Deception Detection for News: Three Types of Fakes

Victoria L. Rubin, Yimin Chen and Niall J. Conroy
Language and Information Technology Research Lab (LIT.RL)
Faculty of Information and Media Studies
University of Western Ontario, London, Ontario, CANADA
vrubin@uwo.ca, ychen582@uwo.ca, nconroy1@uwo.ca

ABSTRACT

A *fake news detection* system aims to assist users in detecting and filtering out varieties of potentially deceptive news. The prediction of the chances that a particular news item is intentionally deceptive is based on the analysis of previously seen truthful and deceptive news. A scarcity of deceptive news, available as corpora for predictive modeling, is a major stumbling block in this field of natural language processing (NLP) and deception detection. This paper discusses three types of fake news, each in contrast to genuine serious reporting, and weighs their pros and cons as a corpus for text analytics and predictive modeling. Filtering, vetting, and verifying online information continues to be essential in library and information science (LIS), as the lines between traditional news and online information are blurring.

Keywords

news verification; deception detection; fake news detection; credibility assessment; reputable sources; fabrication; hoax; satire; natural language processing; text analytics; predictive modeling; corpus construction.

INTRODUCTION

The boundaries between news production and information creation and sharing are gradually blurring in the current online news environments and social media (Chen, Conroy, & Rubin, 2015). Fleeting news becomes part of the static searchable information realm. Broad audiences gain access to news through traditional search engines, digital forms of mainstream news channels and citizen information sharing platforms. While the issue of news verification and fact-checking has traditionally been a matter of journalistic endeavor, LIS offers meaningful insights on credibility assessment and cutting-edge information technologies. As LIS is redefining itself in the age of fast-streaming big data and text analytics, it is turning to automated methods for filtering, vetting, and verifying online information.

Deceptive news, such as fake news, phony press releases and hoaxes, may be misleading or even harmful, especially when they are disconnected from their original sources and contexts. "A 2012 report of Pew Internet Research on the future of big data (Anderson and Rainie, 2012) argues that even though by 2020 big data is likely to have a transformational effect on our knowledge and understanding of the world, there is also danger from inaccurate or false information (called "distribution of harms" (Rumors and Deception in Social Media Workshop, 2015). Occasionally reports of non-existent, surreal, alarming events have been taken seriously. For instance, "Jack Warner, the former FIFA vice president, has apparently been taken in by a spoof article from the satirical website *The Onion*" (Topping, 2015) after *The Onion* had suggested that the FIFA corruption scandal would result in a 2015 Summer Cup in the U.S.

Journalistic deception is "an act of communicating messages verbally (a lie) or nonverbally through the withholding of information with the intention to initiate or sustain a false belief" (Elliot and Culver, 1992, cited in Lee, 2004, p. 98). Cases of outright fabrications by celebrity hosts or journalists have been recently reported (Compton & Benedetti, 2015; Shingler, 2015). The journalism community observes that the responsibility for knowing what is true shifts to individuals, as news consumers, with two predominant overwrought views: "Utopians have heralded this as the end of journalism and the information monopoly of elites and see a citizen media culture that instantly self-corrects – a kind of pure information democracy. Critics see a world without editors, of unfettered spin, where the loudest or most agreeable voice wins and where truth is the first casualty" (Kovach & Rosentiel, 2010, p. 7). With distinctions between genuine and misleading news eroding, "few news verification mechanisms currently exist, and the sheer volume of the information requires novel automated approaches" (Rubin, Conroy, & Chen, 2015).

One of the new branches in LIS and NLP attempts to identify and flag instances of intentionally deceptive news that could be disseminated to unsuspecting audiences. The task of *fake news detection* is defined here as the prediction of the chances of a particular news article (news report, editorial, expose, etc.) being intentionally deceptive (fake, fabricated, staged news, or a hoax). A text analytical system can enhance human abilities to spot lies, provide evidence for its suspicions, and alert users to further fact-checking (for detection methods, see Conroy, Chen & Rubin, 2015).

The objectives of this paper are: 1) to overview fake news corpora requirements (for suitability in textual analysis and predictive modeling); 2) suggest pros and cons of varieties of fake news, as counterparts to serious genuine reporting. This research ultimately supports the development of an automated fake news detection system as part of a broader news verification suite.

ASIST 2015, November 6-10, 2015, St. Louis, MO, USA.

BACKGROUND

Data Collection Practices in Deception Detection R&D

Most of the insights on deception research originate from disciplines without detection automation in mind. We start by overviewing their data collection practices and associated data formats. All of the below-mentioned disciplines are united by the pursuit of understanding of what deception is and how we, humans, can identify liars.

In *interpersonal psychology*, questionnaires, interviews, case scenarios discussions and observation data (often in verbal format) tend to form the basis for analyses. Since 2004 automated cues of deception have been identified in psychological textual data with NLP techniques (e.g., Zhou et al., 2004). Reliability and consistency of the cues are disputed but one fact is regularly cited: in spotting lies humans don't do much better than chance, and machines can slightly outperform humans on restricted tasks (Rubin & Conroy, 2012).

In *law enforcement, credibility assessments, police work* and *homeland security*, researchers gain permissions to use verbal evidence such as court proceedings, police interviews, credibility assessment transcripts, testimonies, pleas, or alibi (e.g., Porter & Yuille, 1996). Court decisions then serve as the determination of "ground truth".

In *computer-mediated communication (CMC),* corpora (of e-mails, forum posts, etc.) are collected via elicitations or submissions of pre-existing messages. Study participants are invited to contribute stories or interact under truthful or deceptive conditions. Such data as deceptive interpersonal e-mail (Keila & Skillicorn, 2005) can be re-purposed for text analytics. Publicly available social media data are a fruitful source for *sentiment analysis.* Tweets are especially suited for detecting such irregular communication behaviors in information sharing as hoaxes (Weissman, 2015).

In *NLP (*or *text analytics),* other types of fake data have been collected by crawling the web or crowdsourcing: fake product reviews (Mukherjee, Venkataraman, Liu, & Glance, 2013), fudged online resumes (Guillory & Hancock, 2012), opinion spamming (Jindal & Liu, 2008), fake social network profiles (Kumar & Reddy, 2012), fake dating profiles (Toma & Hancock, 2012), spamming and phishing (Toolan & Carthy, 2010), and forged scientific work (SciGen, Mathgen). Some dataset are readily available, such as "Boulder Lies and Truth" (Salvetti, 2014), but are restricted in content (e.g., to hotels and electronics reviews). One recent specialized workshop focused on rumor identification (Rumors and Deception in Social Media, 2015), but to the best of our knowledge, in the context of digital news, there has of yet been no concerted effort in classifying potential sources of deceptive news and making dataset available for R&D.

In the reminder of the discussion we first specify R&D data requirements for modeling and testing fake news detection algorithms. Then, we identify and classify news genres and exemplify suitable sources of questionable veracity.

Requirements for Fake News Detection Corpus

Based on the reviewed practices and our own previous research, R&D data should meet the following conditions:

1. *Availability of both truthful and deceptive instances*. Predictive methods should be able to find patterns and regularities in positive and negative data points. Finding counterparts to authentic genuine news is the challenge.

2. *Digital textual format accessibility*. Text is the preferred medium in NLP. Images can be accompanied with texts and audio and video need to be transcribed to be suitable.

3. *Verifiability of "ground truth"*. The question is what constitutes verification and how does one know when the news is clearly genuine or fabricated. Credible news sources that withstood a test of time and have a reputation based on a system of "checks and balances" are advisable as corpora. Crowdsourced ratings (of whether a message is truth or a lie) have been previously used as a metric of success in identifying lies.

4. *Homogeneity in lengths.* Special care should be paid to obtaining comparable article lengths for individual data points. For example, a short tweet with a headline, a Facebook one-paragraph summary and a longform op-ed do not constitute a homogeneous dataset. Normalization can be performed for some unevenly distributed datasets.

5. *Homogeneity in writing matter.* The corpus should be aligned along news genres (e.g., breaking news, editorials, op-eds) and topics (e.g., business, politics, science, health), be using similar types of authors (e.g., professionally trained, mainstream vs. citizen journalists; serious vs. humorous), and be compared across news outlets.

6. *Predefined timeframe.* The corpus has to be collected within a thought-out timeframe. A snapshot of breaking daily news may have greater variation than all of the news on a particular topic over the past 2-3 years.

7. *The manner of news delivery* (e.g., *humor; newsworthiness, believability; absurdity; sensationalism*). Knowing the manner in which the news is provided creates context for situational interpretation. We suspect that normally "truth-biased" readers may, for instance, shift to a "lie-biased" perspective when reading news satire. More research is needed on the interaction of humor and deception.

8. *Pragmatic concerns* include copy-right costs, public availability, ease of obtaining, suitable overall volume of data, degrees of disclosure, and writers' privacy.

9. *Language* and *culture* are silent factors that are often overlooked. English is the predominant language in deception detection with only a handful of other languages explored and reported in deception research (e.g., Spanish, Italian, Mandarin) with little consideration given to language and cultural differences (Rubin, 2014).

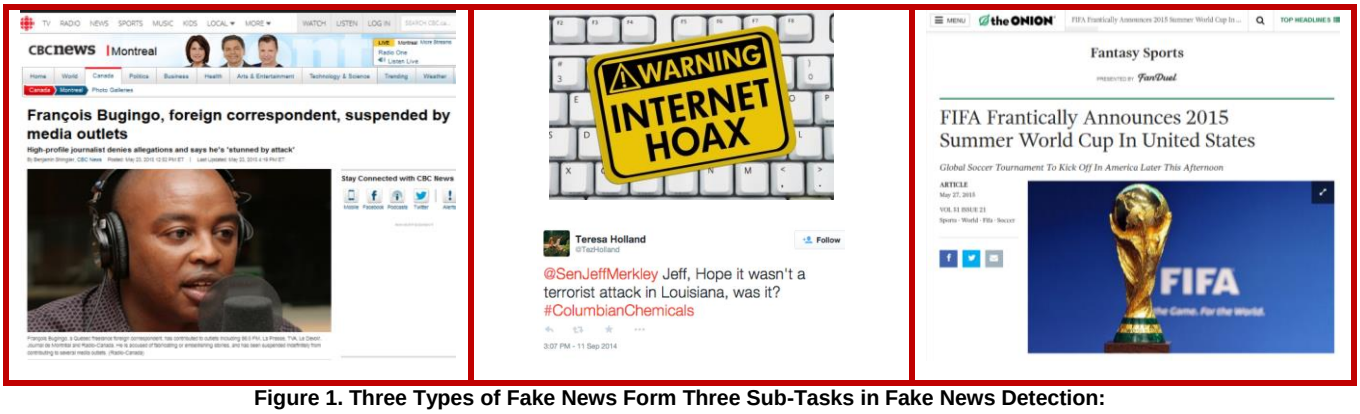

Figure 1. Three Types of Fake News Form Three Sub-Tasks in Fake News Detection: A) exposed fabrications (Shingler, 2015); B) large-scale hoaxes (Matt, 2015); C) news satire (Fan Duel, 2015).

GENUINE NEWS

Truthful reporting in digital journalism is abundant in well-established print or radio "legacy journalism" outlets (e.g., www.nytimes.com, www.bbc.co.uk, www.cbc.ca) or reputable citizen journalist blogs with an established reputation. Such sources can be considered genuine unless proven otherwise, retracted or corrected, but may require permissions or costs for R&D use. Predictive features of veracity may vary by genres and topics, and thus should be kept consistent from truthful to deceptive news.

Finger-pointing

In political events, territorial conflicts, wars or other current controversies, news channels or individual reporters may be accused of partisanship, blindness, or straight out lies. Such situations do not meet the intentional lying criterion, since reporting is likely to be consistent with the reporter's beliefs, worldview, biases, or affiliations. The Israeli-Palestinian or Russian-Ukrainian conflicts are readily offered up as candidates for news verification data, but they lack clear unbiased determination of the "ground truth".

DECEPTIVE NEWS

Deceptive news can be harvested, crowdsourced, or mimicked by qualified study participants. An elicited dataset can be tailored to the above-mentioned nine requirements, but the "ground truth" verification mechanism needs to be built into the data collection process. Below we discuss the three types of fake news, each in contrast to genuine serious reporting, suggesting that there are at least three distinct sub-tasks in fake news detection: a) fabrication, b) hoaxing and c) satire detection.

Serious Fabrications (Type A, Figure 1A)

Fraudulent reporting is not unheard of in both old and new media. Exposed fraudulent journalistic writing, discussed in Compton & Benedetti (2015) or Shingler (2015; Fig. 1A), are ideal for a fake news corpus. It is time-consuming and tedious to collect original copies of uncovered fabricated news; and resulting corpora may be limited in size for NLP.

Yellow press and *tabloids* present a wide spectrum of unverified new and uses eye-catching headlines ("clickbaits"), exaggerations, scandal-mongering, or sensationalism to increase traffic or profits (Yellow Journalism, 2015). *Tabloids* specifically emphasize topics such as sensational crime stories, astrology, gossip columns about celebrities, and junk food news (Tabloids, 2015). *Yellow journalism* is a suitable source for fake news corpus in cases of obvious or exposed falsification, fabrication, or exaggeration, and may require investigation.

Large-Scale Hoaxes (Type B, Figure 1B)

Hoaxing is another type of deliberate fabrication or falsification in the mainstream or social media. Attempts to deceive audiences masquerade as news, and may be picked up and mistakenly validated by traditional news outlets. Brunvand (1998) distinguishes a hoax from *pranking* or *practical joking* as "relatively complex and large-scale fabrications" which may include deceptions that go beyond the merely playful and "cause material loss or harm to the victim" (p. 875). The #Columbian Chemical plant hoax is an example of a harmful multi-platform attack (Chen, 2015) with an identifiable communication pattern (Matt, 2015).

Humorous Fakes (Type C)

We distinguish serious fabricated news from humorous ones. If readers are aware of the humorous intent, they may no longer be predisposed to take the information at face value. Technology can identify humor and prominently display originating sources (e.g., *The Onion*) to alert users, especially in decontextualized news aggregators/platforms.

News Satire (Figure 1C)

News satire sites, news parody, a.k.a. literally *fake news* (e.g., *The Onion* and *CBC's This is That)* is one specific genre with numerous sites that present news "in a format typical of mainstream journalism but rely heavily on irony and deadpan humor to emulate a genuine news source, mimicking credible news sources and stories, and often achieving wide distribution" (News Satire, 2015).

News Game Shows

Unusual news pieces are written by professional journalists for *the NPR's Bluff the Listener* news game show. They are read aloud to listeners who call in to guess which news is true (though often unbelievable). Stylistic elements of absurdity and humor may interfere with the news writers' intention to deceive (Rubin, Conroy, & Chen, 2015). More research is needed on this area.

CONCLUSIONS

This paper separates the task of fake news detection into three, by type of fake: a) serious fabrications (uncovered in mainstream or participant media, yellow press or tabloids); b) large-scale hoaxes; c) humorous fakes (news satire, parody, game shows). Serious fabricated news may take substantial efforts to collect, case by case. Journalistic fraudsters may face harsh consequences for dishonest reporting, and are likely to exhibit cues of deception akin to "verbal leakages" in other contexts (such as law enforcement or CMC). Large-scale hoaxing attacks are creative, unique, and often multi-platform, which may require methods beyond text analytics (e.g., network analysis). Humorous news provides a steady stream of data, but their writers' intentions to entertain, mock, and be absurd may interfere with binary text classification techniques, especially if algorithms pick up cues of believability, sensationalism, or humor instead of cues for deception. As an important development at an intersection of LIS, NLP, big data and journalism, fake news detection (of the three identified types) holds promise in automated news verification and online content credibility assessment.

ACKNOWLEDGMENTS

This research has been funded by the Government of Canada Social Sciences and Humanities Research Council (SSHRC) Insight Grant (#435-2015-0065) awarded to Dr. Rubin for the project entitled *Digital Deception Detection: Identifying Deliberate Misinformation in Online News.*

REFERENCES

Brunvand, J.H. (1998). *American Folklore: An Encyclopedia*. Taylor & Francis.

Chen, A. (2 June 2015). *The Agency.* Retrieved on 16 June 2015 from www.nytimes.com/2015/06/07/magazine/the-agency.html

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). News in an Online World: The Need for an "Automatic Crap Detector". In *The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015)*, Nov. 6-10, St. Louis.

Conroy, N. J., Chen, Y., & Rubin, V. L. (2015). Automatic Deception Detection: Methods for Finding Fake News. In *The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015)*, Nov. 6-10, St. Louis.

Compton, J.R. & Benedetti, P. (5 March 2015) News, Lies & Videotape: The Legitimation Crisis In Journalism. *Rabble*. Retrieved on June, 14 2015 from rabble.ca/news/2015/03/news-lies-and-videotape-legitimation-crisis-journalism

Elliot, D., & Culver, C. (1992). Defining & analyzing journalistic deception. *Journal of Mass Media Ethics*, 7(2), 69–84.

Fan Duel. (27 May 2015) FIFA Frantically Announces 2015 Summer World Cup In United States. *The Onion.*

Guillory, J. & Hancock, J.T. (2012) The Effect of Linkedin on Deception in Resumes, *Cyberpsychology, Behavior, & Social Networking*, 15 (3), 135-140.

Jindal, N. & Liu, B. (2008). Opinion Spam and Analysis, ACM

Keila, P. & Skillicorn, D. (21005). Detecting unusual & deceptive communication in email. Technical Report, School of Computing, Queen's University, Kingston, Canada.

Kovach, B. & Rosentiel, T. (2010). *Blur: How to Know What's True in the Age of Information Overload.* Bloomsbury, NY.

Kumar, N. & Reddy, R.N. (2012) Automatic Detection of Fake Profiles in Online Social Networks, BTech Thesis.

Lee, S. T. (2004). Lying to Tell the Truth: Journalists & Social Context of Deception. *Mass Communic. & Society 7(1)*, 97-120.

"News Satire". *Wikipedia.* Retrieved on 13 June, 2015 from en.wikipedia.org/wiki/ News_satire.

Mathgen: Randomly Generated Mathematics Research Papers! thatsmathematics.com/mathgen/

Matt (June 12, 2015). #ColumbianChemicals Hoax: Trolling the Gulf Coast for Deceptive Patterns. *Cyber Threat Intelligence.* Retrieved on June 24, 2015 from www.recordedfuture.com/columbianchemicals-hoax-analysis/

Mukherjee, A., Venkataraman, Liu, & Glance (2013). Fake Review Detection: Classification & Analysis of Real & Pseudo Reviews. Technical Report, Dept. of CompSci, University of Illinois, Chicago & Google Inc.

Porter, S. & Yuille, J. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. *Law & Human Behavior 20*(4), 443–458.

Rubin, V.L. (2014) Pragmatic and Cultural Considerations for Deception Detection in Asian Languages. *TALIP Perspectives, Guest Editorial Commentary, 13 (2)*.

Rubin, V.L. & Conroy, N. (2012). Discerning truth from deception: Human judgments & automation efforts. *First Monday 17 (*3-5). dx.doi.org/10.5210/fm.v17i3.3933

Rubin, V.L., Conroy, N., & Chen, Y. (2015). Towards News Verification: Deception Detection Methods for News Discourse. The Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, *Hawaii International Conference on System Sciences (HICSS48)*, January 2015.

Rumors and Deception in Social Media: Detection, Tracking, and Visualization Workshop (2015). Workshop Statement & Aims. May 19, 2015, Florence, Italy www.pheme.eu/events/rdsm2015

Salvetti, F. (2014). Boulder Lies and Truth LDC2014T24. Web Download. Philadelphia: *Linguistic Data Consortium*.

Shingler, B. (23 May 2015). François Bugingo, Foreign Correspondent, Suspended By Media Outlets. *CBC News Montreal.*

SciGen: Automatic CS Paper Generator. pdos.csail.mit.edu/scigen/

"Tabloid Journalism". *Wikipedia*. en.wikipedia.org/wiki/Tabloid_journalism

Toma, C.L. & Hancock, J.T. (2012) What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles, *Journal of Communication*, 62(1), 78-97.

Topping, A. (31 May 2015). "Ex-Fifa vice president Jack Warner swallows Onion spoof". *The Guardian*.

Toolan, F., & Carthy, J. (2010). Feature Selection for Spam and Phishing Detection. *E-Crime Researchers Summit*, Dallas,1-12.

Weissman, C. G. (12 June 2015). Computers can now tell the difference between real breaking news and internet trolls trying to dupe us. *Business Insider*.

"Yellow Journalism". *Wikipedia*. en.wikipedia.org/wiki/Yellow_journalism.

Zhou, et al. (2004). "Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications," *Group Decision & Negotiation 13(1)*, 81-106.