Automatic Deception Detection: Methods for Finding Fake News

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

ABSTRACT

This research surveys the current state-of-the-art technologies that are instrumental in the adoption and development of fake news detection. “Fake news detection” is defined as the task of categorizing news along a continuum of veracity, with an associated measure of certainty. Veracity is compromised by the occurrence of intentional deceptions. The nature of online news publication has changed, such that traditional fact checking and vetting from potential deception is impossible against the flood arising from content generators, as well as various formats and genres.

The paper provides a typology of several varieties of veracity assessment methods emerging from two major categories – linguistic cue approaches (with machine learning), and network analysis approaches. We see promise in an innovative hybrid approach that combines linguistic cue and machine learning, with network-based behavioral data. Although designing a fake news detector is not a straightforward problem, we propose operational guidelines for a feasible fake news detecting system.

Keywords

Deception detection, fake news detection, veracity assessment, news verification, methods, automation, SVM, knowledge networks, predictive modelling, fraud

INTRODUCTION

News verification aims to employ technology to identify intentionally deceptive news content online, and is an important issue within certain streams of library and information science (LIS). Fake news detection is defined as the prediction of the chances of a particular news article (news report, editorial, expose, etc.) being intentionally deceptive (Rubin, Conroy & Chen, 2015). Tools aim to mimic certain filtering tasks which have, to this point, been the purview of journalists and other publishers of traditional news content. The proliferation of user-generated content, and Computer Mediated Communication (CMC) technologies such as blogs, Twitter, and other social media have the capacity of news delivery mechanisms on a mass scale— yet much of the information is of questionable veracity (Ciampaglia, Shiralkar, Rocha, Bollen, Menczer & Flammini, 2015). Establishing the reliability of information online is a daunting but critical challenge. Four decades of deception detection research has helped us learn about how well humans are able detect lies in text. The findings show we are not so good at it. In fact, just 4% better than chance, based on a meta-analysis of more than 200 experiments (Bond & DePaulo, 2006). This problem has led researchers and technical developers to look at several automated ways of assessing the truth value of potentially deceptive text based on the properties of the content and the patterns of computer-mediated communication .

Structured datasets are easier to verify than non-structured (or semi-structured) data such as texts. When we know the language domain (e.g., insurance claims or health-related news) we can make better guesses about the nature and use of deception. Semi-structured non-domain specific web data come in many formats and demand flexible methods for veracity verification. For some time, however, the development and evaluation of different methods have remained in isolated corners, relatively unknown in LIS. More recently, efforts of methodological cross-pollination and hybrid approaches have produced promising results (Rubin et al., 2015^{A}). The range of journalistic practices and available news sources (see Rubin et al. (2015^{B}) for an overview) demand consideration of multiple methods since one approach often addresses known weaknesses in another. How then is it possible to gauge the veracity of online news?

This paper provides researchers with a map of the current landscape of veracity (or deception) assessment methods, their major classes and goals, all with the aim of proposing a hybrid approach to system design. These methods have emerged from separate development streams, utilizing disparate techniques. In this survey, two major categories of methods emerge: 1. *Linguistic Approaches* in which the content of deceptive messages is extracted and analyzed to associate language patterns with deception; and 2. *Network Approaches* in which network information, such as message metadata or structured knowledge network queries can be harnessed to provide aggregate deception measures. Both forms typically incorporate machine learning techniques for training classifiers to suit the analysis. It is incumbent upon researchers to understand these different areas, yet no known typology of methods exists in the current literature. The goal is to provide a survey of the existing research while proposing a hybrid approach, which utilizes the most effective deception detection methods for the implementation of a fake news detection tool.

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

LINGUISTIC APPROACHES

Most liars use their language strategically to avoid being caught. In spite of the attempt to control what they are saying, language "leakage" occurs with certain verbal aspects that are hard to monitor such as frequencies and patterns of pronoun, conjunction, and negative emotion word usage (Feng & Hirst, 2013). The goal in the linguistic approach is to look for such instances of leakage or, so called "predictive deception cues" found in the content of a message.

Data Representation

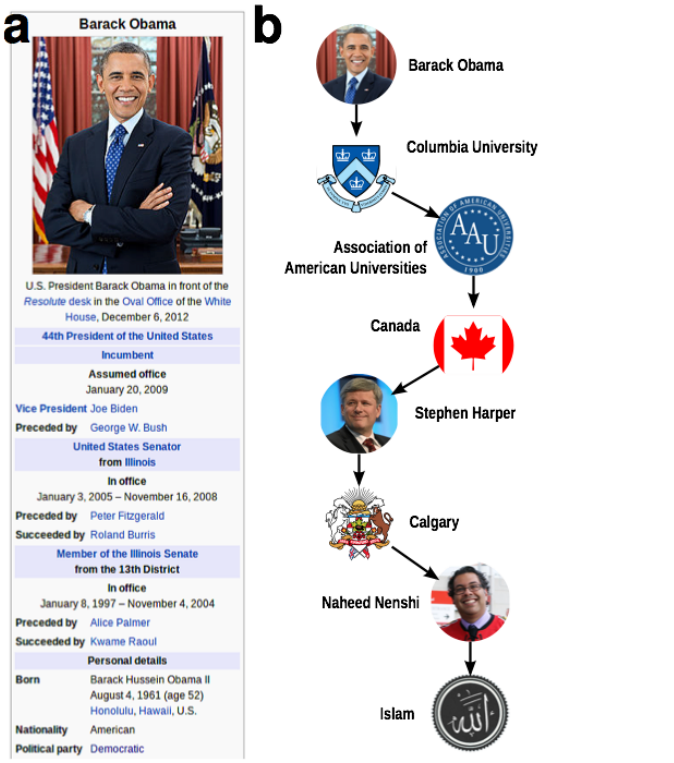
Perhaps the simplest method of representing texts is the "bag of words" approach, which regards each word as a single, equally significant unit. In the bag of words approach, individual words or "n-grams" (multiword) frequencies are aggregated and analyzed to reveal cues of deception. Further tagging of words into respective lexical cues for example, parts of speech or "shallow syntax" (Hancock & Markowitz, 2013), affective dimensions (Vrij, 2006), or location-based words (Hancock, et al, 2013) are all ways of providing frequency sets to reveal linguistic cues of deception.

The simplicity of this representation also leads to its biggest shortcoming. In addition to relying exclusively on language, the method relies on isolated n-grams, often divorced from useful context information. In this method, any resolution of ambiguous word sense remains non-existent (Larcker & Zakolyukina 2012). Many deception detection researchers have found this method useful in tandem with different, complementary analysis (Zhang, Fan, Zeng & Liu, 2012; Lary, Nikitov & Stone, 2010; Ott, Cardi, & Hancock, 2013), several of which are discussed in the remainder of this proposal.

Deep Syntax

Analysis of word use is often not enough in predicting deception. Deeper language structures (syntax) have been analyzed to predict instances of deception. Deep syntax analysis is implemented through Probability Context Free Grammars (PCFG). Sentences are transformed to a set of rewrite rules (a parse tree) to describe syntax structure, for example noun and verb phrases, which are in turn rewritten by their syntactic constituent parts (Feng, Banerjee & Choi, 2012). The final set of rewrites produces a parse tree with a certain probability assigned. This method is used to distinguish rule categories (lexicalized, unlexicalized, parent nodes, etc.) for deception detection with 85-91% accuracy (depending on the rule category used) (Feng et al., 2012).

Third-party tools, such as the Stanford Parser (de Marneffe, MacCartney, Manning, 2006; Rahangdale & Agrawa, 2014), AutoSlog-TS syntax analyzer (Oraby, Reed, Compton, Riloff, Walker, & Whittaker, 2015) and others assist in the automation. Alone, syntax analysis might not be sufficiently capable of identifying deception, and studies often combine this approach with other linguistic or network analysis techniques (e.g., Feng et al., 2012; Feng & Hirst, 2013).



Barack Obama

U.S. President Barack Obama in front of the *Resolute* desk in the Oval Office of the White House, December 6, 2012

44th President of the United States

Incumbent

Assumed office
January 20, 2009

Vice President Joe Biden

Preceded by George W. Bush

United States Senator from Illinois

In office
January 3, 2005 – November 16, 2008

Preceded by Peter Fitzgerald

Succeeded by Roland Burris

Member of the Illinois Senate from the 13th District

In office
January 8, 1997 – November 4, 2004

Preceded by Alice Palmer

Succeeded by Kwame Raoul

Personal details

Born	Barack Hussein Obama II August 4, 1961 (age 52) Honolulu, Hawaii, U.S.
Nationality	American
Political party	Democratic

Figure 1: Fact-checking statements. (a) Structured information about President Obama contained in the "infoboxes" of Wikipedia articles. (b) Shortest knowledge graph path returned for the false statement "Barack Obama is a Muslim". The path traverses high-degree nodes representing generic entities, such as Canada, and is assigned a low truth value. (Ciampiaglia et al., 2015)

Semantic Analysis

As an alternative to deception cues, signals of truthfulness have also been analyzed and achieved by characterizing the degree of compatibility between a personal experience (e.g., a hotel review) as compared to a content "profile" derived from a collection of analogous data. This approach extends the n-gram plus syntax model by incorporating profile compatibility features, showing the addition significantly improves classification performance. (Feng & Hirst, 2013). The intuition is that a deceptive writer with no experience with an event or object (e.g., never visited the hotel in question) may include contradictions or omission of facts present in profiles on similar topics. For product reviews, a writer of a truthful review is more likely to make similar comments about aspects of the product as other truthful reviewers. Extracted content from key words consists of *attribute:descriptor* pair. By aligning profiles and the description of the writer's personal experience, veracity assessment is a function of the compatibility scores: 1. Compatibility with the existence of some distinct aspect (eg. an art museum near the hotel); 2. Compatibility with the description of some general aspect, such as location or service. Prediction of falsehood is shown to be approximately 91% accurate with this method.

Although demonstrated useful in the above context of reviews, this method has so far been restricted to the domain of application. There are two potential limitations in this method: the ability to determine alignment between attributes and descriptors depends on a sufficient amount of mined content for profiles, and the challenge of correctly associating descriptors with extracted attributes.

Rhetorical Structure and Discourse Analysis

At the discourse level, deception cues present themselves both in CMC communication and in news content. A description of discourse can be achieved through the Rhetorical Structure Theory (RST) analytic framework, that identifies instances of rhetoric relations between linguistic elements. Systematic differences between deceptive and truthful messages in terms of their coherence and structure has been combined with a Vector Space Model (VSM) that assesses each message's position in multi-dimensional RST space with respect to its distance to truth and deceptive centers (Rubin & Lukoianova, 2014). At this level of linguistic analysis, the prominent use of certain rhetorical relations can be indicative of deception. Tools to automate rhetorical classification are becoming available, although not yet employed in the context of veracity assessment.

Classifiers

Sets of word and category frequencies are useful for subsequent automated numerical analysis. One common use is for the training of "classifiers" as in Support Vector Machines (SVM) (Zhang et al., 2012) and Naïve Bayesian models (Oraby et al., 2015). Simply put, when a mathematical model is sufficiently trained from pre-coded examples in one of two categories, it can predict instances of future deception on the basis of numeric clustering and distances. The use of different clustering methods and distance functions between data points shape the accuracy of SVM (Strehl, Ghosh & Mooney, 2000), which invites new experimentation on the net effect of these variables. Naïve Bayes algorithms make classifications based on accumulated evidence of the correlation between a given variable (e.g., syntax) and the other variables present in the model (Mihalcea & Strapparava, 2009).

The classification of sentiment (Pang & Lee, 2008; Ott et al., 2013) is based on the underlying intuition-that deceivers use unintended emotional communication, judgment or evaluation of affective state (Hancock, Woodworth, & Porter, 2011). Likewise, syntactic patterns may be used in distinguishing feeling from fact-based arguments by associating learned patterns of argumentation style classes. In studies of business communication, performance is significantly better than a random guess by 16%, and the language of deceptive executives exhibits fewer non-extreme positive emotions (Larcker & Zakolyukina, 2012). Comparison between human judgement and SVM classifiers showed 86% performance accuracy on negative deceptive opinion spam (Ott et al., 2013). Fake negative reviewers over-produced negative emotion terms relative to the truthful reviews. These were deemed not the result of "leakage cues" from the emotional distress of lying, but exaggerations of the sentiment deceivers are trying to convey.

These linguistic approaches all rely on language usage and its analysis, and are promising when used in hybrid approaches. However, findings emerging from topic-specific studies (product reviews, business) may have limited generalizability towards real-time veracity detection of news.

NETWORK APPROACHES

Innovative and varied, using network properties and behavior are ways to complement content-based approaches that rely on deceptive language and leakage cues to predict deception. As real-time content on current events is increasingly proliferated through micro-blogging applications such as Twitter, deception analysis tools are all the more important.

Linked data

The use of knowledge networks may represent a significant step towards scalable computational fact-checking methods. For certain data, false "factual statements" can represent a form of deception since they can be extracted and examined alongside findable statements about the known world. This approach leverages an existing body of collective human knowledge to assess the truth of new statements. The method depends on querying existing knowledge networks, or publicly available structured data, such as DBpedia ontology, or the Google Relation Extraction Corpus (GREC).

The inherently structured data network of entities is connected through a predicate relationship. Fact checking can be effectively reduced to a simple network analysis problem: the computation of the simple shortest path (see Figure 1). Queries based on extracted fact statements are assigned semantic proximity as a function of the transitive relationship between subject and predicate via other nodes. The closer the nodes, the higher the likelihood that a particular subject-predicate-object statement is true.

There are several so-called 'network effect' variables that are exploited to derive truth probabilities (Ciampaglia et al., 2015), so the outlook for exploiting structured data repositories for fact-checking remains promising. From the short list of existing published work in this area, results using sample facts from four different subject areas range from 61% to 95%. Success was measured based on whether the machine was able to assign higher true values to true statements than to false ones (Ciampaglia, et al., 2015). A problem with this method, however, rests in the fact that statements must reside in a pre-existing knowledge base.

Social Network Behavior

Authentication of identity on social media is paramount to the notion of trust. The proliferation of news in the form of current events through mass technologies like micro-blogs invites ways of ascertaining the difference between fake and genuine content. Outside of the analysis of content comes the use of metadata and telltale behavior of questionable sources (Chu, Gianvecchio, Wang & Jajodia, 2010). The recent use of twitter in influencing political perceptions (Cook et al., 2013) is one scenario where certain data, namely the inclusion of hyperlinks or associated metadata, can be compiled to establish veracity assessments. Centering resonance analysis (CRA), a mode of network-based text analysis, represents the content of large sets of texts by identifying the most important words that link other words in the network. This was employed by Papacharissi & Oliviera to identify content patterns in posts about Egypt's elections (2012). Combining sentiment and behaviour studies have demonstrated the contention that sentiment-focused reviews from singleton contributors significantly affects online ranking (Wu, Greene, Smyth & Cunningham, 2010), and that this is an indicator of "shilling" or contributing fake reviews to artificially distort a ranking.

CONCLUSION

Linguistic and network-based approaches have shown high accuracy results in classification tasks within limited domains. This discussion drafts a basic typology of methods available for further refinement and evaluation, and provides a basis for the design of a comprehensive fake news detection tool. Techniques arising from disparate approaches may be utilized together in a hybrid system, whose features are summarized:

- Linguistic processing should be built on multiple layers from word/lexical analysis to highest discourse-level analysis for maximum performance.
- As a viable alternative to strictly content-based approaches, network behavior should be combined to incorporate the 'trust' dimension by identifying credible sources.
- Tools should be designed to augment human judgement, not replace it. Relations between machine output and methods should be transparent.
- Contributions in the form of publicly available gold standard datasets should be in linked data format to assist in up-to-date fact checking.

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

Ciampaglia, G., Shiralkar, P., Rocha, L., Bollen, J. Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks.

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.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is tweeting on Twitter: Human, Bot, or Cyborg? in the Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10, pp. 21-30.

Cook, D., Waugh, B., Abdipanab, M, Hashemi, O., Rahman, S. (2013). Twitter Deception and Influence: Issues of Identity, Slacktivism and Puppetry

de Marneffe, M., MacCartney, B. & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

Hancock, J., Woodworth, M. & Porter, S. (2011). Hungry like a wolf: A word pattern analysis of the language of psychopaths. *Legal and Criminological Psychology*. 113.

Hancock, J. & Markowitz, D. (2014). Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel. PLoS ONE,9(8)

Feng, V. & Hirst, G. (2013) Detecting deceptive opinion with profile compatibility.

Feng, S., Banerjee, R. & Choi, Y. (2012). Syntactic Stylometry for Deception Detection. *50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*, 171–175.

Larcker, D., Zakolyukina, A. (2012). Detecting Deceptive Discussions in Conference Calls. *Journal of Accounting Research*, 50(2), 495540.

Mihalcea, R. & Strapparava, C. (2009). The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. Proceedings of the ACL-IJCNLP Conference Short Papers, pp. 309–312,

Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M. & Whittaker, S. (2015). And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue

Ott, M., Cardie, C. & Hancock, J. (2013). Negative Deceptive Opinion Spam. *Proceedings of NAACLHLT.* pp. 497–501,

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), pp. 1–135.

Papacharissi, Z. & Oliveira, M. (2012).The Rhythms of News Storytelling on #Egypt. *Journal of Communication. 62.* pp. 266–282.

Rahangdale, A. & Agrawa, A. (2014). Information extraction using discourse analysis from newswires. *International Journal of Information Technology Convergence and Services*. 4(3), pp. 21-30.

Rubin, V., Conroy, N. & Chen, Y. (2015)^{A}. Towards News Verification: Deception Detection Methods for News Discourse. *Hawaii International Conference on System Sciences.*

Rubin, V. L. Chen, Y.,& Conroy, N. J. (2015)^{B}. Deception Detection for News: Three Types of Fakes. In *The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015)*, Nov. 6-10, St. Louis.

Rubin, V. & Lukoianova, T. (2014). Truth and deception at the rhetorical structure level. *Journal of the American Society for Information Science and Technology, 66*(5).DOI: 10.1002/asi. 23216 ·

Strehl, A. Ghosh, J. & Mooney, R. (2000). Impact of Similarity Measures on Web-page Clustering. *AAAI Technical Report WS-00-01.*

Wu, G., Greene, D. Smyth, B. & Cunningham P. (2010). Distortion as a Validation Criterion in the Identification of Suspicious Reviews. *1st Workshop on Social Media Analytics.*

Zhang, H., Fan, Z., Zeng, J. & Liu, Q. (2012). An Improving Deception Detection Method in Computer-Mediated Communication. *Journal of Networks*, *7* (11),