Diagnosis, Prevention, and Cure for Misinformation

Ritwik Banerjee
Computer Science
Stony Brook University
New York, USA
rbanerjee@cs.stonybrook.edu

Indrakshi Ray
Computer Science
Colorado State University
Fort Collins, Colodaro
indrakshi.ray@colostate.edu

2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI) | 978-1-6654-1621-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/COGMI52975.2021.00028

***Abstract*—Misinformation has become a widespread problem in contemporary society, harming the most vulnerable sections in particular. It incurs a high cost on everyone – socially, politically, and even financially. Much of the progress made in tackling misinformation, however, has only offered relatively simple solutions by taking a narrow view of "misinformation" itself. We discuss the various nuances of the term, illustrating the technical difficulties that arise from them, and propose a multiplicity of context-sensitive modeling approaches that may prove to be fruitful in addressing these difficulties. A tremendous amount of work remains to be done to ensure our inoculation from misinformation and its harmful effects, and much of this work, we argue, requires collaborative effort across disciplines. It is our hope that this article serves as a call to further such research in this field.**

***Index Terms*—misinformation, deception, linguistics, multi-modal learning, natural language processing**

I. Introduction

Traditional media, and in more recent years, social media, have often propagated incorrect – or at the very least, inaccurate – information. While the attention of quantitative research on this problem is relatively recent, the phenomenon itself has been a social and political concern for centuries.

A. Let them eat cake: the history and advent of misinformation

In 1522, there was an attempt to manipulate the pontifical election by sonnets ridiculing several candidates. These sonnets were pasted for public viewing on the sculpted figure known as Pasquino, in Rome [1]. To this day, "pasquinade" thus means a publicly viewable writing intended to ridicule or calumniate a person. Deliberate acts of propagating falsehood for political manipulation would be seen again during the French Revolution, when the now-famous quote, "let them eat cake", was attributed to Marie Antoinette to add fuel to the popular hatred and outrage against her[1]. As technology changed the nature of mass communication, so did change the nature of propagation of falsehood. These include examples of hoax spread intentionally through newspapers, as well as accidental errors [3], [4]. With the advent of the Internet, the reach as well as the rate of diffusion of false information has, of course, increased dramatically. Such false content ranges from absurd alternative narratives to statements based on emotional appeal. Since prominent stakeholders in many spheres use these conduits to influence the general public, we find ourselves in a 'post-truth' twilight zone where facts are often difficult to distill from a pervasive onslaught of propaganda and counter-propaganda.

The importance of identifying false information has been widely recognized in many distinct but increasingly interconnected fields, including journalism, sociology, computational linguistics, and computer vision. Given the sheer amount of data available in the modern world, computational approaches have been at the forefront of tackling this issue. These include lie detection based on linguistic style, rumor and hoax identification using social network structures, and multi-modal analyses that combine textual and visual information [5]–[9].

B. True, false, both, neither, and other pesky details

Except for a few rare instances, empirical approaches have focused only on distinguishing truth from falsehood. Qualitative work, on the other hand, has delved deeper into the nature of the problem in its modern form, painstakingly analyzing its subtleties and its terminology – some of which entered the English language as recently as the late 1980s[2]. This body of work has always maintained that this is a nuanced and complex problem, going far beyond "prototypical instances" of lying [11].

Debates in information science and philosophy demonstrate that many terms (for example, "disinformation") are often poorly understood, and even eminent dictionaries can offer improper definitions [12, p. 409]. Delving into this large body of qualitative research [12]–[16] offers two critical insights:

1) a majority of the empirical work on disinformation or misinformation have instead focused on the general idea of deception, and
2) subtle forms of falsehood such as "spin" or "half-truth" are more dangerous and prevalent, but have received little attention in empirical work.

To further compound the inadequacies of our current outlook toward the spread of falsehood through traditional or social media, notions such as "bias" (in the form of, among other things, selection or omission), or even the role of individual

This work was supported in part by NSF under awards IIS 2027750 and SES 1834597.

[1]The origin can be traced to the philosopher Jean-Jacques Rousseau including this as an anecdote in his "Confessions" in 1766, attributing it to a "great princess" [2]. Marie-Antoinette was ten years old at the time.

[2]For instance, "disinformation", the translation of the Russian word *dezinformatsiya*, was allegedly coined by Joseph Stalin after WW-II [10].

978-1-6654-1621-4/21/$31.00 ©2021 IEEE
DOI 10.1109/CogMI52975.2021.00028

memory in the recollection of events, indicate that the scope of this problem – *i.e.*, detecting falsehoods – may in itself be somewhat nebulous. The goal of this paper is twofold. First, to underscore the gap between the current computational approaches and the problem of misinformation, and second, to offer a vision of a collective research agenda aimed at bridging this gap, based on collaborative efforts that cross the boundaries of individual disciplines.

We begin by attempting to define a more precise scope of the problem, in Section II, before discussing related contemporary computational research in Section III. We then provide our vision for future research in this direction (Section IV) before concluding.

II. Scope and Definition

Throughout human history, a long lineage of thinkers and philosophers – from Aristotle to St. Augustine and St. Thomas Aquinus to Immanuelle Kant – considered deception, *i.e.*, the intentional assertion of a false statement, to be categorically wrong. Deception, to this *absolutist* school of thought, is always an immoral act. Few subscribe to this view of deception, having given way to the *utilitarian* perspective instead. In this view, deception is morally permissible if and only if there is no other option available to the agent "that would result in better balance of good consequences relative to bad consequences" [17, p. 146][3]. Questions naturally arise, such as who decides what constitutes a better balance, or if such a decision is in itself permissible. To have a defensible notion of the problem definition, such questions must remain beyond the scope of computational and empirical scientists. Thus, for the purpose of this paper, any assertion of a factually incorrect statement will be treated as impermissible.

No matter where an individual's perspective on deception resides, there is nearly universal acceptance that the most egregious deception comprises three factors: *intent* – the communicator intends to propagate the falsehood; *literal meaning* – the statement is false in its literal meaning; and *effect* – the statement will very likely cause the listener to believe in something untrue. *Disinformation* – a term now in common use within contemporary related literature – is this form of deception, where we find a confluence of all three. On the other hand, *misinformation* is commonly understood as unintentional lying, where the literal meaning and/or effect are present, but the communicator is not deliberately propagating the falsehood. A large body of empirical work does not address the distinction between these two terms. Perhaps more dangerous, though, is the absence of empirical work about the more nuanced aspects of deception, which lie somewhere between disinformation and misinformation, or combines features of both. As a mundane illustration, let us consider the following:

1) Mary is a painter. Her friend John knows Mary is a painter.
2) John sees a new painting in Mary's living room, and says, "This is beautiful. Perhaps your best work."
3) The painting happens to be store-bought, but Mary responds, "Thank you."

Mary did not utter anything *literally* false, but the intent is clearly deceptive, and the effect makes John believe something untrue. Such examples are numerous, and have been characterized as selection bias, whitewashing, slanting, distorting, etc. Schauer and Zeckhauser call these things "less than lying" [16, p. 39], but point out that they are perhaps even more harmful than claims that are literally false. Especially because these are significantly more pervasive, and have been for many years before computational research on fake news and deception garnered attention. Due to the ubiquity of this phenomenon in traditional and social media, and the potential harm it causes to society, we include these forms of deception within the term "misinformation". The urgent social need, consequently, is to identify such misinformation, prevent its propagation, and finally – since the prevention is unlikely to be complete – repair, as far as possible, the damage caused by it. Before diving into these three components of research, we discuss some noteworthy related work in related areas, in order to establish the context of our vision.

[3]This is distinct from the school of *ethical egoism*, which views deception as the right act if it maximally promotes one's own welfare.

III. Related Work

The most common form of research in the recent years has investigated the "fake news" phenomenon. At its core, this body of work culminates in a binary classification (*true* vs *fake*) or regression problem (scoring the truthfulness in an interval). Being a largely supervised learning problem, there is an explicit reliance on knowledge bases, which are used to fact-check against a repository of assertions that are *a priori* known to be true. These knowledge bases are usually curated by experts, and are the output of a manual fact-checking process. A large number of automated fact-checking approaches use this curated output (*e.g.*, `PolitiFact` or `Snopes`) to train their machine learning models, while others have investigated crowd-sourced fact-checking [18]–[20]. There are several noteworthy advances made in this vertical, as detailed in recent surveys on fact-checking (Li et al. [21], among others). Similar data-driven approaches have also led to advances in the detection of rumor and hoax [7], [22].

A. Fact-checking and credibility: epistemological concerns

The fact-checking approach, however, has been criticized by others who have raised epistemological questions about the fact-checking process and the fact-checkers. This body of work contends that events or claims may not always be universally acceptable as true of false, and even when they are, simply having good inter-rater agreement is not a measure of ground-truth [23]–[25]. Others have used the frequentist notion of probability to model credibility and trustworthiness of sources [26]–[28], but as Weikum [29] remarks, this brings forth another set of potential flaws. Moreover, credibility has either been taken at face value, or is based on the assumption that credible news articles employ objective language. While this is evidently true for conspicuous propaganda from sources

like DC Gazette (see Rashkin et al. [30]), it rarely holds true when facing sophisticated forms of deception.

For instance, whether or not RT or Xinhua are good candidates for ground-truth is less obvious: criticism from organizations such as the Poynter Institute for Media Studies speak of "selection bias", while others have argued that they may act as sources of propaganda [31], [32]. Equally worrisome are indications that trusted sources like the New York Times (NYT) have published articles with selection bias along with claims, only to refute those claims later [33]. It is worth noting that some earlier computational approaches in automated fake news detection have treated articles from organizations like Xinhua and NYT as universally true [30].

As various instances demonstrate, data-driven approaches much be extremely cautious before arriving at conclusions based on empirical evidence, since the nature of trustworthy evidence is not always obvious in this work. Communication that appears objective, for instance, cannot be assumed as *prima facie* credible. Indeed, in some data-driven approaches, it has been observed that creators of misinformation attempt to imitate the communication style (including the use of objective language) of truthful agents [8, p. 600].

As the idea of misinformation draws from a large body of work in psychology, philosophy, information science, communication theory, and linguistics, we must necessarily pay attention to developments in these fields in conjunction with the recent developments in computational approaches (*e.g.*, the use of deep neural networks to train increasingly sophisticated discriminative models).

B. Pragmatics in communication

In various modes of communication, be it textual or visual, literal falsehood is not by itself an indicator of misinformation. As such, misinformation must be distinguished from pragmatic[4] communication constructs such as the use of metaphors, sarcasm, irony, satire, or humor. As social media has become an important conduit for misinformation, it has also seen extensive use of these constructs, be it in Tweets, memes, comments on platforms such as YouTube, or relatively more extensive discussions on platforms like Reddit.

Clearly, in order to distinguish such pragmatic constructs from misinformation, one must accurately identify them. Much of the work in this direction focuses on textual data [34]–[37]. More recently, however, significant advances have been made toward recognizing these communication constructs in multimodal data that combines two or more of audio, image, and text [38]–[41].

Detection of figurative language (and similarly, of images), is a difficult problem. Even more so when there is a need to separate it from misinformation, because such constructs (*e.g.*, sarcasm or irony) may completely decouple – and even contrast – the communicator's intent from the communicated content [42], rendering it indistinguishable from misinformation unless empirical methods can aptly model the context surrounding a specific communication.

[4]We use the term *pragmatics* in the sense of its definition in linguistics: the study of communication in a specific context, and how context contributes to, or provides, meaning to the communication.

C. Cognitive load and deception

In parallel to, and perhaps complementing, the computational detection of pragmatics in language, a sizeable body of work in psychology has studied the interplay between communication, cognitive load, and deception. The connection to cognitive load is particularly important when the deception is intentional. This was noted in several studies, which demonstrated the deceiver frequently exhibiting behavior belying the content of the communication. This body of work provides important linguistic as well as paralinguistic cues of misinformation [5], [43]–[45]. Depaulo et al. [43] and Vrij et al. [46] also observe that on one hand, the increased cognitive burden in intentional deception produces factual inconsistency in a narrative, but on the other hand, the recipient of the communication feels that the communication is rehearsed, when such inconsistencies do not exist. Perhaps not surprisingly, these findings are consistent with observations made by criminal investigators outside the world of academic research.

A clear signal from this body of work is thus the need to model longer narratives, providing a temporal view that can distill any inherent inconsistencies, instead of trying to identify the truth or falsity of individual claims in isolation.

D. Perception

While discussing the epistemological concerns with fact-checking and credibility-based diagnosis of misinformation (Section III-A), it becomes clear that even scientists may be susceptible to misinformation. At least in part, this is because *perceived* credibility has an immense impact on how much the recipient believes the message [47]. Markers of credibility have thus been studied across various disciplines. For electronic media, and text in particular, the *perception* of credibility is based largely on two factors: the authority of the author, and referrals to texts from credible external sources [48]–[51]. The perception of credibility thus has a cyclical dependency on ones prior perception of credibility. Intuitively, this leads to a network effect. Indeed, this network effect has been observed in studies on disinformation [52], and the role of perception in the formation and propagation of misinformation in networks has been investigated by some (*e.g.*, [53]). But significant gaps remain, which we discuss next in Section IV.

Our perception, of course, is closely related to the stance we take on an issue or even another person, whether while consuming information, creating it, or propagating it (see, for example, Manusov [54] and Mingkun et al. [55]). The connection between stance and veracity has been explored in some recent work. Particularly interesting is the approach taken by Jin et al. [56], who contrast conflicting viewpoints from social media in their attempt to distill the truth.

IV. A Vision for the Future

In spite of the rapid advances made in multiple directions in our attempts to tackle misinformation, the problem has proven to be far more complex than a straight-forward issue of categorizing a piece of information as either *true* or *false*. The gaps that remain are due in part to the technical difficulty of modeling human communication precisely enough to perform such a categorization, but even more to the nuances of misinformation that have not yet been incorporated into any data-driven empirical approach. Our research agenda is motivated by identifying these gaps, modeling them, and thereby providing a far more universal empirical picture of (i) what constitutes misinformation, (ii) how to prevent it, and (iii) how to repair the damage it causes.

If there is a single fundamental concept whose introduction can plug most, if not all, the gaps we find in this field, it is correct modeling of **context**. Context, however, is a complex notion by itself. Based on what we glean from the current literature, it can be interpreted and modeled in five primary ways: temporal, linguistic and pragmatic, extralinguistic, domain-specific, and cross-genre. These are broad categories, and not always completely distinct from each other, but we believe this categorization nevertheless manages to provide an overview of the research agenda.

A. Temporal context

As discussed in Section III-C, there is a need to model longer narratives that have a temporal span, so that events that are created by the proverbial "spin doctors" or propagandists can be modeled in ways that can distill inconsistencies in the narrative. Arguably, an inconsistent narrative can be spotted by identifying semantic changes. While the precise nature of modeling such a change will vary on the modality and the length of the time scale, our view is that such models can all be based on the dominant philosophical definition of semantic change, as proposed by the classical work by Charlie Broad [57]. This is known as the *successive view*, and defines semantic change in terms of juxtapositions. This notion has been implicitly adopted in recent work that studies diachronic changes in word meanings [58]–[60]. In a similar vein, juxtapositions within the scope of a single narrative may also be constructed to identify semantic changes of the important aspects of an event of claim.

B. Linguistic and pragmatic context

Since pragmatics see pervasive usage in social media, it is crucial that we develop technology capable of accurately modeling figurative language (textual or visual) use and distinguishing it from misinformation. Some advances have already been made in the detection of such linguistic constructs, such as the detection of irony and sarcasm in Twitter responses while modeling the entire thread of Tweets as the appropriate context [34], the generation of sarcastic responses to visual prompts, or the detection of figurative language use on the basis of incongruous linguistic contexts [59], [61], [62]. The datasets used in this body of work, however, were relatively simplistic in the sense that figurative language use was not mixed with non-figurative misinformation. Therefore, it remains to be seen whether ideas such as incongruous linguistic contexts need to be refined further, and if so, in what ways.

Pragmatics also feature by means of implicature. That is, when a meaning is conveyed without its explicit mention in the communication. More often than not, this is due to correlation being presented while causation gets implied. To illustrate with an example:

> **News Headline:** *Kanika Kapoor met Prince Charles before he tested positive for COVID-19.*
> **Implication:** Kanika Kapoor infected Prince Charles (or vice versa).

The above implication did, indeed, find its way into social media, and hundreds of Tweets were devoted to a discussion of which one of the two personalities was responsible for infecting the other. The news article, however, never literally conveyed causation.

C. Extralinguistic context

Communication through textual and visual language almost always takes place within a larger structure. In social media, this structure is the social network itself, which offers extra-linguistic features such as the metadata of the users engaged in the communication and properties (local and glocal) of the network graph. While some work has been done that connects these features to disinformation and political stance detection, a comprehensive incorporation of these features into the study of misinformation has aroused relatively less interest.

The modeling of extra-linguistic context in social networks could, however, be largely due to the lack of availability of adequate data to model the properties of the graph. Some social networks offer extremely limited amounts of information through their free or relatively cheap API services (while costing significantly more otherwise, as is the case with enterprise APIs provided by Twitter), but others do not provide such APIs at all. Due to privacy concerns, these restrictions may be fully justified, and therefore, progress along this line of research may only be possible under collaboration with social media enterprises.

Extra-linguistic context may exist outside of social networks as well. Markers of credibility, for example, could be present through visual cues such as the design of a website, the mention of credible sources in a piece of text, and other such cues. Not enough attention has been paid yet to such cues vis-à-vis misinformation, with some very recent exceptions [63].

D. Domain-specific and cross-genre context

Even though misinformation was at first largely studied as a political phenomenon, the large-scale misinformation that accompanied the COVID-19 pandemic resulted in a lot of attention toward medical misinformation. Albeit just one example of domain-specific misinformation, it is a crucially important one. Thus, we can view the medical domain as an exemplar for modeling domain-specific context as an important aspect of identifying misinformation in that specific

domain. Literally false assertions exist, of course, and must be identified with urgency. This has been done with some success during the pandemic.

In such specialized domains, however, a more complex and subtle problem exists that pertains to misinformation. This is the distortion of medical facts, as opposed to outright false claims. Manual assessment of this phenomenon has a long history, and has consistently shown that distortions are commonplace [64]–[67]. One of the earliest such studies found that according to scientists whose work was being presented or discussed in news, only 8.8% of the articles had no errors. Several other analyses found that distortions, exaggerated claims, overstatement of risks, sensationalism, and other types of partial falsehoods, were regularly present in medical news.

The modeling of such distortions in a specialized domain requires collaborative efforts between the medical community and the empirical researchers who work in machine learning and related fields. To some extent, however, computer science researchers can avail resources generated by medical professionals and healthcare journalists and leverage them to build supervised learning systems to model domain-specific knowledge. There are a few initial steps taken in this direction, which indicate that domain-specific context may need to be modeled by looking across multiple genres [63], [68].

In the domain of healthcare and medicine, for instance, new findings are first communicated through research publications, some of which are then propagated through news articles. Thus, fact-checking – whether automated, manual with expert help, or semi-manual with human-in-the-loop mechanisms – necessitates verifying information present in two genres that communicate information using vastly different vocabularies. Furthermore, when an expert rater considers some assertion as misinformation, it may not be obvious to a lay person why it is so. This may be within the grasp of immediate future research, however. Machine learning and data-driven research on fact-checking has used knowledge-bases, (see Section III). But, due to the narrow focus on binary classification or regression, little attention has been paid to the more detailed expert analysis provided by these knowledge bases. At the simpler end of such analysis lies the identification of *check-worthy* content [69], while at the other end of the spectrum lies deeper and more subtle nuances of misinformation such as the use of sensational language, or the use of low-quality evidence to market a medical intervention. The former has received considerable attention in recent years, but the latter needs further study. One notable exception is the recent work investigating various qualitative aspects of health information [70].

E. Diagnosis and prevention of misinformation

Diagnosing misinformation demands technology capable of highly accurate detection of these nuanced forms of misinformation that go beyond the traditional binary classification. We conjecture that the incorporation of contexts – broadly along the lines of the categories described above – will benefit this. Moreover, such modeling may lead to more interpretable models. This needs researchers working on misinformation to couple deep learning models with kernel machines, or further the research into the modeling of contexts in kernel machines [71]. As such, advancing research into misinformation cannot be decoupled from the advances that need to be made into understanding the modality of communication, be it textual or visual or multimodal. Error analyses on such interpretable models of misinformation can then spur a positive feedback loop of progressively more interpretable models, which will in turn enable us to model the more subtle and nuanced forms of misinformation, as discussed by Schauer and Zeckhauser [16].

The ability to detect misinformation more accurately, by itself does not lead to its prevention. However, we anticipate that the ability to model various types of contexts will foster and spawn research into co-learning between various modalities and contexts, which will in turn allow us to preemptively identify harmful content at early stages of their propagation through social networks. The identification of local and global graphical contexts could, for instance, lead to subgraph matching algorithms being used for such preventive measures. Extra-linguistic context like user-metadata could achieve similar objectives. Indeed, some work on disinformation has achieved remarkable success with this approach (for example, the identification of bots, sockpuppet accounts, and trolls on Twitter who spread misinformation during elections or about vaccines).

F. Is there a cure for misinformation?

Prevention of misinformation is an ambitious goal, and a cure is even more so. It is almost common knowledge – an adage, if you will – that trust, once broken, cannot fully recover. And the perverse success of misinformation and its harmful effects is perhaps a testimony to the deep trust deficit that permeates our society today. However, we remain optimistic that some repair is still possible.

There is a single anecdote we would like to share, to illustrate both the trust deficit as well as possible (even if partial) resolution.

In recent years, Detroit went forth with a tree planting initiative. To their surprise, a large number (over 7,000) of of non-white residents refused to have tree planted, free of cost to them, in front of their property. Upon investigating, it was discovered that the residents retained the memory of the city cutting down large trees after the 1967 race rebellion so that (allegedly) their homes could be under easier surveillance from helicopters and other high vantage points. Another reason for refusal was that the people offering to plant trees were not locals of Detroit. However, the survey that discovered this reason behind the refusal also received positive responses about the survey itself, where the residents were satisfied that such a survey was, indeed, done. [72]

Similar anecdotes and qualitative surveys show a consistent pattern – the trust deficit exists because of historical events and collective memory of such events (*e.g.*, trust deficit about vaccines in low-income neighborhoods), but when local residents form the bridge between researchers and the population

actually affected by the harm, there is at least some mitigation of the distrust. Collaborative efforts must therefore be the norm in the future, and this collaboration must extend outside the world of academic research to include "boots on the ground". Combined with preemptive measures such as training through workshops [73], [74], this may be the cure we need.

V. Conclusion

In this article, we underscore the limitations of current computational research in the detection and prevention of misinformation. Furthermore, we discuss the reasons for these limitations, and how they may be overcome, by delving into various nuances of misinformation that go beyond a simplistic binary classification. In particular, we highlight the need to incorporate context in various forms, and stress on the need for collaborative efforts across disciplines, both in order to develop and model some important types of context (such as cross-genre fact-checking for medical findings) as well as to bridge the trust deficit in our society that makes people more susceptible to misinformation.

Our vision may be optimistic, perhaps even overly optimistic. But we hope that it has shed some light on various aspects overlooked in quantitative studies on misinformation, and that it aids further research in a multiplicity of vertical avenues to tackle the problem of widespread misinformation.

Acknowledgement

This work was supported in part by NSF under awards IIS 2027750 and SES 1834597.

References

[1] R. Darnton, "The True History of Fake News," The New York Review, February 13 2017. [Online]. Available: https://www.nybooks.com/daily/2017/02/13/the-true-history-of-fake-news

[2] J.-J. Rousseau, *Confessions*. London: Aldus Society, 1903.

[3] W. N. Griggs, *The Celebrated "Moon Story", Its Origins and Incidents*. Bunnell and Price, 1852.

[4] T. Jones, "Dewey defeats Truman: The most famous wrong call in electoral history," Chicago Tribune, December 9 2007. [Online]. Available: https://www.chicagotribune.com/featured/sns-dewey-defeats-truman-1942-20201031-5kkw5lpdavejpf4mx5k2pr7trm-story.html

[5] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting decetion from linguistic styles," *Personality and Social Psychology Bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[6] S. Feng, R. Banerjee, and Y. Choi, "Characterizing Stylistic Elements in Syntactic Structure," in *Proceedings of the 2012 Joint Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 1522–1533.

[7] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1589–1599.

[8] S. Kumar, R. West, and J. Leskovec, "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, p. 591–602.

[9] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal Variational Autoencoder for Fake News Detection," in *The World Wide Web Conference*. Association for Computing Machinery, 2019, p. 2915–2921.

[10] A. Taylor, "Before 'fake news', there was soviet 'disinformation'," The Washington Post, November 26 2016. [Online]. Available: https://www.washingtonpost.com/news/worldviews/wp/2016/11/26/before-fake-news-there-was-soviet-disinformation

[11] D. Fallis, "A functional analysis of disinformation," in *iConference 2014 Proceedings*, 2014.

[12] ——, "What is disinformation?" *Library Trends*, vol. 63, no. 3, pp. 401–426, 2015.

[13] L. Floridi, *The Philosophy of Information*. Oxford University Press, 2011.

[14] T. L. Carson, *Lying and deception: Theory and practice*. Oxford University Press, 2010.

[15] F. Giglietto, L. Iannelli, L. Rossi, and A. Valeriani, "'Fake news' is the invention of a liar: How false information circulates within the hybrid news system," *Current Sociology*, vol. 67, no. 4, pp. 625–642, 2019.

[16] F. Schauer and R. Zeckhauser, "Paltering," in *Deception: From Ancient Empires to Internet Dating*, B. Harrington, Ed. Stanford University Press, 2009, pp. 38–54.

[17] T. L. Carson, "The Range of Reasonable Views about the Morality of Lying," in *Lying: Language, Knowledge, Ethics, and Politics*, E. Michaelson and A. Stokke, Eds. Oxford University Press, 2018, ch. 7, pp. 145–160.

[18] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, "A context-aware approach for detecting worth-checking claims in political debates," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2017, pp. 267–276.

[19] N. Nakashole and T. M. Mitchell, "Language-aware truth assessment of fact candidates," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2014, pp. 1009–1019.

[20] F. Arslan, N. Hassan, C. Li, and M. Tremayne, "A Benchmark Dataset of Check-Worthy Factual Claims," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 821–829, 2020.

[21] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," *ACM SIGKDD Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016.

[22] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–36, 2018.

[23] E. Ostermeier, "Selection bias? PolitiFact rates Republican statements as false at three times the rate of Democrats," *Smart Politics*, vol. 10, 2011.

[24] J. E. Uscinski and R. W. Butler, "The epistemology of fact checking," *Critical Review*, vol. 25, no. 2, pp. 162–180, 2013.

[25] J. E. Uscinski, "The epistemology of fact checking (is still naïve): Rejoinder to amazeen," *Critical Review*, vol. 27, no. 2, pp. 243–252, 2015.

[26] A. Abbasi, F. M. Zahedi, and S. Kaza, "Detecting fake medical web sites using recursive trust labeling," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 4, pp. 1–36, 2012.

[27] J. Pasternack and D. Roth, "Latent credibility analysis," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1009–1020.

[28] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Credibility assessment of textual claims on the web," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 2173–2178.

[29] G. Weikum, "What computers should know, shouldn't know, and shouldn't believe," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1559–1560.

[30] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 2931–2937.

[31] A. Mantzarlis and A. Valeeva, "Is Russia Today a legitimate fact-checker? We did the math," The Poynter Institute for Media Studies, July 27 2017. [Online]. Available: https://www.poynter.org/fact-checking/2017/is-russia-today-a-legitimate-fact-checker-we-did-the-math/

[32] Reporters Without Borders, "Xinhua: the world's biggest propaganda agency," January 20 2016. [Online]. Available: https://rsf.org/en/reports/xinhua-worlds-biggest-propaganda-agency

[33] S. Hersh, "Whose sarin?" *London Review of Books*, vol. 35, no. 24, 2013.

[34] A. Ghosh and T. Veale, "Fracking Sarcasm using Neural Network," in *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, 2016, pp. 161–169.

[35] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in Twitter," in *Language Resources and Evaluation*, vol. 47. Springer Nature, 2013, pp. 239–268. [Online]. Available: https://link.springer.com/article/10.1007/s10579-012-9196-x

[36] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as Contrast between a Positive Sentiment and Negative Situation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2013, pp. 704–714. [Online]. Available: https://aclanthology.org/D13-1066

[37] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 2367–2376.

[38] D. Bertero and P. Fung, "Deep learning of audio and language features for humor prediction," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), 2016, pp. 496–501.

[39] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2506–2515.

[40] D. Das and A. J. Clark, "Sarcasm detection on flickr using a cnn," in *Proceedings of the 2018 International Conference on Computing and Big Data*, 2018, pp. 56–61.

[41] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1136–1145.

[42] E. Camp, "Sarcasm, Pretense, and The Semantics/Pragmatics Distinction," *Noûs*, vol. 46, no. 4, pp. 587–634, 2012.

[43] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[44] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[45] J. T. Hancock, L. E. Curry, S. Goorha, and M. T. Woodworth, "Lies in conversation: An examination of deception using automated linguistic analysis," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 26, no. 26, 2004.

[46] A. Vrij, R. Fisher, S. Mann, and S. Leal, "A cognitive load approach to lie detection," *Journal of Investigative Psychology and Offender Profiling*, vol. 5, no. 1-2, pp. 39–43, 2008.

[47] R. E. Petty and J. T. Cacioppo, "Involvement and Persuasion: Tradition Versus Integration," *Psychol. Bull.*, vol. 107, no. 3, pp. 367–374, 1990.

[48] C. L. Corritore, S. Wiedenbeck, B. Kracher, and R. P. Marble, "Online Trust and Health Information Websites," *Int J Tech and Hum. Interaction*, vol. 8, no. 4, pp. 92–115, 2012.

[49] A. Flanagin and M. J. Metzger, "From Encyclopaedia Britannica to Wikipedia: Generational Differences in the Perceived Credibility of Online Encyclopedia Information," *Inf., Commun. & Soc.*, vol. 14, no. 3, pp. 355–374, 2011.

[50] J. Olaisen, "Information Quality Factors and the Cognitive Authority of Electronic Information," in *Information Quality: Definitions and Dimensions*, I. Wormwell, Ed. Taylor Graham, 1990, pp. 91–121.

[51] S. A. Rains and C. D. Karmikel, "Health information-seeking and perceptions of website credibility: Examining Web-use orientation, message characteristics, and structural features of websites," *Comput. in Hum. Behav.*, vol. 25, no. 2, pp. 544–553, 2009.

[52] K. Starbird, "Disinformation's spread: bots, trolls and all of us," *Nature*, vol. 571, no. 7766, p. 449, 2019.

[53] J. Zhang, L. Cui, Y. Fu, and F. B. Gouza, "Fake news detection with deep diffusive network model," *arXiv preprint arXiv:1805.08751*, 2018.

[54] V. Manusov, ""It Depends on Your Perspective": Effects of Stance and Beliefs about Intent on Person Perception," *Western Journal of Communication*, vol. 57, no. 1, pp. 27–41, 1993.

[55] M. Gao, Z. Xiao, K. Karahalios, and W.-T. Fu, "To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, nov 2018.

[56] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[57] C. D. Broad, *Examination of McTaggart's Philosophy (Volume 2)*. Cambridge University Press, 1938.

[58] H. Dubossarsky, D. Weinshall, and E. Grossman, "Outta control: Laws of semantic change and inherent biases in word representation models," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 1136–1145.

[59] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016, pp. 1489–1501.

[60] A. Jatowt and K. Duh, "A framework for analyzing semantic change of words across time," in *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, 2014, pp. 229–238.

[61] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, 2015, pp. 757–762.

[62] S. Skalicky and S. Crossley, "Linguistic features of sarcasm and metaphor production quality," in *Proceedings of the Workshop on Figurative Language Processing*. Association for Computational Linguistics, 2018, pp. 7–16.

[63] C. Zuo, K. Mathur, D. Kela, N. S. Faramarzi, and R. Banerjee, "Beyond Belief: A Cross-Genre Study on Perception and Validation of Health Information Online," unpublished.

[64] J. W. Tankard Jr and M. Ryan, "News Source Perceptions of Accuracy of Science Coverage," *Journalism Quarterly*, vol. 51, no. 2, pp. 219–225, 1974.

[65] B. Moore and M. Singletary, "Scientific Sources' Perceptions of Network News Accuracy," *Journalism Quarterly*, vol. 62, no. 4, pp. 816–823, 1985.

[66] F. Molitor, "Accuracy in Science News Reporting by Newspapers: The Case of Aspirin for the Prevention of Heart Attacks," *Health Commun.*, vol. 5, no. 3, pp. 209–224, 1993.

[67] R. Moynihan, L. Bero, D. Ross-Degnan *et al.*, "Coverage by the News Media of the Benefits and Risks of Medications," *New Eng J Med*, vol. 342, no. 22, pp. 1645–1650, 2000.

[68] C. Zuo, N. Acharya, and R. Banerjee, "Querying across genres for medical claims in news," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 1783–1789.

[69] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, and T. Mandl, "The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news," in *Advances in Information Retrieval*, D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, Eds. Cham: Springer International Publishing, 2021, pp. 639–649.

[70] C. Zuo, Q. Zhang, and R. Banerjee, "An empirical assessment of the qualitative aspects of misinformation in health news," in *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. Online: Association for Computational Linguistics, 2021, pp. 76–81.

[71] M. Belkin, S. Ma, and S. Mandal, "To understand deep learning we need to understand kernel learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 541–549.

[72] B. Mock, "Why Detroit Residents Pushed Back Against Tree-Planting," Bloomberg CityLab, January 11 2019. [Online]. Available: https://www.bloomberg.com/news/articles/2019-01-11/why-detroiters-didn-t-trust-city-tree-planting-efforts

[73] S. Van Der Linden, E. Maibach, J. Cook, A. Leiserowitz, and S. Lewandowsky, "Inoculating against misinformation," *Science*, vol. 358, no. 6367, pp. 1141–1142, 2017.

[74] G. Vaidyanathan, "News feature: Finding a vaccine for misinformation," *Proceedings of the National Academy of Sciences*, vol. 117, no. 32, pp. 18 902–18 905, 2020. [Online]. Available: https://www.pnas.org/content/117/32/18902