Some Considerations on Fake News Detection

Diana HORNOIU
Ovidius University of Constanta

Abstract
The paper addresses the phenomenon of fake news. First, it provides an overview of definitions of fake news in recent research, with special focus on three categories of fake news, depending on the intent behind falsification: a) fabrication, b) hoaxing and c) satire. Second, it looks into the methodological issues related to gathering a relevant corpus for fake news detection, highlighting the conditions such a corpus is supposed to meet. Last but not least, the paper argues for a model of analysis for the detection of fake news within the framework of Natural language processing.

Keywords: *deception detection; fake news; fabrication; hoax; satire; fact-checking; natural language processing*

1. Introduction

We live in the midst of the "fake news era". This problem can hardly be settled by relying on writer's honesty and integrity and/or on readers' critical thinking and determination to verify everything they read with multiple sources.

Media deception, whether it takes the form of fake news, phony press releases and hoaxes, is notoriously misleading and even harmful, especially when they are taken out of their original contexts. To make matters worse, traditional barriers to publishing content have by and large disappeared and so have some of the traditional quality control procedures. Basic journalistic principles like source verification, fact checking and accountability can be easily bypassed or simply ignored by even by some newspapers, not to mention individuals and organizations publishing content on various social media networks, such as Facebook, Twitter, etc.

The impact of this situation has led to the emergence of terms such as "trolls", "fake news", "post-truth media" and "alternative facts" to describe this state of affairs. There is evidence that these developments have far-reaching consequences which are far from being harmless and which may have a significant impact on real-world events, as illustrated by Allcott and Gentzkow's (2017) study on the role of social media in the 2016 US presidential election, and by a study on the mystifications misinformation, and disinformation spread over social media by the anti-vaccine movement (Broniatowski *et al*., 2018).

The rest of this paper is divided into five sections. Section 2 provides an overview of definitions of fake news. Section 3 reviews the types of fake news, Section 4 looks into the methodological issues related to gathering a relevant corpus for fake news detection. Section 5 outlines a theoretical framework for the detection of fake news.

2. Definitions of Fake News

According to (Elliot and Culver, 1992), 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".

In a narrow sense, fake news is defined as news articles that are intentionally and verifiably false and can mislead readers. Authenticity and intent are thus the key features of fake news under the narrow interpretation (Conroy, Rubin, and Chen, 2015; Klein and Wueller, 2017). First, fake news includes false information that can be verified and proved as such. Second, fake news is created with the dishonest intention of misleading readers. Broader definitions of fake news focus on either authenticity or intent.

On a different approach, satire news is regarded as fake news due to its false contents and despite its entertainment-oriented nature and acknowledged deceptiveness (Rubin, Conroy, Chen, and Cornwell, 2016). Still others treat deceptive news as fake news including in this larger category fabrications, hoaxes and satire (Rubin, Chen, and Conroy, 2015).

3. Types of fake news

Among journalists, the responsibility for knowing what is true rests with news consumers. In this context, Kovach & Rosentiel (2010: 7) argue that this shift in responsibility could signal the end of journalism pointing to "a world without editors, of unfettered spin, where the loudest or most agreeable voice wins and where truth is the first casualty". According to Rubin, Conroy and Chen (2015) "few news verification mechanisms currently exist, and the sheer volume of the information requires novel automated approaches".

There are various types of fake news depending on the intent behind falsification. Regardless of the category, deceptive news tends to build narratives rather than report facts. In what follows, we briefly address the three types of fake news, contrasting each type to genuine reporting: a) fabrication, b) hoaxing and c) satire.

3.1 Fabrications

Fabrications are an extreme kind of disinformation which reports what is blatantly false. Thus fabrications deliberately deceive readers or promote a biased agenda. They include post generated and distributed on social media from propaganda and the so-called clickbait ("eye-catching" headlines) accounts. The intent behind propaganda and clickbait varies from opinion manipulation to attention redirection and increasing traffic on social media. Exposed fraudulent journalistic writing, discussed in Compton and Benedetti (2015) or Shingler (2015), are ideal for a fake news corpus.

Yellow press and *tabloids* are an appropriate source for fake news corpus since they present a wide range of unverified news using eye-catching headlines

("clickbaits"), exaggerations, scandal-mongering, or sensationalism to increase traffic or profits.

3.2 Hoaxes

Hoaxing is another type of disinformation that deliberately deceives the reader (Tambuscio et al., 2015; Kumar et al., 2016) present in both the mainstream or social media. Created with the intent of going viral, hoaxes masquerade as genuine news. They can be picked up and mistakenly validated by traditional news outlets. Brunvand (1998) draw a distinction between hoaxing and *pranking* or *practical joking* arguing that the former can be characterized 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).

3.3 News Satire/Humorous Fakes

News satire or news parody (e.g., *The Onion* and *CBC's This is That)* is a specific genre 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). Thus, in news satire, the writer's primary intent is not to mislead the reader, but rather to criticize or entertain (Conroy et al., 2015). However, Rubin et al. (2015) point out the harmful nature of news satire or hoaxes when they are taken out of context.

A distinction should be drawn between fabricated news and news satire. As long as news consumers are aware of the humorous intended meaning, they may no longer take the information literally and interpret it at face value. Technology can identify news satire and display originating sources (e.g., *The Onion*) to alert users especially it is decontextualized on news platforms.

4. Data Collection Practices in Deception Detection for news

Requirements for Fake News Detection Corpus in Natural Language Processing (NLP)

Recent research (Rubin, Chen, and Conroy, 2015; Rubin, Conroy, Chen, and Cornwell, 2016) has shown that a corpus of empirical data relevant for fake news detection should meet the following conditions:

1. *Availability of both truthful and deceptive news*. The corpus should include both authentic genuine news and their fake counterparts in order for the machine to be able to find patterns and regularities.
2. *Digital textual format accessibility*. The preferred medium in NLP is text. Thus, it is mandative that audio and video data be transcribed.
3. *Verifiability of "ground truth"*. When collecting a corpus for fake news detection, the question that arises is what constitutes verification and how does one decide whether the news is genuine or fabricated. To answer the question one may

rely on news sources that are based on a system of "checks and balances". Such news sources qualify as appropriate as corpora since they have withstood the test of time.

4. *Homogeneity in length.* The dataset should be homogeneous in terms of length for individual news articles since this will make the news items comparable. For instance, a one-paragraph summary on Facebook, a short tweet with a headline and a lengthy op-ed article do not qualify for comparable news items.

5. *Homogeneity in style/ writing matter.* A corpus designed for NLP applications for fake news detection should be aligned along news genres (e.g. editorials, op-ed articles, breaking news) and topics (science, health, politics, business). Moreover, the articles should be written by similar types of authors. Meeting these requirements ensures the items are comparable, the comparison being made across news outlets.

6. *Predefined timeframe.* A collection of breaking daily news has been shown to be more relevant and have more variation than a collection of the news on a particular topic over an extended period of time (Rubin, Chen and Conroy 2015).

7. *The manner of news delivery* (e.g., *humor; newsworthiness; absurdity; sensationalism*). The manner of delivery is instrumental in creating context for interpretation. For instance, "truth-biased" readers may be expected to shift to a "lie-biased" perspective when reading news satire.

8. *Pragmatic concerns* Data collection is influenced by various external factors such as copy-right-related costs, public availability, ease of accessibility, suitable overall volume of data, and writers' privacy.

9. *Consideration given to language and cultural specificity* (Rubin, 2014). Research on fake news detection has mainly focused on English disregarding other languages, with few notable exceptions explored and reported in deception research (e.g., Spanish, Italian, Mandarin). Thus it is essential that language and culture specificity should be taken into account when addressing the phenomenon of fake news.

5. Fact checking

Fact checking is defined as the task of assessing the truthfulness of a claim made by a public figure in a particular context. Under this definition, fact checking appears to be a binary classification task. However, it is often the case that statements are not completely true or false. For example, the claim in (1) is has been assessed as "mostly true" because some of the sources dispute it.

(1)
Claim (by President Barack Obama): "For the first time in over a decade, business leaders around the world have declared that China is no longer the world's No. 1 place to invest; America is."
Verdict: MOSTLY TRUE (by Politifact)

"The president is accurate by citing one particular study, and that study did ask business leaders what they thought about investing in the United States. A broader look at other rankings doesn't make the United States seem like such a powerhouse, even if it does still best China in some lists."
(Vlachos and Riedel 2014)

On the other hand, for the claim in (2) the statistics can be manipulated to support or disprove it as desired.

(2)
Claim (by Chancellor George Osborne): "Real household disposable income is rising."
Verdict: HALF TRUE (by Channel 4 Fact Check)
"RHDI did grow in latest period we know about (the second quarter of 2013), making Mr Osborne arguably right to say that it is rising as we speak. But over the last two quarters we know about, income was down 0.1 per cent. If you want to compare the latest four quarters of data with the previous four, there was a fall in household income, making the chancellor wrong. But if you compare the latest full year of results, 2012, with 2011, income is up and he's right again."
https://www.channel4.com/news/factcheck

Thus, according to Frank and Hall (2001) fact checking should be viewed as an ordinal classification task in order to capture all its nuances.

5.1 Manual fact-checking

Research (Vlachos and Riedel 2014; Potthast, Kiesel, Reinartz, Bevendorff and Stein 2018) has shown that conceptualizing news in terms of a binary distinction is hardly feasible since any piece of fake news is not entirely false. Conversely, pieces of real news may not be entirely flawless.

Thus, there is a tendency among journalists working on the manual fact-checks of news articles to rate news as "mostly true," "mixture of true and false" or "mostly false". Opinion-driven posts which lack a factual claim are rated as "no factual content." The ratings "true and false" and "mostly false" have to be accounted for and when a piece of news raises doubts regarding the rating a second opinion is required. Disagreements are resolved on the basis of a third opinion. All news articles rated "mostly false" undergo a final check to ensure the rating is justified.

The journalists working on the manual fact-checks of news articles use the following as guidelines for rating the articles:

(a) Mostly true. The news article or the post does not include unsupported speculation or claims. This rating is used for news articles and any related links or images which are based on factual information and portray it accurately. The

authors may offer a personal interpretation of the event as long as the events, numbers, quotes, reactions, etc., are not misrepresented or made up in any way.
(b) Mixture of true and false (mix, for short). This rating applies to news articles or posts including unfounded claims mixed with real events, numbers, quotes, etc. It also applies to news articles or posts whose headline makes a false claim even when the text of the story is largely accurate. However, it is important to point out that it is only used on condition that the unsupported or false information be roughly equal to the accurate information in the post or link.
(c) Mostly false. This rating is used when most or all the information in the news article is inaccurate. It also applies to a post whose central claim is proved false.
(d) No factual content (n/a, for short). This rating is reserved for any type of news articles that are based on unconfirmed information. Such items of news may include pure opinion posts, comics and satire that do not make a factual claim.

5. Fake news detection and natural language processing

Natural language processing could prove useful for fake news detection. This approach enables the researcher to develop an algorithm for fake news detection (Feng and Hirst, 2013; Markowitz and Hancock, 2014; Ruchansky et al. 2017). The NLP framework of analysis include the following stages: collecting a corpus of both fake news and real news; feeding the corpus to the machine; building an algorithm to parse sentence structure; training the algorithm on the text itself to distinguish between fake news and real news on the basis of specific patterns or linguistic cues.

Recent research (Bachenko et al., 2008; Larcker and Zakolyukina, 2012) has demonstrated the effectiveness of linguistic cue identification, as the language of real news is known to differ from that of fake news. The analysis of empirical data has shown that fake news articles are rich in lexical items and phrases referring to feelings or senses (e.g., seeing, touching), negative emotion words as well as other-oriented pronouns as opposed to self-oriented pronouns. Similarly, fake news articles have been shown to have lower cognitive complexity.

On the other hand, the linguistic indicators of fake news across different types of fake news and across different media platforms are still challenging and less understood. Each of the types of fake news discussed in section 3 has its own potential textual indicators (Rubin 2015).

The manual fact checking process is an approach that decomposes the task into the following stages: (1) extracting statements to be fact-checked; (2) constructing appropriate questions; (3) obtaining the answers from relevant sources; (4) reaching a verdict using based on the answers obtained.

Natural language processing offers a theoretical framework well-suited for the stages of fact-checking. Approaches similar to those proposed for speculation detection (Farkas et al., 2010) and veridicality assessment (de Marneffe et al., 2012) can be applied for statement extraction. Semantic parsing can offer the

solution for the task of obtaining answers to questions from databases. Compiling the answers into a verdict could be approached in a way similar to logic-based textual entailment (Bos and Markert, 2005).

6. Conclusions

The task of fake news detection may be separated into three, according to the type of fake: a) fabrications (uncovered in mainstream or yellow press or tabloids); b) hoaxes; c) humorous fakes (news satire, parody). Serious fabricated news may require considerable effort to collect, case by case. Authors of fabrications are likely to use cues of deception similar to "verbal leakages" in other contexts (such as law enforcement or computer-mediated-communication) in order to avoid the consequences for dishonest reporting. Hoaxes are creative, unique, and often multi-platform. Consequently, this type of fake news requires detection methods beyond text analytics (e.g., network analysis). With regard to humorous news, their entertaining or mocking nature may interfere with binary text classification (real vs. fake news), especially if the algorithm mistakes cues of sensationalism, or humor for cues for deception.

The nine requirements discussed in section 4 indicate that an algorithm built to detect fake news should also detect non-fake news and account for factors such as developing news and language and cultural interpretations. Using linguistic cues for deception detection in news articles is not only laborious but also topic/media dependent, resulting in the limitation of the scalability of these solutions.

Works Cited

Allcott, Hunt and Gentzkow, Matthew. "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31(2):211–36, 2017.

Bachenko, Joan, et al. "Verification and implementation of language-based deception indicators in civil and criminal narratives." *Proceedings of the 22nd International Conference on Computational Linguistics*-Volume 1, 41–48. Association for Computational Linguistics, 2008.

Bos, Johan and Markert, Katja. "Recognising textual entailment with logical inference." *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2005), 628–635, 2005.

Broniatowski *et al*. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate." *American Journal of Public Health*, 2018.

Brunvand, Jan Harold. *American Folklore: An Encyclopedia*. Taylor & Francis, 1998.

Compton, James R. and Benedetti, Paul. "News, Lies and Videotape: The Legitimation Crisis in Journalism." Rabble. Retrieved from

http://rabble.ca/news/2015/03/news-lies-and-videotape-legitimation-crisis-journalism, 2015.

Conroy, Niall J, Rubin, Victoria L and Chen, Yimin. "Automatic deception detection: Methods for finding fake news." *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

Elliot, Deni., and Culver, Charles. "Defining and analyzing journalistic deception." *Journal of Mass Media Ethics*, 7(2), 69–84, 1992.

Farkas, Richard et al. "The CoNLL 2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text." *Proceedings of the Fourteenth Conference on Computational Natural Language Learning: Shared Task*, 1–12, Uppsala, Sweden, 2010.

Feng, Vanessa Wei and Hirst, Graeme. "Detecting deceptive opinions with profile compatibility." Ruslan Mitkov, Jong C. Park Ed. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 338–346, Nagoya, Japan, 14-18 October 2013. https://www.aclweb.org/anthology/I13-1.pdf

Klein, David and Wueller, Joshua. "Fake News: A Legal Perspective." *Journal of Internet Law*, 20 (10), 2017 Available at SSRN: https://ssrn.com/abstract=2958790

Kovach, Bill and Rosenstiel, Tom. *Blur: How to Know What's True in the Age of Information Overload*, New York, Bloomsbury, 2010.

Kumar, Srijan et al. "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes." *Proceedings of the 25th international conference on World Wide Web*, 591–602. International World Wide Web Conferences Steering Committee, 2016.

Larcker, David F and Zakolyukina, Anastasia A. "Detecting deceptive discussions in conference calls." *Journal of Accounting Research*, 50(2):495–540, 2012.

Markowitz, David M. and Hancock, Jeffrey T. 2014. "Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel." *PLoS ONE* 9(8), 2012, e105937.

de Marneffe, Marie-Catherine, et. al. "Did it happen? The pragmatic complexity of veridicality assessment." *Computational Linguistics*, 38(2):301–333, 2012.

Potthast, Kiesel, Reinartz, Bevendorff and Stein. "A Stylometric Inquiry into Hyperpartisan and FakeNews." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 231–240 Melbourne, Australia, July 15 - 20, 2018.

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

Rubin, Victoria L, Conroy, Niall J., Chen, Yimin, and Cornwell, Sarah. "Fake news or truth? Using satirical cues to detect potentially misleading news." *Proceedings of NAACL-HLT*, pp. 7–17, 2016.

Ruchansky, Natali et al. "Csi: A hybrid deep model for fake news." *arXiv* preprint arXiv:1703.06959, 2017.

Shingler, B. 2015. "François Bugingo, Foreign Correspondent, Suspended by Media Outlets." CBC News Montreal.

Tambuscio et al. "Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks." *Proceedings of the 24th International Conference on World Wide Web*, 977-982, Florence, 2015.

Vlachos, Andreas and Riedel, Sebastian. "Fact Checking: Task definition and dataset construction." *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 18–22, Baltimore, Maryland, USA, June 26, 2014.

https://www.channel4.com/news/factcheck