pose the question of their applicability to the context of social media. This chapter is divided into five parts as follows.

The Problem Statement section sets the stage for why deception detection methods are needed in the social media context by calling attention to the pervasiveness of social media and its potential role in manipulating user perceptions.

In the Background segment I define deception and talk briefly about the roots and more contemporary forms of deception research. I provide the necessary background for what is currently known about people's overall abilities to spot lies and what constitutes predictive cues to tell the liars apart from truth tellers.

The Methodological Solutions part outlines some principles by which deception can be identified outside of social media context. I elaborate on predictive linguistic cues and methods used to identify deception. I follow up with an overview of several Online Tools that tackle the problem of deception detection and argue that more research and development of deception detection tools is needed, taking into account the specificity of each type and format of the social media stream.

In Broader Content Verification I consider several important related concepts: rumors, credibility, subjectivity, opinions, and sentiment, evaluating appropriate techniques and method for identifying these phenomena.

Open Research and Development Problems are discussed in terms of the needed methodologies for three most recent social media phenomena: potential fraud on collaborative networking sites, pervasiveness of clickbaiting, and astroturfing by bots in social media. I briefly explain how each phenomenon relates to deception detection efforts, identifying these areas as most up-to-date niches requiring research and development.

I conclude that social media requires content verification analysis with a combination of previously known approaches for deception detection, as well as novel techniques for debunking rumors, credibility assessment, factivity analysis and opinion mining. Hybrid approaches may include text analytics with machine learning for deception detection, network analysis for rumor debunking and should incorporate world knowledge databases to fully take advantage of the linguistic, interpersonal, and contextual awareness.

PROBLEM STATEMENT: DECEPTION IN SOCIAL MEDIA CONTEXT

Although social media is difficult to define precisely as a phenomenon (McCay-Peet and Quan-Haase, Forthcoming 2016), most popular social networking and microblogging sites such as Facebook, Twitter, and LinkedIn include the function of online community building, personal messaging, and information sharing (Guadagno and Wingate, 2014). Digital news environments,

I argue, are now becoming increasingly social, if not in the way they have been written, at least in the way they are accessed, disseminated, promoted, and shared.

The boundary between mainstream media news and user generated content is slowly blurring (Chen et al., 2015b). Kang, Höllerer and O'Donovan (2015) observe that microblogging services have recently transformed from "online journal or peer-communication platforms" to "powerful online information sources operating at a global scale in every aspect of society, largely due to the advance of mobile technologies. Today's technology enables instant posting and sharing of text and/or multimedia content, allowing people on-location at an event or incident to serve as news reporters". They cite studies of traditional media journalist practices showing that journalists rely heavily on social media for their information, and report about 54% of all U.S. journalists use microblogs to collect information and to report their stories (Kang et al., 2015).

It is also common for lay news readers to receive news by via their social peers on networks like Facebook and Twitter. Thus, it is reasonable to consider tools and methodologies from fuller-form communication formats such as news in "pre-social media era" as a starting point for deception detection and rumor debunking methodologies for the newer platforms and formats, whether they are shorter or longer, hashed or not, video-based or image-based. In this chapter I focus primarily on text-based social media application, those that use text primarily as their medium of communication, while image-sharing (such as Flickr and Picassa) and video-sharing websites (such as YouTube and Vimeo) are left aside for separate consideration of potential ways to manipulate non-textual content.

A 2013 PEW research report (Holcomb et al., 2013) showed an increase in the number of users that keyword-search blogs as opposed to the traditional content streams. It means that a larger portion of information comes from complete strangers rather than from known or trusted sources (Kang et al., 2015). "With the massive growth of text-based communication, the potential for people to deceive through computer-mediated communication has also grown and such deception can have disastrous results" (Fuller et al., 2009, p. 8392).

The majority of social media contributors presumably communicate their messages to the best of their knowledge, abilities, and understanding of the situation at hand. Their messages primarily match their own beliefs. Social media streams are awash in biased, unreliable, unverified subjective messages, as well as ads and solicitations. Most social media consumers are typically aware of the subjective nature of social media streams, as well as the typical promotional intentions to attract online traffic and revenue. What is rarer is the realization that there are (albeit rarer) instances of posts, tweets, links, and so forth, that are designed to create false impressions or conclusions. Creating false beliefs in the social media consumers' minds can be achieved though disseminating outright lies, fake news, rumors or hoaxes, especially when the message appears to come from "a friend" or another in-group member. Spam and phishing attacks in e-mail messages are more recognizable now that most users have experience receiving and filtering them, while the issue of information manipulation via social media is still poorly understood and rarely atop of users' minds. Malevolent intentions manifest themselves in inter-personal deception and can be damaging in person-to-person communication.

Social media users often hold a general presumption of goodwill in social media communication. Morris et al. (2012) found that, for instance, Twitter users "are poor judges of truthfulness based on content alone, and instead are influenced by heuristics such as user name when making

credibility assessments". Some social media users may sacrifice caution for the sake of convenience, which may result in them being vulnerable to those who intend to deceive by disseminating false rumors or hoaxes in an effort to alter users' decision making and patterns of behavior (beyond incentivizing to purchase via pushed advertising).

There are well-documented instances of deceptive, unconfirmed, and unverified tweets being picked up by main-stream media, giving them undeserving weight and credibility. In October 2008, three years prior to Steve Jobs' death, a citizen journalist posted a report falsely stating that Jobs had suffered a heart attack and had been rushed to a hospital. The original deliberate misinformation was quickly "re-tweeted" disregarding the fact that it originated from CNN's iReport.com which allows "unedited, unfiltered" posts. Although the erroneous information was later corrected, the "news" of Jobs' alleged health crisis spread fast, causing confusion and uncertainty, and resulting in a rapid fluctuation of his company's stock on that day (per CBC Radio "And the Winner Is", 31 March 2012). This is just one, albeit very public, example of deceptive information being mistaken for authentic reporting, and it demonstrates the very significant negative consequences such errors can create. Earlier examples of companies "struck by phony press releases" include the fiber optic manufacturer, Emulex, and Aastrom Biosciences (Mintz, 2002). Research further cites evidence of false tweets recently discovered in U.S. Senate campaigns, in reporting of the Iranian election protests, and in the coverage of unfolding natural disasters such as the Chilean earthquake (Morris et al., 2012). Social search tools (such as Bing Social Search [bing.com/social] and Social Seaking [socialseeking.com]) can also amplify undesirable memes, and while some false reporting is relatively harmless (such as celebrity deaths), "increased reliance on social media for actionable news items (*Should I vote for candidate X? Should I donate to victims of disaster Y?*)", makes credibility a nontrivial concern" (Morris et al., 2012).

A 2015 Pew report documents that "about six-in-ten online Millennials (61%) report getting political news on Facebook in a given week, a much larger percentage than turn to any other news source, according to a new Pew Research Center analysis." About the same ratio of Baby Boomers [born 1946-1964] (60%) rely by contrast on local TV sources for political news (Pew 2015 Report by Mitchell and Page, 2015). Considering that younger users tend to rely on social media to inform themselves on breaking news, political issues, local and international events, the potential for harm from being intentionally misinformed over the internet is evident.

Researchers and developers for social media platforms are starting to consider methods and tools for filtering out intentionally manipulative messages and prompting unsuspecting users to fact-check. The context of social media is unique, diverse in formats, and relatively new, but lying and deceiving has been at play in other forms of human communication for ages. The next section overviews the roots of deception studies and the contemporary interpretation of the phenomenon in deception research. I also outline how deception can be detected in texts, specifically, with the use of state-of-the-art text analytics. Though no "bullet-proof" mechanism currently exists to screen out all memes, hoaxes, rumors, and other kinds of malevolent manipulative messages, it is perhaps time to consider what methodologies can be harnessed from the previous years in deception detection research, and how those methods can be successfully ported to the new context of social media.

BACKGROUND: DECEPTION AND TRUTH BIAS

Since the ancient times, the concepts of *truth*, *falsehood*, *lying*, and *deception* have been pondered over great thinkers, from the ancient Greek philosophers (Socrates, Plato, Aristotle) to central figures in modern Western philosophy (Emmanuel Kant, Ludwig Wittgenstein). Sissela Bok writes in her analysis of morality (1989) that "lying has always posed a moral problem"; for instance, Aristotle believed falsehood in itself to be "mean and culpable", and Kant regarded truthfulness as an "unconditional duty which holds in all circumstances".

In 21st century *truthfulness* and *honesty* remain essential for successful communication, while deception is still largely frowned upon and widely condemned (Walczyk et al., 2008). Deception (with or without computer mediation) violates the cooperative principle for successful communication, expressed as a failure to observe at least one of the four maxims, as postulated by a philosopher of language, Paul Grice (1975) (1975): say what you believe to be true (Maxim of Quality), do not say more than needed (Maxim of Quantity), stay on the topic (Maxim of Relevance), and do not be vague (Maxim of Manner) (Rubin, 2010b).

Recent Inter-Personal Psychology and Computer-Mediated Communication studies define *deception* as an intentional and knowing attempt on the part of the sender of the message to create a false belief or false conclusion in the mind of the receiver of the message (e.g., Buller and Burgoon, 1996, Zhou et al., 2004). The definition typically excludes self-deception and unintentional errors since in those exceptions the senders' beliefs still match the intended communicated message. Lying is considered to be just one kind of deception – that of falsification – as opposed to other deceptive varieties such as omission, equivocation, or concealment.

From Inter-Personal Psychology studies we also know that people are generally *truth-biased*, or more predisposed towards to veracity than deception. "*The truth bias* is the presumption of truth in interpersonal interactions and the tendency to judge an interpersonal message as truthful rather than deceptive, irrespective of the actual truth of the message. Communicators are initially assumed to be truthful, and this assumption is possibly revised only if something in the situation evokes suspicion" (Van Swol, 2014). "Numerous studies have found that independent of actual message veracity, individuals are much more likely to ascribe truth to other's messages than deceit" (Levine et al., 1999). There is no reason to presume that a subset of the general population wouldn't exhibit similar truth-bias tendencies.

Truth bias is also one of the potential explanations for why people are so inept at distinguishing truths from deception. Humans are notoriously poor lie detectors even when they are alerted to the possibility of being lied to (Vrij, 2004, Vrij, 2000, Vrij et al., 2012). A widely cited source that conducted a meta-analytical review of over 100 experiments with over 1,000 participants (DePaulo et al., 1997), concludes that on average people are able to distinguish a lie from a truthful statement with a mean accuracy rate of 54%, slightly above chance (Rubin and Conroy, 2012).

On the other hand, current theories of deceptive communicative behaviors suggest that deceivers communicate in qualitatively different ways from truth-tellers. Stable differences are found in behaviors of liars versus truth-tellers, especially evident in the verbal aspects of behavior (Ali and

Levine, 2008). Liars are said to be identified by their words – not by what they say but by how they say it (Newman et al., 2003). There have been efforts to compile, test, and cluster predictive cues for deceptive messages in order to translate those findings into text analytical tools for detecting lies, primarily in longer forms of Computer Mediated Communication such as e-mail.

METHODOLOGICAL SOLUTIONS: DECEPTION DETECTION WITH LINGUISTIC PREDICTORS

Deception Detection researchers generally agree that it is possible to detect deception based on linguistic cues. Several successful studies on deception detection have demonstrated the effectiveness of linguistic cue identification, as the language of truth-tellers is known to differ from that of deceivers (e.g., Bachenko et al., 2008, Larcker and Zakolyukina, 2012).

Though there is no clear consensus on reliable predictors of deception, deceptive cues can be identified in texts, extracted and clustered conceptually, for instance, to represent diversity, complexity, specificity, and non-immediacy of the analyzed texts. For instance, (Zhou et al., 2004) developed Text-based Asynchronous Compute-Mediated Communication (TA-CMC) and reviewed five main systems developed for the analysis of the deception detection in textual communication: Criteria-Based Content Analysis (CBCA), Reality Monitoring (RM), Scientific Content Analysis (SCAN), Verbal Immediacy (VI) and Interpersonal Deception Theory (IDT). Each of the systems developed criteria for classifying textual information either as deceptive or truthful and contributed towards creation of the list of 27 linguistic features in eight broad conceptual clusters, as shown in Figure 1 (Zhou et al 2004).

Figure 1. Summary of Zhou et al's (2004) Linguistic Features for Deception Detection.
Twenty seven linguistic-based features, amenable to automation, were grouped into nine linguistic constructs: quantity, complexity, uncertainty, nonimmediacy, expressivity, diversity, informality, specificity, and affect. All the linguistic features are defined in terms of their measurable dependent variables. (Redrawn from Zhou et al., 2004).

I. QUANTITY

1. Word: a written character or combination of characters representing a spoken word.
2. Verb: a word that characteristically is the grammatical center of a predicate & expresses an act, occurrence, or mode of being.
3. Noun phrase: a phrase formed by a noun, its modifiers & determiners.
4. Sentence: a word, clause, or phrase or a group of clauses or phrases forming a syntactic unit which expresses an assertion, a question, a command, a wish, an exclamation, or the performance of an action, which usually begins with a capital letter & concludes with appropriate end punctuation.

II. COMPLEXITY

5. Average number of clauses: $\frac{\text{total \# of clauses}}{\text{total \# of sentences}}$
6. Average sentence length: $\frac{\text{total \# of words}}{\text{total \# of sentences}}$
7. Average word length: $\frac{\text{total \# of characters}}{\text{total \# of words}}$
8. Average length of noun phrase: $\frac{\text{total \# of words in noun phrases}}{\text{total \# of noun phrases}}$
9. Pausality: $\frac{\text{total \# of punctuation marks}}{\text{total \# of sentences}}$

III. UNCERTAINTY

10. Modifiers: describes a word or makes the meaning of the word more specific. There are two parts of speech that are modifiers - adjectives & adverbs.
11. Modal verb: an auxiliary verb that is characteristically used with a verb of predication & expresses a modal modification.
12. Uncertainty: a word that indicates lack of sureness about someone or something.
13. Other reference: third person pronoun.

IV. NON-IMMEDIACY:

14. Passive voice: a form of the verb used when the subject is being acted upon rather than doing something.

IV. NON-IMMEDIACY (continued)

15. Objectification: an expression given to (as an abstract notion, feeling, or ideal) in a form that can be experienced by others & externalizes one's attitude.
16. Generalizing terms: refers to a person (or object) as a class of persons or objects that includes the person (or object).
17. Self-reference: first person singular pronoun.
18. Group reference: first person plural pronoun.

V. EXPRESSIVITY

19. Emotiveness: $\frac{\text{total \# of adjectives + total \# of adverbs}}{\text{total \# of nouns + total \# of verbs}}$

VI. DIVERSITY

20. Lexical diversity: $\frac{\text{total \# of different words or terms}}{\text{total \# of words or terms}}$, which is the percentage of unique words in all words.
21. Content word diversity: $\frac{\text{total \# of diff. content words}}{\text{total \# of content words}}$, where content words primarily express lexical meaning.
22. Redundancy: $\frac{\text{total \# of function words}}{\text{total \# of sentences}}$ where function words express primarily grammatical relationships.

VII. INFORMALITY

23. Typographical error ratio: $\frac{\text{total \# of misspelled words}}{\text{total \# of words}}$

VIII. SPECIFICITY

24. Spatio-temporal information: information about locations or the spatial arrangement of people and/or objects, or information about when the event happened or explicitly describes a sequence of events.
25. Perceptual information: indicates sensorial experiences such as sounds, smells, physical sensations & visual details.

IX. AFFECT

26. Positive affect: conscious subjective aspect of a positive emotion apart from bodily changes.
27. Negative affect: conscious subjective aspect of a negative emotion apart from bodily changes.

When implemented with standard classification algorithms (such as neural nets, decision trees, and logistic regression), such methods achieve 74% accuracy (Fuller et al., 2009). Existing psycholinguistic lexicons (e.g., LIWC by Pennebaker and Francis, 1999) have been adapted to perform binary text classifications for truthful versus deceptive opinions, with classifiers demonstrating a 70% average accuracy rate (Mihalcea and Strapparava, 2009).

Human judges, by a rough measure of comparison, achieved only 50 – 63% success rates in identifying deception, depending on what is considered deceptive on a seven-point scale truth-to-

deception continuum: the more extreme degrees of deception are more transparent to judges (Rubin and Conroy, 2011).

Deception Detection researchers also widely acknowledge a variation in linguistic cues as predictors across situations (Ali and Levine, 2008), across genres of communication, communicators (Burgoon et al., 2003) and cultures (Rubin, 2014). The main lesson we are learning is that the contexts in which deceptive communications occurs matter greatly. For example, in synchronous text-based communication, deceivers produced more total words, more sense-based words (e.g., seeing, touching), and used fewer self-oriented but more other-oriented pronouns (Hancock et al., 2007). Compared to truth-tellers, liars showed lower cognitive complexity and used more negative emotion words (Newman et al., 2003).

In conference calls of financiers, Larcker and Zakolyukina (2012)found deceptive statements to have more general knowledge references and extreme positive emotions, and also fewer self-references, extreme negative emotions, as well as certainty and hesitation words.

In police interrogations, Porter & Yuille (1996b) found three significantly reliable, verbal indicators of deception (based on Statement Validity Analysis techniques used in law enforcement for credibility assessments): amount of detail reported, coherence, and admissions of lack of memory.

In descriptions of mock theft experiments, Burgoon and colleagues (2003) found deceivers' messages in their text-based chats were briefer (i.e., lower on quantity of language), less complex in their choice of vocabulary and sentence structure, and lacked specificity or expressiveness.

Deception is prominently featured in several domains (e.g., politics, business, personal relations, science, journalism (Rubin, 2010a). The use of language changes under the influence of different situational factors, genre, register, speech community, text and discourse type (Crystal, 1969). Therefore, the verbal cues for deception detection across various knowledge domains and various formats of social media may differ, though the computational algorithms or broader concept (such as Zhou's clusters of diversity, complexity, specificity, and non-immediacy) may remain constant. When predictive linguistic cues are developed based on general linguistic knowledge (Höfer et al., 1996), linguistic cues could be portable to social media contexts (for instance, from e-mail to full-sentences forum posts). Nevertheless, if the subject areas are highly specialized, then when deciphering predictive cues, researchers should account for context specificity and format (Höfer et al., 1996, Porter and Yuille, 1996a, Köhnken and Steller, 1988, Steller, 1989). Table 1 summarizes various types of discourses or types of data that were addressed within various disciplines that study deceptive behaviors and their linguistic predictors. Notice that only a limited portion of data types can be found on social media (such as dating profiles and product and services reviews), while several non-social media types of discourse bare closer resemblance to each other (such as confessions and diary-style blogs).

DECEPTION RESEARCH DISCIPLINE	DATA TYPES
Inter-Personal Psychology, Computer-Mediated Communication	Elicited data Data generated with imaginary tasks Questionnaire data Interview data Case scenario discussions Observation data Lying games data Pre-existing messages (e-mails, diaries, etc.) Digitized records of any of the above Transcripts of oral interactions
Law Enforcement, Credibility Assessment, Police Work, Homeland Security	Court proceedings Police interrogations Credibility assessment transcripts Testimonies Pleas Alibi Court decisions
Deception Detection with Natural Language Processing and Machine Learning	Digital forms (texts, transcripts) of any of the above Crowdsourced data (e.g., Mechanical Turk) Crawled and harvested Web data Social media data in contexts of *fake product and service reviews* *fake dating profiles* *fudged online resumes* *fake social network profiles* *fake news* *spamming and phishing* *forged scientific work*

Table 1. Deception Research Disciplines and Associated Data Type Examples.
Various contemporary disciplines that study deception for the purpose of its detection are listed with corresponding typical data types that have been studied in the recent past. The data type distinctions are made based on the mode of obtaining data and the predominant discourse variety in the data type. Both columns are non-exhaustive.

How predictive cues of deception in microblogs (Twitter) may be different from more verbose formats (e-mails or conference call records) is yet to be studied. The social nature of the media can also provide other affordances that are typically inaccessible to face-to-face communication studies (such as past track-record, profiles, geolocation, and associated imagery) which could and should be matched against known truths or general world knowledge (encapsulated in such sources as Wikipedia and Wiktionaries). In other words, since context appears to be paramount to obtaining appropriate linguistic predictors of deceptive messages, contextual information should be intensely explored for social media deception detection. Past behaviors and profiles afford a more holistic interpretation of one's linguistic behavior and its correspondence to reality, since (ethical issues of surveillance, tracking and profiling aside) incongruities can be directly identified based on one's "footprints" in social networking and communication.

ONLINE DECEPTION DETECTION TOOLS

In the past several years, conceptual tools dealing with language accuracy, objectivity, factuality and fact-verification have increased in importance in various subject areas due to rising amounts of digital information and the number of its users. Journalism, online marketing, proofreading and politics are to name a few. For example, in politics, *Politifact* (although based on manual fact-checking) and *TruthGoggles* sort the true facts in politics, helping citizens to develop better understanding of politicians statements.

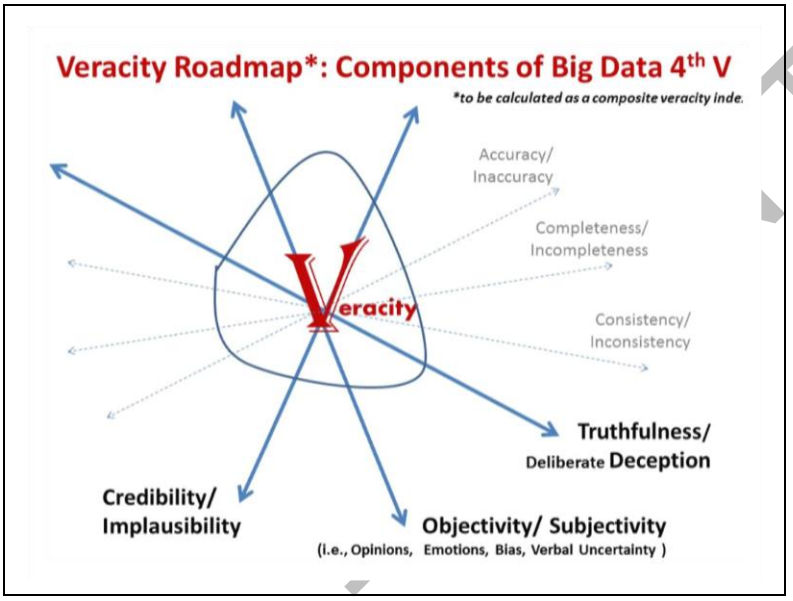
In proofreading, *Stylewriter* and *AftertheDeadline* help users to identify stylistic and linguistic problems related to their writings. These tools use not only linguistic cues to resolve expression uncertainty problems, but also establish the factuality of events and statements using experts' opinions and additional necessary sources. For an overview of related content annotation and automation efforts, see (Morante and Sporleder, 2012) and (Sauri and Pustejovsky, 2009, Sauri and Pustejovsky, 2012).

Building on years of Deception Detection research in Interpersonal Psychology, Communication Studies, and Law Enforcement, a cutting-edge technology is emerging from the fields of Natural Language Processing and Machine Learning. Spurred by demand from practitioners for stable, quick and accurate deception detection tools, scholars have begun to create software for deception detection. A limited number of automated (or partially automated) online deception detection tools became available for the public by around 2010, including those by Chandramouli and Subbalakshmi (2012), Ott et al. (2011), Moffit and Giboney (2012) (evaluated by Rubin and Vashchilko, 2012).

The majority of the text-based analysis software uses different types of linguistic cues. Some of the common linguistic cues are the same across all deception software types, whereas other linguistic cues are derived specifically for specialized topics to generate additional linguistic cues. The complete automation of deception detection in written communication is mostly based on the linguistic cues derived from the classes of words from the Linguistic Inquiry and Word Count (LWIC) (Pennebaker et al., 2001). The main idea of LWIC coding is text classification according to truth conditions. LWIC has been extensively employed to study deception detection (Vrij et al., 2007, Hancock et al., 2007, Mihalcea and Strapparava, 2009).

In 2014 Rubin and Lukoianova proposed that veracity should be considered as an important component of big data assessment, assuming that social media posts, tweets, reviews and other platform messages are a large component of big data (see Figure 2 for an explanation of the proposed veracity index calculation). Passing the deception detection test in Social Media can verify the source's intention to create a truthful impression in the readers' mind, supporting sources trustworthiness and credibility. On the other hand, failing the test immediately alerts the user to potential alternative motives and intentions and necessitates further fact verification (Lukoianova and Rubin, 2014).

Figure 2. Conceptualization of the Components of Big Data Veracity. Veracity – as the 4th V in addition to Volume, Velocity and Variety – portrayed across three primary orthogonal dimensions in the conceptual space – objectivity, truthfulness, credibility. The dimensions intersect in the center and the nebula represents a certain degree of variability within the phenomena that together constitute the big data veracity. Secondary dimensions of lesser concern in textual data are presented in dotted lines. The tree main components of veracity index are normalized to the (0,1) interval with 1 indicating maximum objectivity, truthfulness and credibility, and 0, otherwise. Then, the big data veracity index is calculated as an average of the three, assuming that each of the dimensions equally contributes to veracity establishment. The three primary dimensions reduce "noise" and potential errors in subsequent inferences from the textual big data due to minimization of bias, intentional misinformation, and implausibility.



As of early 2016, researchers declared that the field of automated detection as applied to "social media is a relatively new one. There have so far been only a handful of works that address this problem." (Vosoughi, 2015). Even if automated social media verification tools are on the market or in research and development, they are not particularly well known to general social media users. Nor have they received much attention in mainstream North American media coverage, or in the scientific community. The wealth of predictive linguistic cues knowledge has yet to be tested in the social media context. It is worth noting that other terminology may have been used to refer to deception detection, such as *veracity prediction* and *rumor debunking* or *rumor busting*, and those methodologies – as well as several other "close relatives" pertaining to content verification – will be explored in the next section.

BROADER CONTENT VERIFICATION: RUMORS, CREDIBILITY, AND OPINONS

There are several ways to look at the problem of social media content verification. Detection of deceptive messages based on what has been said (or linguistic cues) is only one part of the

problem. The broader context – in terms of positioning of the message sources in the network, their reputation, trustworthiness, credibility, expertise, as well as propensity for spreading rumors – should be taken into account. How accurate, well-informed and objective are the sources? Ideally, for decision-making, social media users should rely on truthful, accurate, and complete information from credible expert sources.

Rumors and Rumor Debunking

Social media often amplifies and disseminates word-of-mouth rumors by reaching wider audiences. Rumors should not be directly equated to deceptive messages, even though most people as well as experts in the rumor debunking research agree that rumors are harmful. Among undesirable responses to rumors Matthews (2013) lists defamation, protests, and destruction of properties, spread of fear, hate, or euphoria. Rumors on Twitter have been known to influence the stock market. "Perhaps, one of the most infamous cases is of the hacked AP account tweeting a rumor that Barack Obama had been injured in an explosion at the White House. The tweet caused the S&P to decline and wipe $130 Billion in stock value in a matter of seconds" (Liu et al., 2015).

The defining feature of a rumor is lack of verifiability at the moment of dissemination. Merriam Webster's Dictionary defines *a rumor* as "a statement or report current without known authority for its truth" or "talk or opinion widely disseminated with no discernible source" (Merriam-Webster Online Dictionary, 2016). Some dictionary definitions emphasize "the word of mouth" as the method of spreading hearsay (Thr Free Dictionary, 2016) disregarding how prevalent the spread of rumors can be over social networks. Though it is still the dawn of rumor detection studies, there have been further clarifications in a handful of current works which take into account social media reality. For instance, Vosoughi (2015) makes it clear that *a rumor* is "an unverified assertion that starts from one or more sources and spreads over time from node to node in a network." In his recent dissertation on the topic, he continues to explain the subtleties of the rumor spread on Twitter, how rumor is related to deception, and, most importantly, what it means to resolve a rumor algorithmically: "On Twitter, a rumor is a collection of tweets, all asserting the same unverified statement (however the tweets could be, and almost assuredly, are worded differently from each other), propagating through the communications network (in this case Twitter), in a multitude of cascades. A rumor can end in three ways: it can be resolved as either true (factual), false (non-factual) or remain unresolved. There are usually several rumors about the same topic, any number of which can be true or false. The resolution of one or more rumors automatically resolves all other rumors about the same topic. For example, take the number of perpetrators in the Boston Marathon bombings; there could be several rumors about this topic:

1. Only one person was responsible for this act.
2. This was the work of at least 2 or more people.
3. There are only 2 perpetrators.
4. It was at least a team of 5 that did this.

Once rumor number 3 was confirmed as true, it automatically resolved the other rumors as well. (In this case, rumors 1 and 4 resolved to be false and rumor 2 resolved to be true)" (Vosoughi, 2015).

Traditionally, rumors have been resolved with either by common sense judgements or with further investigations by professionals. There are several existing examples of rumor detection systems, some with real time algorithmic veracity prediction that is potentially faster than human verification by professionals. For instance, Liu and his colleagues from the Thompson Reuters R&D group (2015), observed the need to invent tools for journalists to verify rumors. They thus proposed a method to automatically debunk rumors on Twitter using social media. Figure 3 shows the types of features that the rumor debunking system considers in its real-time analysis.

Figure 3. Verification Feature for Rumor Debunking on Twitter (Liu et al., 2015).
The six proposed categories of verification features largely based on insights from journalists.

CATEGORY	FEATURE NAME
SOURCE CREDIBILITY	Is trusted/satirical news account Has trusted/satirical news url Profile has url from top domains Client application name
SOURCE IDENTITY	Profile has person name Profile has location Profile includes profession information
SOURCE DIVERSITY	Has multiple news/non-news urls after dedup Deduped tweets' text is dissimilar
SOURCE LOCATION & WITNESS	If tweet location matches event location If profile location matches event location Has witness phrases, i.e., “I see" and “I hear"
MSG. BELIEF	Is support, negation, question or neutrality
EVENT PROPAGATION	Event Topic Retweet, mention, hashtag h-index Max reply/retweet graph4 size/depth

Most existing algorithms for debunking rumors, however, follow Castillo, Mendoza, and Poblete’s work (Castillo et al., 2011, Mendoza et al., 2010) employing variations on data used and features extracted (Wu et al., 2015, Yang et al., 2012). Qazvinian and colleagues (2011) focus on rumor-related tweets to match certain regular expression of the keyword query and the users' believing behavior about those rumor-related tweets; both pieces of information are instrumental in isolating rumors. Mendoza and colleagues (2010) analyze user behavior through tweets during the Chilean earthquake that year: “they analyze users' retweeting topology network and the difference in the rumor diffusion pattern on Twitter environment than on traditional news platforms” (Yang et al., 2012). Moving away from Twitter, Yang and colleagues (2012) studied Sina Weibo, China's leading micro-blogging service provider that functions like a Facebook-Twitter hybrid. They collected and annotated a set of rumor-related microblogs based on the Weibo's rumor-busting service, as a result proposed extra. Figure 4 lists the features used for the Weibo *rumor buster*.

Figure 4. Rumor Busting Features for Sina Weibo Microblogs (Yang, 2012).
Grouped into five broad types (content-based, client-based, account-based, propagation-based, and location-based), rumor detection features were extracted on Sina Weibo, the Chinese leading micro-blogging platform, for the binary classification purposes (rumor or not). Each predictive feature is described in terms of its implementation.

Category	Features	Description
CONTENT	HAS MULTIMEDIA	Whether the microblog contains pictures, videos, or audios
	SENTIMENT	The numbers of positive and negative emoticons used in the microblog
	HAS URL	Whether the microblog includes a URL pointing to an external source
	TIME SPAN	The time interval between the time of posting and user registration
CLIENT	CLIENT PROGRAM USED	The type of client program used to post a microblog: web-client or mobile-client
ACCOUNT	IS VERIFIED	Whether the user's identity is verified by Sina Weibo
	HAS DESCRIPTION	Whether the user has personal descriptions
	GENDER OF USER	The user's gender
	USER AVATAR TYPE	Personal, organization, and others
	NUMBER OF FOLLOWERS	The number of user's followers
	NUMBER OF FRIENDS	The number of users who have a mutual following relationship with this user
	NUMBER OF MICROBLOGS POSTED	The number of microblogs posted by this user
	REGISTRATION TIME	The actual time of user registration
	USER NAME TYPE	Personal real name, organization name, and others
	REGISTERING PLACE	The location information taken at user's registration
LOCATION	EVENT LOCATION	The location where the event mentioned by rumor-related microblogs happened
PROPAGATION	IS RETWEETD	Whether the microblog is original or is a retweet of another microblog
	NUMBER OF COMMENTS	The number of comments on the microblog
	NUMBER OF RETWEETS	The number of retweets of the microblog

Credibility and Credibility Assessment

Credibility assessment tools have explored broader contextual profiles than deception detection and rumor debunking methods. The concept of credibility is intrinsically linked to believability, which is not necessarily equivalent to truthfulness or veracity but is rather a reflection of perceived truth.

Much has been written in Library and Information Science and Human-Computer Interaction on credibility assessment and a variety of checklist schemes to verify the credibility and stated cognitive authority of the information providers. See Rieh (2010) for a summary of the historical development of the credibility research in such fields as Psychology and Communication, and a recent overview of credibility typologies in LIS (e.g., source credibility, message credibility, and media credibility) and HCI (e.g., computer credibility: presumed credibility, reputed credibility, surface credibility, and experienced credibility.

The concept of *trust* is often used in everyday language and communication in making trustworthiness decisions. Hardin (Hardin, 2001) noticed a pervasive conceptual slippage that involves a misleading inference from the everyday use of trust: many ordinary-language statements about trust seem to conceive trust, at least partly, as a matter of behavior, rather than an expectation or a reliance. Trust, in Inter-Personal and Organizational Psychology, is seen as a positive expectation of a trusting entity regarding the behavior of the trustee (the trusted entity) in a context that entails risk to the trustor (e.g., Marsh and Dibben, 2003). Fogg and Tseng

(1999) firmly equate *credibility* to *believability* and *trust* to *dependability* (p. 41). *Content trust* is a trust judgment about a particular piece of information in a given context (Gil and Artz, 2006), e. g., any statement regarding upcoming or ongoing political upheaval. While an entity can be trusted on the whole, each particular piece of information provided by the entity may still be questioned.

In relation to information shared on social media, *trust* is an assured reliance on the character, ability, strength, or truth of trusted content (Merriam-Webster.com). In the Semantic Web literature, two types of trust are distinguished, one concerned with trust judgments about the providers of the information, and the other concerned with the nature of the information provided (Gil and Artz, 2006), e.g., a judgment about the US Government provided by the activists of the 99% movement.

Rieh (2010) also underscores the importance of *trustworthiness* and *expertise*, as the two widely recognized components of credibility, although according to her, they are not always perceived together. "An expert with the title of doctor or professor might have a reputation of being knowledgeable in a certain area but still might not be considered trustworthy for the tendency to unreliability or bias. A person may think of a friend as being honest and trustworthy in general, but the advice that the friend gives is not necessarily considered credible for the friend's lack of expertise" (Rieh, 2010, p. 1338).

Trustworthiness refers to the goodness or morality of the source and can be described with terms such as well-intentioned, truthful, or unbiased. Expertise refers to perceived knowledge of the source and can be described with terms such as knowledgeable, reputable, and competent (Tseng and Fogg, 1999).

Since the early 2000s credibility tools have proliferated in the form of varying measures for credibility predictions, computational models, and algorithms. In 2011 Castillo, Mendoza, and Poblete (2011), proposed an algorithm that predicts the credibility of an event based on a set of features of a given set of tweets: they analyzed tweets related to "trending topics" and use a binary supervised classification method from machine learning to place them into one of the two bins: credible or not credible. Kang, Höllerer, and O'Donovan (2015) identify and evaluate key factors that influence credibility perception on Twitter and Reddit (such as time spent posting or time spent reading posts of others). For their ground truth measure of the credibility of microblog data to achieve a "more stable" estimate of credibility, Sikdar and colleagues (2013) combine manually annotated scores with observed network statistics (such as retweets).

Rubin and Liddy's (2006) short influential work on modeling credibility of blogs set out a framework for assessing blog credibility, with 25 indicators outlined within four main categories: blogger expertise and offline identity disclosure; blogger trustworthiness and value system; information quality; and appeals and triggers of a personal nature (see Table 2). Weerkamp and de Rijke (2008) estimated several of the indicators proposed in Rubin and Liddy (2006) and integrated them into their retrieval approach, ultimately showing that combining credibility indicators significantly improves retrieval effectiveness.

1) Blogger's Expertise and Offline Identity Disclosure

a) Name and geographic location (connecting on-line and off-line identities)
b) Credentials
c) Affiliations (personal and institutional)
d) Blogrolls (both dynamic and static links)
e) Stated competencies
f) Mode of knowing ((observation, deduction, trusted sources, etc.)
g) Certainty level trends over time

Desired effect: knowledgeable, reputable, and competent blogger (Tseng and Fogg, 1999)

2) Blogger's Trustworthiness and Value System

a) Biases (stated or otherwise displayed priorities)
   e.g., *"I don't care much for political correctness; I do care for accuracy and honesty (what people actually do rather than what they believe or say)"*
b) Beliefs
c) Opinions
   e.g., *"I had never got the hang of academic writing. The personal voice on blogs appealed to me so much more"*
d) Honesty indicators
e) Preferences
f) Habits and behavioral patterns
g) Slogans

3) Information Quality

a) Completeness
b) Accuracy
c) Appropriateness
d) Timeliness
e) Information organization style (by categories, chronology, etc.)
f) Match to prior expectations
g) Match to information need
h) Use of rhetoric devices *beneficial* to blogger's credibility
   - projecting concerns for readers' viewpoints
   - expressing modesty
i) Use of rhetoric devices *detrimental* to blogger's credibility
   - having prior inaccuracies or errors
   - using artificially adorned figurative speech

Desired effect: complete, accurate, and appropriate information (Van House, 2004).

4) Appeals and Triggers of a Personal Nature

a) Aesthetic appeal (i.e., design layout, typography, and color schemes)
b) Literary appeal (i.e., writing style, wittiness, "coolness" factor)
c) Curiosity trigger
d) Memory trigger (i.e., shared experiences)
e) Personal connection (e.g., the source is an acquaintance or a competitor)
f) Match between information need and availability
g) Match to prior expectations
h) Personal connection (e.g., the source is an acquaintance or a competitor of the blog-reader)

"The Wild Card": information-seeker and information provider's interaction; hard to elicit and harvest automatically

Table 2. Blog Credibility Assessment Factors
(Redrawn from Rubin and Liddy (2006) with additional details added from the associated presentation).

Even though certain features have been proven to be beneficial for more accurate blog retrieval in early work on weblog credibility in information retrieval, subjectivity research, and sentiment analysis (Rubin and Liddy, 2006, Weerkamp and de Rijke, 2008), the research has not yet resonated with the rumor debunking community, probably due the isolation of the literatures or perhaps due to the differences between blogs and micro-blogs formats.

When analysing social media platforms and formats of interaction, these two components of credibility should be considered separately. In summary, two credibility components,

trustworthiness and expertise, are essential to making credibility (i.e., believability) judgments about trustworthiness (i.e., dependability) of sources and information on social media, regardless of whether such judgments are expressed lexically with a vocabulary of trust as being trustworthy (i.e., dependable) or credible (i.e., believable).

Subjectivity and Opinion Mining, or Sentiment Analysis

Some fields, such as Media Theory, differentiate objectivity from credibility, both of which have been part of traditional journalistic practices since 1950s, with credibility equated to believability (Johnson and Wiedenbeck, 2009). The main two reasons for using automation in deception detection are to increase objectivity by decreasing potential human bias in detecting deception (reliability of deception detection), and improve the speed in detecting deception (time processing of large amounts of text) (Hauch et al., 2012).

The concept of separating subjective judgments from objective became of great interest to Natural Language Processing researchers and gave rise to a very active area of sentiment analysis, or opinion mining, which concerns with analyzing written texts for people's attitudes, sentiments, and evaluations with text analytical techniques. "Rubin (2006b) traces the roots of subjectivity identification tools to the work of (Wiebe et al., 2001) who proposed one of the first annotation schemes to classify and identify subjective and objective statements in texts. Prior to this work on subjectivity, Rubin (2006b) continues, an NLP system needed to determine the structure of a text – normally at least enough to answer "Who did what to whom?" (Manning and Schütze, 1999). Since early 2000s the revised question was no longer just "Who did what to whom?" but also "Who thinks what about somebody doing what?" (Lukoianova and Rubin, 2014).

The majority of current text analytical tools operating on social media datasets are disproportionally focused on sentiment analysis or polarity of opinions (positive, negative, or neutral), while the issues of credibility and verifiability are addressed less vigorously. (For a comprehensive overview of the field of opinion-mining and/or sentiment analysis, see Pang and Lee (2008) and a more recent survey by Liu (2012) as well as the introductory article by Thelwall (Forthcoming, 2016) in this book which is specifically focused on sentiment analysis tools for social media. The work on identification of factuality or factivity in text-mining ((e.g., Sauri and Pustejovsky, 2009, Sauri and Pustejovsky, 2012, Morante and Sporleder, 2012) stems back to the idea that people exhibit various levels of certainty (or epistemic modality) in their speech, and that these levels are marked linguistically (e.g., "maybe", "perhaps" vs "probably" and "for sure") and can be identified with text analytical techniques (Rubin, 2006a, Rubin et al., 2006, Rubin et al., 2004). Text analysis for factuality and writer's certainty is more beneficial to enhance deception detection capabilities than currently acknowledged in the field. For instance, opinion mining should not disregard factivity, objectivity, and certainty in stated opinions, since lack of those properties in personal claims may render them useless and may skew aggregate analyses of social media data (such as product and services reviews).

Open Research and Development Problems

Outside of the previously discussed studies, there have been surprisingly few well-known efforts to verify information in social media feeds. Notable exceptions are studies of fake social network profiles (Kumar and Reddy, 2012), fake dating profiles (Toma and Hancock, 2012) and fake product reviews (Mukherjee et al., 2013), though the interactive social component may be less prominent in these studies as compared more mainstream micro-blogging platforms such a Twitter and Sina Weibo.

Three relatively recent social media phenomena call for further investigations: the rise of collaborative networking sites and their openness to potential fraud, pervasiveness of clickbaiting, and astroturfing by social bots to influence users. Each is discussed in turn here.

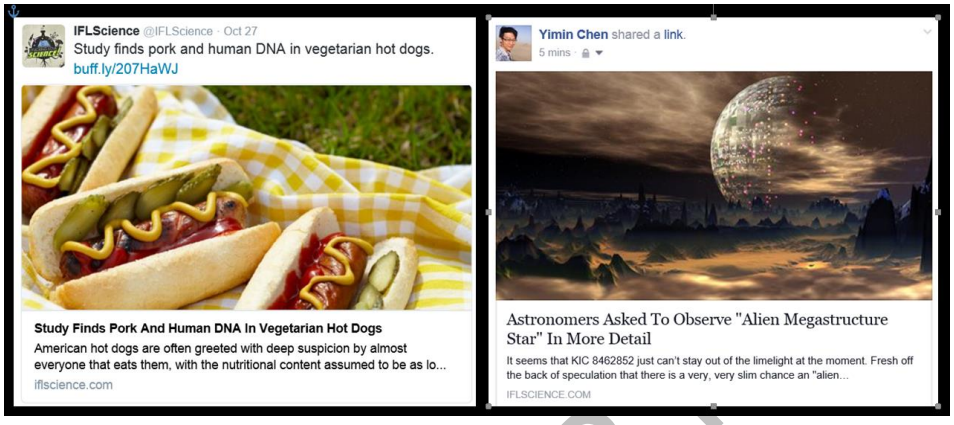
Fraud on Academic Collaborative and Networking Sites

Relatively new academic collaborative and networking platforms (such as ResearchGate, Academia.edu, Mendeley, or ORCID) are yet to be studied for potential content manipulation and fraud. To the best of my knowledge, no deception detection tools are yet available within these profession-based collaborative scholarly sharing systems. Inaccurate self-presentation or presentation of others on their behalf (with or without their knowledge) can have ramifications for perceptions of scholars' productivity when socially shared data is used for altmetrics (bibliometrics and webometrics combined) of scholarly output. For instance, Ortega (2015) firmly links social and usage metrics at the authors' level to the authors' productivity and treats such metrics as a proxy for research impact. The newly coined field of *altemtrics* has not yet considered the margins of errors related to fraud, as most of the collaborative platform data seem to be currently taken for its face value.

Clickbaiting

Another issue that received little attention thus far is the prevalence of "clickbait" in news streams (see Figure 5 for examples).

Figure 5. Examples of Clickbait via Twitter. (Chen, Conroy and Rubin, 2015 presentation). Two examples of *clickbaits* or online content that primarily aims to attract attention and encourage visitors to click on a link to a particular web page. The claims made in headlines of clickbaits and the associated imagery are often outrageous and/or misleading, as the reader finds out from reading the full message.



Clickbait refers to "content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page" ['clickbait,' n.d.] and has been implicated in the rapid spread of rumor and misinformation online. *Clickbaiting* can be identified through a consideration of the existence of certain linguistic patterns, such as the use of suspenseful language, unresolved pronouns, a reversal narrative style, forward referencing, image placement, reader's behavior and other important cues (Chen et al., 2015a).

Several social sharing platforms have standardized formats and visual presentation of delivery, regardless the source. Be it a satirical news piece from the Onion or a mainstream news piece from the New York Times, when "liked" and "shared" on Facebook or Twitter, the visual clues for potentially misleading information are minimal. The source's attribution is barely visible (see bottom of Figure 5). Tabloidization of news production and the shift towards digital content incentivizes the use of clickbait (Chen et al., 2015a), and it is yet unclear how skilled news readers on social media are in distinguishing this variety of content manipulation from legitimate news.

More work is necessary to distinguish fake news from authentic ones, and clickbaiting practices are just the tip of the iceberg. Other potential threats to veracity include such fake news as fraudulent journalistic reporting, hoaxes, and misleading satirical news taken at face value (Rubin et al., 2015).

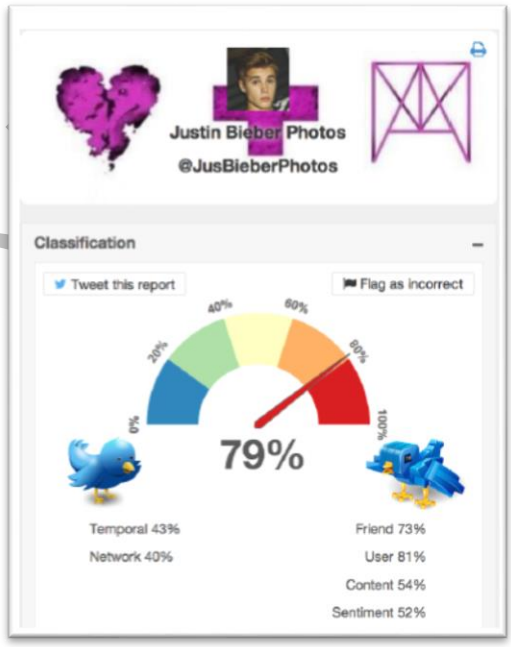
Astroturfing by Social Bots

Astroturfing is a recent phenomenon and, by a definition found in an off-beat dictionary, is an attempt "to create the impression of public support by paying people in the public to pretend to be supportive" (The Urban Dictionary, 2016). A new computerized form of such false support is slowly spreading on social media. Some social media platforms allow *sybil accounts* or *social bots* which rely on computer algorithms to imitate humans by automatically producing content

and interacting with other users. Such social bots pollute authentic content and spread misinformation by manipulating discussions, altering user popularity ratings, and "even perform[ing] terrorist propaganda and recruitment actions" (Davis et al., 2016).

Subrahmanian and colleagues (2016) identified three types of Twitter bots that engage in deceptive activities: 1) "*Spambots* spread spam on various topics"; 2) "*Paybots* illicitly make money. Some paybots copy tweet content from respected sources like @CNN but paste in micro-URLs that direct users to sites that pay the bot creator for directing traffic to the site; 3) "*Influence* bots try to influence Twitter conversations on a specific topic. For instance, some politicians have been accused of buying influence on social media". Subrahmanian and colleagues (2016) also notice that influence bots can "pose a clear danger to freedom of expression", citing examples of spread of radicalism, political disinformation and propaganda campaigns. The challenge has just been recently identified in the U.S. DARPA Social Media in Strategic Communications program competition to test the effectiveness of influence bot detection methods. Three most successful teams found machine learning techniques alone were insufficient because of lack of training data, but thought a semi-automated process that included machine learning was useful. Their feature set is reminiscent of a variety of features discussed in this chapter thus far. For instance, *BotOrNot*, a publicly-available service since May 2014 (see Figure 6), leverages *more than one thousand features* to evaluate the extent to which a Twitter account exhibits similarity to the known characteristics of social bots (Davis et al., 2016).

Figure 6. Dashboard-Style Interface of the *BotOrNot* System (Davis et al 2015). The system evaluates the extent to which a Twitter account exhibits similarity to the known characteristics of social bots.



The organizers of the DARPA challenge predict that as "bot developers are becoming increasingly sophisticated", "we can expect a proliferation of social media influence bots as advertisers, criminals, politicians, nation states, terrorists, and others try to influence populations" in the next few years. This trend necessitates the need for significant enhancements in the analytical tools that help analysts detect influence bots (Subrahmanian et al., 2016).

CONCLUSION

In conclusion, social media with its new mechanisms for interaction and information flow requires a variety of content verification mechanisms, perhaps in combination with previously known deception detection approaches as well as novel techniques for rumor debunking, credibility assessments, and opinion mining. When analyzing social media for potentially deceptive content, it is important to apply methods that consider not just what is being said, but also how the message is presented, by who, and in what format and context. The hybrid approach should include text analytics, network analysis and world knowledge database incorporation to fully take advantage of linguistic, interpersonal, and contextual awareness. This chapter is a call for further research in developing further, as well as modifying and applying existing deception detection methods and rumor debunking technologies towards various social media forms and formats.

REFERENCES

ALI, M. & LEVINE, T. 2008. The Language of Truthful and Deceptive Denials and Confessions. *Communication Reports,* 21**,** 82 - 91.

BACHENKO, J., FITZPATRICK, E. & SCHONWETTER, M. Verification and implementation of language-based deception indicators in civil and criminal narratives. Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008. Association for Computational Linguistics, 41-48.

BOK, S. 1989. *Lying: Moral choice in public and private life.,* New York, Vintage.

BULLER, D. B. & BURGOON, J. K. 1996. Interpersonal Deception Theory. *Communication Theory,* 6**,** 203-242.

BURGOON, J. K., BLAIR, J. P., QIN, T. T. & NUNAMAKER, J. F. 2003. Detecting deception through linguistic analysis. *Intelligence and Security Informatics, Proceedings,* 2665**,** 91-101.

CASTILLO, C., MENDOZA, M. & POBLETE, B. Information credibility on twitter. Proceedings of the 20th international conference on World wide web, 2011. ACM, 675-684.

CBC RADIO "AND THE WINNER IS". 31 March 2012. *"News 2.0, Part II", Retrieved from* http://www.cbc.ca/andthewinneris/ [Online]. [Accessed].

CHEN, Y., CONROY, N. J. & RUBIN, V. L. Misleading Online Content: Recognizing Clickbait as "False News". Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, 2015a. ACM, 15-19.

CHEN, Y., CONROY, N. J. & RUBIN, V. L. News in an online world: the need for an automatic crap detector. Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, 2015b. American Society for Information Science, 81.

CRYSTAL, D. 1969. *What is linguistics?,* Edward Arnold.

DAVIS, C. A., ONUR VAROL, O., FERRARA, E., FLAMMINI, A. & MENCZER, F. 2016. BotOrNot: A System to Evaluate Social Bots. *WWW'16 Companion.* Montréal, Québec, Canada.

DEPAULO, B. M., CHARLTON, K., COOPER, H., LINDSAY, J. J. & MUHLENBRUCK, L. 1997. The Accuracy-Confidence Correlation in the Detection of Deception. *Personality and Social Psychology Review,* 1**,** 346-357.

FOGG, B. J. & TSENG, H. The elements of computer credibility. SIGCHI conference on Human factors in computing systems: the CHI is the limit, 1999 Pittsburgh, Pennsylvania, United States. ACM.

FULLER, C. M., BIROS, D. P. & WILSON, R. L. 2009. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems,* 46**,** 695-703.

GIL, Y. & ARTZ, D. Towards content trust of web resources. 15th international conference on World Wide Web, 2006 Edinburgh, Scotland. 1135861: ACM, 565-574.

GRICE, H. P. 1975. Logic and conversation. *In:* COLE, P. & MORGAN, J. (eds.) *Syntax and semantics 3: Speech acts.* New York: Academic Press.

GUADAGNO, R. E. & WINGATE, V. S. 2014. Internet: Facebook and Social Media Sites. *In:* LEVINE, T. (ed.) *Encyclopedia of Deception.* Thousand Oaks, California: SAGE Publications.

HANCOCK, J. T., CURRY, L. E., GOORHA, S. & WOODWORTH, M. 2007. On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes,* 45**,** 1-23.

HARDIN, R. 2001. Conceptions and Explanations of Trust. *In:* COOK, K. S. (ed.) *Trust in Society.* New York, NY: Russell Sage Foundation.

HAUCH, V., MASIP, J., BLANDON-GITLIN, I. & SPORER, S. L. Linguistic cues to deception assessed by computer programs: a meta-analysis. Proceedings of the workshop on computational approaches to deception detection, 2012. Association for Computational Linguistics, 1-4.

HOLCOMB, J., GOTTFRIED, J. & MITCHELL, A. 2013. News use across Social Media Platforms. *Pew Research Journalism Project*.

HÖFER, E., AKEHURST, L. & METZGER, G. Reality monitoring: A chance for further development of CBCA. Proceedings of the Annual Meeting of the European Association on Psychology and Law, Sienna, Italy, 1996.
JOHNSON, K. A. & WIEDENBECK, S. 2009. Enhancing Perceived Credibility of Citizen Journalism Web Sites. *Journalism & Mass Communication Quarterly,* 86**,** 332-348.
KANG, B., HÖLLERER, T. & O'DONOVAN, J. Believe it or Not? Analyzing Information Credibility in Microblogs. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015. ACM, 611-616.
KUMAR, N. & REDDY, R. N. 2012. *Automatic Detection of Fake Profiles in Online Social Networks.* BTech Thesis.
KÖHNKEN, G. & STELLER, M. 1988. The evaluation of the credibility of child witness statements in the German procedural system. *Issues in Criminological & Legal Psychology*.
LARCKER, D. F. & ZAKOLYUKINA, A. A. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research,* 50**,** 495-540.
LEVINE, T. R., PARK, H. S. & MCCORNACK, S. A. 1999. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs,* 66**,** 125-144.
LIU, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies,* 5**,** 1-167.
LIU, X., NOURBAKHSH, A., LI, Q., FANG, R. & SHAH, S. Real-time Rumor Debunking on Twitter. Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015 Melbourne, Australia. ACM, 1867-1870.
LUKOIANOVA, T. & RUBIN, V. L. 2014. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances in Classification Research Online,* 24**,** 4.
MANNING, C. D. & SCHÜTZE, H. 1999. *Foundations of statistical natural language processing,* MIT Press.
MARSH, S. & DIBBEN, M. R. 2003. The role of trust in information science and technology. *Annual Review of Information Science and Technology,* 37**,** 465-498.
MATTHEWS, C. 2013. How Does One Fake Tweet Cause a Stock Market Crash. *Time.*
MCCAY-PEET, L. & QUAN-HAASE, A. Forthcoming 2016. What is social media and what questions can social media research help us answer? *In:* SLOAN, L. & QUAN-HAASE, A. (eds.) *Handbook of Social Media Research Methods.* London, UK: Sage.
MENDOZA, M., POBLETE, B. & CASTILLO, C. Twitter Under Crisis: Can we trust what we RT? Proceedings of the first workshop on social media analytics, 2010. ACM, 71-79.
MERRIAM-WEBSTER ONLINE DICTIONARY. 2016. Available: http://www.merriam-webster.com/dictionary/ [Accessed].
MERRIAM-WEBSTER.COM Trust. *Merriam-Webster.*
MIHALCEA, R. & STRAPPARAVA, C. The lie detector: Explorations in the automatic recognition of deceptive language. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009. Association for Computational Linguistics, 309-312.
MINTZ, A. 2002. *Web of Deception: Misinformation on the Internet,* Medford, N.J., CyberAge Books.
MITCHELL, A. & PAGE, D. 2015. State of the News Media 2015. *Pew Research Journalism, Project for Excellence in Journalism*.
MORANTE, R. & SPORLEDER, C. 2012. Modality and Negation: An Introduction to the Special Issue. *Computational Linguistics,* 38**,** 223-260.
MORRIS, M. R., COUNTS, S., ROSEWAY, A., HOFF, A. & SCHWARZ, J. Tweeting is believing?: understanding microblog credibility perceptions. Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, 2012. ACM, 441-450.
MUKHERJEE, A., VENKATARAMAN, V., LIU, B. & GLANCE, N. 2013. Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. Technical Report.: Department of Computer Science, University of Illinois at Chicago, and Google Inc.

NEWMAN, M. L., PENNEBAKER, J. W., BERRY, D. S. & RICHARDS, J. M. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin,* 29**,** 665-675.

ORTEGA, J. L. 2015. Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics,* 9**,** 39-49.

PANG, B. & LEE, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval,* 2**,** 1-135.

PENNEBAKER, J. W. & FRANCIS, M. E. 1999. *Linguistic inquiry and word count: LIWC.*, Erlbaum Publishers.

PENNEBAKER, J. W., FRANCIS, M. E. & BOOTH, R. J. 2001. Linguistic inquiry and word count (LIWC): A computerized text analysis program. *Mahwah (NJ),* 7.

PORTER, S. & YUILLE, J. C. 1996a. The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior,* 20**,** 443.

PORTER, S. & YUILLE, J. C. 1996b. The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior,* 20**,** 443-458.

QAZVINIAN, V., ROSENGREN, E., RADEV, D. R. & MEI, Q. Rumor has it: Identifying misinformation in microblogs. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011. Association for Computational Linguistics, 1589-1599.

RIEH, S. Y. 2010. Credibility and Cognitive Authority of Information. *In:* BATES, M. J. (ed.) *Encyclopedia of library and information sciences.* Taylor & Francis.

RUBIN, V. L. 2006a. Identifying certainty in texts. *Unpublished Doctoral Thesis, Syracuse University, Syracuse, NY.*

RUBIN, V. L. 2006b. *Identifying Certainty in Texts. Thesis.* Syracuse University.

RUBIN, V. L. 2010a. Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts. *Information Processing & Management,* 46**,** 533-540.

RUBIN, V. L. On deception and deception detection: content analysis of computer-mediated stated beliefs. 73rd ASIS&T Annual Meeting: Navigating Streams in an Information Ecosystem, 2010b Pittsburgh, Pennsylvania. American Society for Information Science.

RUBIN, V. L. 2014. TALIP Perspectives, Guest Editorial Commentary. *ACM Transactions on Asian Language Information Processing,* 13**,** 1-8.

RUBIN, V. L., CHEN, Y. & CONROY, N. J. Deception detection for news: three types of fakes. Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community, 2015. American Society for Information Science, 83.

RUBIN, V. L. & CONROY, N. 2011. *Challenges in Automated Deception Detection in Computer-Mediated Communication* [Online]. New Orleans, Louisiana. [Accessed].

RUBIN, V. L. & CONROY, N. 2012. Discerning truth from deception: Human judgments and automation efforts. *First Monday* [Online], 17. Available: http://firstmonday.org.

RUBIN, V. L., KANDO, N. & LIDDY, E. D. Certainty categorization model. AAAI spring symposium: Exploring attitude and affect in text: Theories and applications, Stanford, CA, 2004.

RUBIN, V. L. & LIDDY, E. Assessing Credibility of Weblogs. AAAI Symposium on Computational Approaches to Analyzing Weblogs, 2006 Stanford, CA. AAAI Press.

RUBIN, V. L., LIDDY, E. D. & KANDO, N. 2006. Certainty identification in texts: Categorization model and manual tagging results. *Computing attitude and affect in text: Theory and applications.* Springer.

RUBIN, V. L. & VASHCHILKO, T. 2012. Extending information quality assessment methodology: A new veracity/deception dimension and its measures. *Proceedings of the American Society for Information Science and Technology,* 49**,** 1-6.

SAURI, R. & PUSTEJOVSKY, J. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation,* 43**,** 227-268.

SAURI, R. & PUSTEJOVSKY, J. 2012. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics***,** 1-39.

SIKDAR, S., KANG, B., O'DONOVAN, J., HOLLERER, T. & ADAH, S. Understanding information credibility on twitter. Social Computing (SocialCom), 2013 International Conference on, 2013. IEEE, 19-24.

STELLER, M. 1989. Recent developments in statement analysis. *Credibility assessment.* Springer.

SUBRAHMANIAN, V., AZARIA, A., DURST, S., KAGAN, V., GALSTYAN, A., LERMAN, K., ZHU, L., FERRARA, E., FLAMMINI, A. & MENCZER, F. 2016. The DARPA Twitter Bot Challenge.

THE URBAN DICTIONARY. 2016. *astroturfing* [Online]. Available: http://www.urbandictionary.com/define.php?term=astroturf [Accessed].

THELWALL, M. Forthcoming, 2016. Sentiment analysis. *In:* SLOAN, L. & QUAN-HAASE, A. (eds.) *The Handbook of Social Media Research Methods.* London, UK: Sage.

THR FREE DICTIONARY 2016. rumor. *The Free Dictionary*.

TOMA, C. L. & HANCOCK, J. T. 2012. What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles. *Journal of Communication,* 62**,** 78-97.

TSENG, H. & FOGG, B. J. 1999. Credibility and computing technology. *Communications of the ACM,* 42**,** 39-44.

VAN SWOL, L. 2014. Truth Bias. *In:* LEVINE, T. (ed.) *Encyclopedia of Deception.* Thousand Oaks, California: SAGE Publications.

VOSOUGHI, S. 2015. *Automatic detection and verification of rumors on Twitter.* Doctor of Philosophy, Massachusetts Institute of Technology.

VRIJ, A. 2000. *Detecting Lies and Deceit,* New York, John Wiley and Sons.

VRIJ, A. 2004. Why professionals fail to catch liars and how they can improve. *Legal and criminological psychology,* 9**,** 159-181.

VRIJ, A., MANN, S., KRISTEN, S. & FISHER, R. P. 2007. Cues to deception and ability to detect lies as a function of police interview styles. *Law and human behavior,* 31**,** 499.

VRIJ, A., MANN, S. & LEAL, S. 2012. Deception Traits in Psychological Interviewing. *Journal of Police and Criminal Psychology,* 28**,** 115-126.

WALCZYK, J. J., RUNCO, M. A., TRIPP, S. M. & SMITH, C. E. 2008. The Creativity of Lying: Divergent Thinking and Ideational Correlates of the Resolution of Social Dilemmas. *Creativity Research Journal,* 20**,** 328 - 342.

WEERKAMP, W. & DE RIJKE, M. Credibility improves topical blog post retrieval. HLT-NAACL, 2008 Columbus, Ohio. 923-931.

WIEBE, J., BRUCE, R., BELL, M., MARTIN, M. & WILSON, T. A corpus study of evaluative and speculative language. Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16, 2001. Association for Computational Linguistics, 1-10.

WU, K., YANG, S. & ZHU, K. Q. False Rumors Detection on Sina Weibo by Propagation Structures. IEEE International Conference on Data Engineering, ICDE, 2015.

YANG, F., LIU, Y., YU, X. & YANG, M. Automatic Detection of Rumor on Sina Weibo. Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, 2012 Beijing, China. ACM, 13.

ZHOU, L., BURGOON, J. K., NUNAMAKER, J. F. & TWITCHELL, D. 2004. Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications. *Group Decision and Negotiation,* 13**,** 81-106.