On Unsupervised Methods for Fake News Detection

Deepak P

Abstract In this chapter, we consider a reasonably underexplored area in fake news analytics, that of unsupervised learning. We intend to keep the narrative accessible to a broader audience than machine learning specialists and accordingly start with outlining the structure of different learning paradigms vis-à-vis supervision. This is followed by an analysis of the challenges that are particularly pertinent for unsupervised fake news detection. Third, we provide an overview of unsupervised learning methods with a focus on their conceptual foundations. We analyze the conceptual bases with a critical eye and outline other kinds of conceptual building blocks that could be used in devising unsupervised fake news detection methods. Fourth, we survey the limited work in unsupervised fake news detection in detail with a methodological focus, outlining their relative strengths and weaknesses. Lastly, we discuss various possible directions in unsupervised fake news detection and consider the challenges and opportunities in the space.

Keywords Unsupervised learning · Fake news detection

1 Introduction

Fake news, the topic of this book, is a phenomenon of increasing concern over the last many years. Unlike the vast majority of machine learning tasks that seek to automate tasks that humans are quite adept at, such as image segmentation [7], action recognition [10], and emotion analysis [30], fake news identification [25] is a task of a different nature. Humans often find it hard to assess the veracity of news they come across due to a plurality of factors. First, in certain cases such as those of magic cures and anti-vaccination news, laypersons do not have enough knowledge of the domain to assess the veracity of a given news piece. Second, the news may pertain to real-time events that have not had time to gain enough of a footprint in public discourse, so there is no reference point to judge its veracity. Third, much fake news is carefully tailored to exploit human cognitive biases such as confirmation

Deepak P et al., *Data Science for Fake News*, The Information Retrieval Series 42, https://doi.org/10.1007/978-3-030-62696-9_2

bias, echo chamber effects, and negativity bias; some discussions appear in the literature (e.g., [6, 26]). There are various other challenges that undermine the lay-person's ability to fact check for herself without the aid of additional technology or knowledge, but for the purposes of this chapter, it is enough to emphasize that humans could legitimately find the task difficult. In a way, the machine learning (ML) models for fake news detection seek to surpass the accuracy levels achieved by humans within reasonable time, effort, and knowledge limits.

1.1 Paradigms of Machine Learning vis-à-vis Supervision

The two broad streams of machine learning, viz., supervised and unsupervised, differ in terms of whether they assume the availability of historical labelled data to enable learning a statistical model that would then be used to label new data. Supervised learning, broadly construed, can be thought of as a mechanism of taking a training dataset of input–output pairs $\mathcal{T} = \{\ldots, [I, O], \ldots\}$ and producing a statistical model that embodies a mapping from the domain of inputs to outputs, $\mathcal{F} : D(I) \rightarrow D(O)$. For the task of fake news detection, the target domain is a veracity label, which could be one of $\{Fake, Legitimate, Doubtful\}$ or a number in a [0, 1] range with the ends indicating *fake* and *legitimate*, respectively. The shape and form of the statistical model is *guided* by the labels in the training data but is constrained in ways to ensure its generalizability and/or conformance to knowledge about how the domain functions. On the other hand, the raw material for unsupervised methods is simply a set of unlabelled data objects, $\mathcal{T} = \{\ldots, I, \ldots\}$, from which the statistical models should learn to differentiate fake news from legitimate news. In contrast to supervised learning, the unsupervised methods may not necessarily produce a mapping from an input data object to a veracity label but could instead provide a grouping or representation whose subspaces are homogeneous with respect to veracity. For example, a clustering that is able to group a set of articles into two unlabelled groups, one of which is all fake articles and the other one all legitimate ones, could be considered successful from the perspective of fake news detection despite not being able to indicate which cluster is fake and legitimate. That said, producing a label along with output clusters only enhances the usefulness of the clustering with respect to the task.

There are other paradigms of machine learning that can make use of different flavors of supervision rather than the all-or-nothing cases discussed above. These include semi-supervised learning [39], active learning [24], and reinforcement learning [12]. Our focus in this chapter will be on unsupervised approaches to the task.

1.2 Challenges for Unsupervised Learning in Fake News Detection

When considering any analytics task, it may be observed that addressing the task in the unsupervised setting is obviously much more challenging than addressing it in a supervised setting. The former does not have the luxury of *label guidance* to complement or supplement domain knowledge-based directions in searching for effective statistical modeling. Thus, unsurprisingly, the effectiveness of unsupervised learning often falls well-short of that of supervised models.

We now consider some challenges for unsupervised learning for fake news detection. To offset the unavailability of labelled data, a natural pathway would be to develop a deeper understanding of the nature of fake news. This could be along dimensions such as *author*, *metadata* (e.g., article category, time, location), *propagation*, and *content*. For example, we may want to identify authors who regularly post content of limited veracity and categories (e.g., magic cures) that regularly get populated with disinformation. Similarly, if the news propagation is deeply dichotomous on the emotional aspect (e.g., either extremes of love or hate, without much in the middle ground), it may suggest correlation with disinformation or other aspects such as highly opinionated or divisive content. Some patterns in the content could itself be highly revealing; examples include clickbait-ish contents where the title and the article are highly divergent, or a sensationalist image placed strategically. Broadly speaking, the unavailability of label guidance could be offset by identifying some high-level patterns that correlate with veracity, which could then be folded into an unsupervised method. It may, however, be noted that such high-level patterns are unlikely to generalize across domains. For example, a fake news that deals with celebrity gossip may have a different structure than disinformation that deals with a COVID-19 cure (the fake news around COVID-19 has been called an infodemic [38]). Thus, the unsupervised methods that embed deeper domain knowledge could implicitly be very specific to the domain given that the deeper domain knowledge would itself be domain-specific. This may be contrasted with supervised learning where the label-guided learning framework may be generalizable across domains; concretely, it may learn *different models* for different domains using the *same* learning strategy since the labels in different domains could *pull* the learner in different directions that are suited for those domains.

The discussion suggests that efforts toward crafting unsupervised learning algorithms for fake news detection would entail the following:

– **Deeper Efforts at Understanding the Domain:** It would be useful, if not necessary, to understand the dynamics of the target domain through extensive studies. These may involve other scholarly realms beyond computer science; for example, the usage of confirmation bias as a tool may be more prevalent among xenophobic, anti-minority, and far-right rhetoric in political fake news (e.g., [16]) and thus could naturally be an effective factor in fake news identification too. On

the other hand, an authoritative or assertive linguistic style with an abundance of anecdotes may characterize medical fake news. Such explorations may be situated within other disciplines such as psychology and linguistics or at their intersection with computing. This would likely make the body of literature around unsupervised fake news detection (UFND) more interdisciplinary than its supervised counterpart.

- **Empirical Generalizability:** While making use of insights from across disciplines as well as through extensive data analysis, there should also be an unrelenting focus toward empirical generalizability. If we focus on a single dataset and try out various combinations from a vocabulary of insight-driven heuristics, it is possible to be able to arrive at a spurious technique that performs very well for that dataset. This is often due to the well-understood mechanism of spurious pattern discovery called *data dredging* [29]. The vocabulary of fake news patterns that come from a deep understanding of specific domains may not be amenable to manual audit due to vocabulary size, complexity, and the deep expertise required for such analyses. Thus, there should be a particular focus on empirical generalizability to ensure that the developed methods are practically usable as well as legitimate. This may be achieved through verification over a large number of datasets from the target domain or by vetting for the validity of patterns with scholarly expertise in the target domain. This is particularly crucial when there is reliance on patterns identified through extensive empirical experimentation.
- **Ethical Considerations:** Machine learning methods more often rely on empirical than analytical analyses to make their point. Crudely put, it considers that the past is predictive of the future and develops techniques that project historical patterns for usage in unseen data from the future. This makes it systematically less capable of identifying novel and emerging patterns, something which has been very well understood in machine learning, with phenomena such as *concept drift* [32] and methods such as *transfer learning* [18] being well explored. When machine learning is used for tasks such as fake news detection, there is a chance that its widespread adoption would itself skew the data. For example, a novel pattern of legitimate news may be mistaken for fake news and may never be shown to users, leading to it never being labelled by humans anymore. Thus, the next generation of algorithms that work on the data would not be able to correct for it, given the lack of feedback. Such data bias and how they are exacerbated through algorithms have been well studied in the law enforcement domain [21]. Furthermore, the patterns embodied in the method could possibly be differentially equipped to identify fake news in subspaces; for example, a model incompetent at detecting fake remedies for tuberculosis, a predominant disease in some parts of Africa, may still fare well on the overall accuracy when tested over a dataset procured from the Western world where tuberculosis is rare. Unlike the case of supervised learning, there is an increased likelihood of biased high-level heuristics, over and above biased data, to be embedded in unsupervised learning algorithms.

– **Continuous Refinement:** Supervised learning systems can be retrained with new and updated datasets to some extent despite issues such as algorithms affecting the dataset, as described above; however, the analogous refinement of unsupervised learning algorithms requires updates to the algorithm design itself. Such refinements with changing data and societal discourse would require, as in the case of the algorithm design process, identifying and updating high-level heuristics with continuous vetting with domain expertise. We will return to this issue later on in this chapter.

2 Unsupervised Fake News Detection: A Conceptual Analysis

We now consider a conceptual positioning of the various research efforts on unsupervised fake news detection (UFND). As outlined in the discussion above, each unsupervised fake news detection method is invariably driven by high-level assumptions about patterns in the data that correlate with the veracity of news. In this section, we target to position the methods at a conceptual level, without getting into technical details; the technical and methodological details would form the topic of a subsequent section.

2.1 Conceptual Basis for UFND Methods

Given the paucity of UFND methods in the literature, we are able to consider the conceptual basis of each work separately. We have come across four research papers proposing UFND methods, which we use in our discussion as state-of-the-art methods. We have italicized the high-level heuristics employed, as and when discussed, for convenient reference. These are as follows:

– **Truth Discovery:** Truth discovery is the task that deals with estimating the veracity of an information nugget when it is reported by multiple sources (e.g., multiple websites), with conflicts existing across the multiple reports; a survey appears here [14]. An early work, perhaps the first UFND technique [37], makes use of truth discovery heuristics in estimating the veracity of information. It makes use of the high-level heuristic that *a piece of news is likely to be true when it is provided by many trustworthy websites*. Trustworthiness is not assumed as given a priori but estimated in an iterative fashion along with veracity estimation of various news pieces.
– **Differentiating User Types:** Many social media websites such as Twitter provide a way for users to be labelled as *verified*. This label is regarded as broadly honorific and could be interpreted as indicating a higher status or trustworthiness. UFND [36] exploits this user verification process in fake news detection. In particular, it models news veracity as being determined by

user opinion, modeling the way user opinion is factored into veracity analyses differently for verified and unverified users. Their heuristic, as quoted verbatim from Sect. 3.1, is the following: "an implicit assumption is imposed that verified users, who may have large influences and high social status, may have higher credibility in differentiating between fake news and real news." They make use of a generative framework to employ the above assumption into a veracity detection framework.

- **Propagandist Patterns:** The first unsupervised method to make use of behavioral analyses of user groups is a work that targets identifying propagandist misinformation in social media [17]. Their task is motivated by the increasing prevalence of orchestrated political propaganda and misinformation in social media, possibly facilitated by authoritarian governments and usually driven by large groups of users who work collectively to enhance acceptability of the official version. The proposed method for detecting propagandist misinformation relies on identifying *groups of users who write political posts that are textually and temporally synchronized, and aligned with the "official" vision or "party line."* Their method makes use of repeated invocations of clustering and frequent itemset mining [11], both of which are popular unsupervised learning methods.
- **Inter-user Dynamics:** GTUT [8], Graph Mining over Textual, User and Temporal Data, a recently proposed graph-based method for fake news detection, makes use of a phased approach that relies on heuristics that exploit assumptions on user dynamics, in what may be seen as a generalization of the user dynamics approach in [17] to cover a broader spectrum of fake news. In the first phase, they assume that *a set of articles posted by the same users at similar times through textually similar posts are fake*. This assumption follows, as they point out, from orchestrated behavior that is often observed in sharing fake news. Once such a core set of fake news articles are identified, the labels are propagated to other articles based on both *user correlation* and *textual similarity*. Thus, the heuristic beyond the first phase can be summarized as *articles that are similar to core fake articles based on posting users and textual similarity are likely to be fake*. The above heuristics are also analogous (i.e., as vice versa) to identifying a core set of trustworthy/legitimate articles and propagating trustworthiness labels.

Any single pattern or a single cocktail of patterns embedded in an algorithm being used in a widespread manner to counter fake news has high potential risks. This is best understood when fake news detection is seen from the perspective of gamification. When a single technique becomes widespread, the heuristics used by it would become well understood, and the authors of fake news would consequently *game* it by identifying ways to circumvent being caught by this. User dynamics heuristics could be circumvented by automated or semiautomated staggered posting of messages, while majority-oriented heuristics can be circumvented by organizing an orchestrated posting of messages aided by blackmarket services [5]. This makes any single static solution infeasible for effective fake news debunking in the long run. The existence of multiple methodologies for fake news detection that are

continuously refined to be in tune with the current realities of the social media ecosystem is likely the best way to tackle the disinformation menace.

2.2 Critical Analysis of UFND Conceptual Bases

In the following discussion, we consider the relative merits and demerits of the conceptual basis of the techniques discussed above. This is not to undermine their value in being part of a mix of effective methodologies for UFND, but just to ensure a more nuanced understanding. We consider each of the techniques discussed above, in turn.

Truth Discovery

The truth discovery approach has a distinctly majoritarian flavor, whereby a more widespread opinion is likely to be regarded as truer than a narrowly shared one. While the authors in [37] explicitly clarify their assumption that they expect *a higher divergence of false facts* (Heuristic 3 in Sect. 2.2), the validity of their assumption may be challenged if multiple sources may be persuaded, with the aid of a mushrooming market around blackmarket services (e.g., [5]), to post the same fake content. This is plausible especially in narrow-domain topics such as fake news intended to malign a particular local enterprise. Such a situation could persuade the algorithm to consider the fake version as true and vice versa. However, this possibility is somewhat limited by the fact that trustworthy services are less likely to engage in such blackmarket orchestration, which places their trustworthiness at stake in the long run.

Differentiating User Types

The user type differentiation and the assumption of *enhanced credibility of verified users* employed by Yang et al. [36] are an interesting heuristic to analyze. Account verification in social media, according to Wikipedia,[1] was initially a feature for public figures and accounts of public interest, individuals in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas. It was introduced by Twitter in June 2009, followed by Google Plus in 2011, Facebook in 2012, Instagram in 2014, and Pinterest in 2015. On YouTube, users are able to submit a request for a verification badge once they obtain 100,000 or more subscribers. In July 2016, Twitter announced that, beyond public figures, any individual would be able to apply for account verification. With the

[1] https://en.wikipedia.org/wiki/Account_verification—accessed 28 June, 2020.

focus of [36] on Twitter, we will consider Twitter verified users more carefully. Twitter's *request verification* service was temporarily suspended in February 2018, following a backlash over the verification of one of the organizers of the far-right *Unite the Right* rally due to a perception that verification conveys "credibility" or "importance." As of June 2020, Twitter is reportedly still working on bringing back the *request verification* feature.[2] Given this background, the usage of verified accounts as those with enhanced credibility raises some concerns. First, the authors in [36] say: "... in preparing our data, we only consider the tweets created by verified users and the related social engagements (like, retweet, and reply) of the unverified users." This data preparation principle severely limits the ability of their method to detect fake news within narrow domains that may involve very few or no verified users. While the techniques proposed are generalizable, in principle, to any kind of classification of higher-status users, it is yet to be empirically verified for the general case. Second, given that verified users were intended to involve public figures in areas such as politics, religion, music, acting, fashion, journalism, media, etc., the definition could exclude domain experts who may be best positioned to provide credible and well-studied opinions. For example, academics who may be able to provide credible analyses of science fake news, or doctors who may be able to identify health fake news, are kept out of the ambit of verified users. This also likely renders the method to be of limited utility even for many broad domains. Third, while we have not found any analyses of verified user distribution across geographies, it may be reasonably assumed that it is skewed in favor of areas of deep social media penetration such as the developed world. This geographic skew would reflect in the method and could dent its applicability for pressing issues in the global south, such as Africa and South Asia.

Propagandist Patterns

The paper that considers identifying propagandist patterns [17] is quite friendly for analysis in that it explicitly lays down the assumptions. We re-produce them below: *We assume that propaganda is disseminated by professionals who are centrally managed and who have the following characteristics:*

1. *They work in groups.*
2. *Disseminators from the same group write very similar (or even identical) posts within a short timeframe.*
3. *Each disseminator writes very frequently (within short intervals between posts and/or replies).*
4. *One disseminator may have multiple accounts; as such, a group of accounts with strikingly similar content may represent the same person.*
5. *We assume that propaganda posts are primarily political.*

[2] https://www.theverge.com/2020/6/8/21284406/twitter-verified-back-badges-blue-check.

Fig. 1 A propaganda-based misinformation from the Indian context

6. *The content of tweets from one particular disseminator may vary according to the subject of an "assignment," and as such, each subject is discussed in disseminator's accounts during some temporal frame of its relevance.*
7. *Propaganda carries content similar to an official governance "vision" depicted in mass media.*

The above observations, partly motivated in the paper through examples from the Russian social network *VK*,[3] are likely to hold true for most regimes with shallow democracies and autocratic tendencies. Figure 1 shows a political fake news from the Indian context, which illustrates agreement to most of the assumptions above. The easiest way to game the system that works using the above assumptions would be to make the posts textually dissimilar; however, this would require much work and could undermine the ability of such fake news armies to mass produce fake tweets with high throughput. This makes the assumption fairly robust, at least in the short run. The limitations of the approach are largely engrained in the assumptions themselves, in that these apply only to fake news in the political domain produced in favor of the authoritarian regimes. In particular, in a federal governance system

[3] https://vk.com/.

such as those in the USA, India, or Spain, with different political parties leading different provincial governments, there may not be a *coherent official governance vision*, undermining assumption #7 above to some extent. It is likely that a subset of such assumptions above also apply to some other domains, such as religion-based fake news, but more studies may be needed to evaluate those aspects.

Inter-user Dynamics

The recent work on using inter-group dynamics in UFND, called GTUT [8], makes use of three phases, with the core assumption embedded in the first phase of identifying a core set of fake news and legitimate news. Their key assumption is that a core set of fake news articles can be identified as *a set of news articles shared by across a set of users using tweets that are temporally and textually similar*. This resembles some parts of the behavioral identification assumptions used in [17]; however, by relaxing the assumptions of *official vision adherence* and certain others, this is likely applicable to a broader set of scenarios. Analogous to the above, they use a curiously analogous assumption for identifying a core set of legitimate articles. In essence, *a set of news articles shared across a set of users using tweets that are temporally and textually dissimilar* are identified as a core set of legitimate news articles. While a reasoning for this is not adequately described, it is unclear as to the nature of legitimate news articles that would be shared in a temporally and textually dissimilar fashion. Clearly, this heuristic would have limited applicability in the political realm where legitimate news and fake news are often shared synchronously, when the event is in public memory. However, it is notable that these heuristics are only used in order to identify a core set of fake and legitimate articles (around 5% of the dataset, as mentioned in Sect. 4.1). In the subsequent phases, the fake and legitimate news labels are propagated using similarity between articles estimated as a mix of commonality between users and textual content of tweets. Another aspect of the method that may limit the applicability is the reliance on textual similarity. The method assumes that there is accompanying text along with an article over which textual similarity is assessed in the core set finding phase. It is not uncommon to simply share articles without posting any comment in social media; the applicability of GTUT over such posts would be evidently limited. On the positive side, much like observed in the case of [17], inter-user behavioral heuristics are harder to circumvent, making that a strong point of this method.

2.3 Building Blocks for UFND

While end-to-end techniques for UFND have evidently been limited in the literature, empirical analyses that could provide some building blocks for UFND have been explored lately. These are generally one of two types: (1) *computational social science studies* that seek to computationally verify a hypothesis rather than building

a technology for a particular task or (2) work on supervised learning methods that establish the utility of certain features that implicitly indicate fertile directions for UFND research. Work of the latter kind typically is limited in making an observation that a particular feature is useful without indicating the nature of difference between fake news and legitimate news along that feature. For example, if *punctuation* is found to be a useful feature, it does not tell us whether fake news is better or worse in punctuation vis-à-vis real news (though one may be able to guess easily, in this case, as to which is more likely). We consider a few such works below, without claiming to provide a comprehensive overview:

- **Satirical Cues:** Rubin et al. [22] study the usage of satirical cues in supervised fake news detection and provide evidence that *absurdity*, *grammar*, and *punctuation* are useful features.
- **Propagation Patterns:** Vosoughi et al. [34] present evidence that "Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information."
- **Topical Novelty:** Vosoughi et al. [34], in the same study as above, illustrate the utility of topical novelty against recent history as a useful way of identifying fake news, with fake news expected to be more novel topically.
- **Political Orientation and Age:** In a study based on Facebook, Guess et al. [9] say: "Conservatives were more likely to share articles from fake news domains, which in 2016 were largely pro-Trump in orientation, than liberals or moderates. We also find a strong age effect that persists after controlling for partisanship and ideology: On average, users over 65 shared nearly seven times as many articles from fake news domains as the youngest age group."
- **Effect of Fake News Based on Behavioral Traits of the Reader:** In a recent work, Pennycook and Rand [19] identify personality traits with respect to fake news vulnerability and say: "individuals who overclaim their level of knowledge also judge fake news to be more accurate." While this does not necessarily form a building block for UFND, it potentially indicates who may benefit more from the methods.
- **Psychological Appeal:** Acerbi [1] analyzes the cognitive appeal of online misinformation and suggests that misinformation may be correlated with *psychological appeal* in that it aims to exploit various cognitive inclinations of humans.
- **Language Style:** Rashkin et al. [20] illustrate that language style modeled through lexical features can help differentiate fake news from legitimate ones in a supervised task. Linguistic cues were also explored in [4].
- **Network Patterns:** An analysis [27] of dissemination patterns of news through the network indicates that the type of network formed through propagation can be revealing of the veracity of news.
- **Emotions:** Anoop et al. [2] report a computational social science study providing empirical evidence that the emotion profile of fake news differs from legitimate news, through an innovative mechanism that illustrates that emotion-amplified

fake news is farther away from their legitimate counterparts. Emotions and sentiments were also found to be useful in detecting fake reviews in another study [15].

- **Users Who Like:** In a large-scale study of Facebook likes, Tacchini et al. [31] suggest that *users who like* a post is a reasonable predictor of post veracity. This likely points to the existence of some consistent patterns of *liking* activity across the veracity dimension, which may be of use in UFND.
- **Lexical Coherence:** A recent computational social science study [28] considers the various ways of quantifying lexical coherence, and observes that word embedding based on coherence analyses is best suited to tease out the differences between fake and legitimate news.

The above is by no means an exhaustive list but serves to indicate the diversity of directions to explore toward building effective UFND methods. While several minor building blocks, even when packaged into a UFND method, may not have the muscle to compete with the state of the art in UFND, such efforts nevertheless contribute to building a diversity of UFND methods, diversity being an important factor as pointed out earlier. We may also add here that such research efforts are likely more suited to avenues focused on computational social science, such as the many avenues that have been instituted recently, viz., *Journal of Computational Social Science*,[4] *ACM Transactions on Social Computing*,[5] and *IEEE Transactions on Computational Social Systems*.[6]

3 Unsupervised Fake News Detection: A Methodological Analysis

Having introduced the various methods for UFND at the conceptual level in the previous section, we now endeavor to provide a tutorial overview of their methodological details. As in the previous case, we cover each method in turn.

3.1 Truth Discovery

The approach proposed in [37] makes use of an iterative approach toward veracity identification. The approach attacks two estimation problems concurrently:

- *Trustworthiness Estimation of Websites:* Estimating a non-negative trustworthiness score for each website as $T'(w)$.

[4] https://www.springer.com/journal/42001.

[5] https://dl.acm.org/journal/tsc.

[6] https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570650.

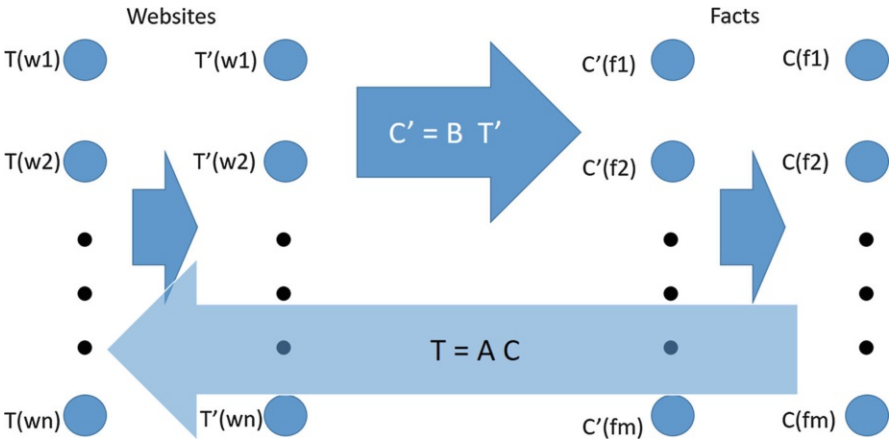

Fig. 2 Truth discovery approach from [37]. Figure adapted from across illustration in the paper

– *Confidence Estimation for Facts:* Estimating a confidence score for each fact as $C'(f)$.

The scores are directly related to trustworthiness and confidence, respectively; that is, higher scores indicate higher trustworthiness and higher confidence. There is also an additional construct, the *objects* associated with each fact, that is also used in the estimation. While the estimation process bears resemblance to the hub-authority score estimation in Hyperlink-induced Topic Search [13], the actual estimation process, as the authors say, is quite different in mathematical character. We provide an overview of the methodology employed in [37], to aid understanding of the spirit of the approach. The exact details are in the paper.

Figure 2 depicts an overview of the method. The set of websites are $W = \{w_1, w_2, \ldots, w_n\}$, across which a number of facts are mentioned, $F = \{f_1, f_2, \ldots, f_m\}$. For each website, there are two trustworthiness scores, $T(w)$ and $T(w')$; these are easily convertible across each other and serve to simplify the iterative computation process only. Analogously, there are two confidence scores, $C(f)$ and $C'(f)$, for facts that are also similarly inter-convertible.

The method starts with initializing all websites to be of equal trustworthiness, say 0.9, for $T(w)$. This is used to estimate $T'(w)$, which is then followed by two key matrix multiplication operations that form the key steps within each iteration:

– *Confidence from Trustworthiness:* Consider the $\{\ldots, T'(w), \ldots\}$ as an $n \times 1$ vector. This vector is transformed using an $m \times n$ matrix $\boldsymbol{B}$ that is structured as follows:

$$\boldsymbol{B}_{ij} = \begin{cases} 1 & if\ f_i\ is\ provided\ by\ w_j \\ \rho \times imp(f_k \rightarrow f_i) & if\ w_j\ provides\ f_k\ and\ o(f_k) = o(f_i) \\ 0 & otherwise \end{cases}$$

As obvious, $\boldsymbol{B}_{ij}$ quantifies the support from website w_j toward the fact f_i. The second case above takes care of the scenario where w_j does not directly provide the fact f_i but provides a related fact f_k that relates to the same object as f_i ($o(f)$ denotes the object the fact f relates to). In that case, the strength of the implication from f_k to f_i (which could be negative when f_k conflicts with f_i), denoted by $imp(f_k \rightarrow f_i)$, is scaled by a factor ρ. The transformation operation is

$$\overrightarrow{C'} = \boldsymbol{B}\ \overrightarrow{T'}$$

– *Trustworthiness from Confidence:* The estimation of trustworthiness from confidence is quite straightforward. In particular, the trustworthiness of a website is simply the average confidence of facts provided by it. In terms of matrix operations, this is modeled as a matrix $\boldsymbol{A}$ that is $n \times m$ whose entries are as follows:

$$\boldsymbol{A}_{ij} = \begin{cases} \frac{1}{|F(w_i)|} & if\ f_j \in F(w_i) \\ 0 & otherwise \end{cases}$$

where $F(w)$ is the set of facts provided by the website w. The transformation is then

$$\overrightarrow{T} = \boldsymbol{A}\ \overrightarrow{C}$$

The iterative process is stopped when the trustworthiness scores do not change much, and the confidence scores are returned as an estimation of veracity for each fact.

The empirical analysis of this method has been predominantly performed over datasets involving books and movies, and it is not clear about the applicability of this method for social media fake news debunking. One way to use this, however, would be to treat each profile as the equivalent of a website, and the *facts* contained with each post as similar to the facts provided by websites. A particular notable aspect of this method is that it provides a trustworthiness estimate along with confidence scores; thus, this could be used in order to assess the trustworthiness of social media profiles, when considering profiles as the equivalent of websites, as outlined above.

3.2 Differentiating User Types

We now consider the approach proposed in [36] and describe the methodological framework. The cornerstone of this work, as outlined earlier, is the differentiation between the *verified* and *unverified* users. They limit their remit to assessing the veracity of news stories that have been tweeted by at least one verified user. Each

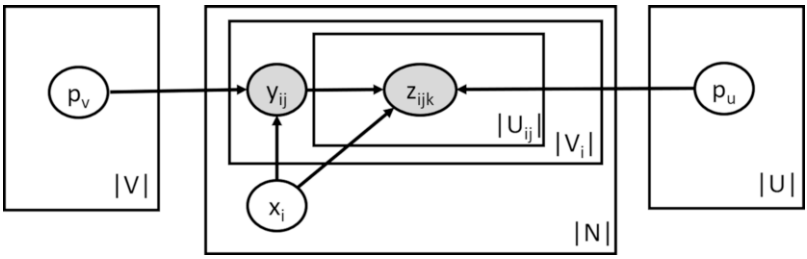

Fig. 3 Simplified graphical model from [36]

tweet of a news story by a verified user could be commented on or reacted to by one or more unverified users.

We introduce some notation to make the ensuing narration easier. Let $N = \{\ldots, n_i, \ldots\}$ be a set of news stories. Each news story n_i has an associated truth value $x_i \in \{0, 1\}$, estimating which forms the core target of the learning process in UFND. Let the opinion made by a verified user v_j on n_i be $y_{ij} \in \{0, 1\}$. The technique considers this opinion as an observed variable, since y_{ij} can be identified using sentiment or opinion analysis techniques. When a verified user v_j expresses an opinion on n_i, it is by means of a *tweet* or a *post* onto which *unverified* users can then engage and express their own opinion. Let z_{ijk} be the opinion expressed by the unverified user u_k on the v_j's post with n_i. This $z_{ijk} \in \{0, 1\}$ is also an observed variable estimated using sentiment or opinion analysis methods. The task is now to estimate x_is given the various y_{ij}s and z_{ijk}s. The authors use a probabilistic graphical model for this purpose.

Figure 3 depicts a simplified version of the graphical model omitting the details as well as hyperparameters for narrative simplicity. Each verified user is represented by a set of parameters p_v, and each unverified user by a different set p_u. The observed opinion y_{ij} is modeled as being influenced by both the truth value of the news x_i and the personal parameters of the user v_j. Similarly, the opinion z_{ijk} is influenced by all of (1) the truth value of x_i, (2) the opinion of the verified user y_{ij}, and (3) the parameters of the unverified user u_k. The parameters for verified and unverified users are modeled differently. The verified users are modeled using their *true positive rate* and *false positive rate*. Given that unverified users can only interact with a news within the context of a verified user's post, the unverified user has four parameters: the positivity rate for each combination of truth value of the article and opinion polarity of the verified user. For example, $p_u(z_{ijk} = 1|x_i = 0, y_{ij} = 0)$ indicates the likelihood of the unverified user expressing a positive opinion on a fake article (fake article since $x_i = 0$) to which the verified user has expressed a negative opinion (since $y_{ij} = 0$). The authors use a Gibbs sampling approach to estimate the latent parameters in the model, details of which are in the paper.

We had indicated in an earlier section that the authors of [36] had opined that "an implicit assumption is imposed that verified users, who may have large influences and high social status, may have higher credibility in differentiating between fake

news and real news." However, nothing in the methodology, as far as we understand, prevents verified users from having lower true positive rates (or higher false positive rates) than unverified users. There is evidently differentiated modeling of verified and unverified users, which may be implicitly pushing toward configurations that confer higher credibility to verified users, though it is far to reason analytically as to how such configurations are favored.

They evaluate the method against the truthfinder method as well as other baselines over two public datasets, LIAR [35] and BuzzFeed News data, and report accuracies of around 70% or higher.

3.3 Propagandist Patterns

The third work we describe, from [17], looks at using propagandist patterns in order to tackle misinformation that is aligned with the *official version*, probably inspired by scenarios in shallow democracies around the world. The technique itself is structured as a human-in-the-loop method that targets to identify patterns that need to be vetted by humans in order to complete the misinformation detection pipeline.

The automated part of the process follows the illustration in Fig. 4. We trace the process in reference to the seven assumptions outlined in Sect. 2.2. The target domain is Twitter, with the tweets ordered in temporal order indicated on the left-hand side. Tweets are split into temporal buckets to align with assumption #2. The tweets inside each time window are then clustered to ensure the textual similarity

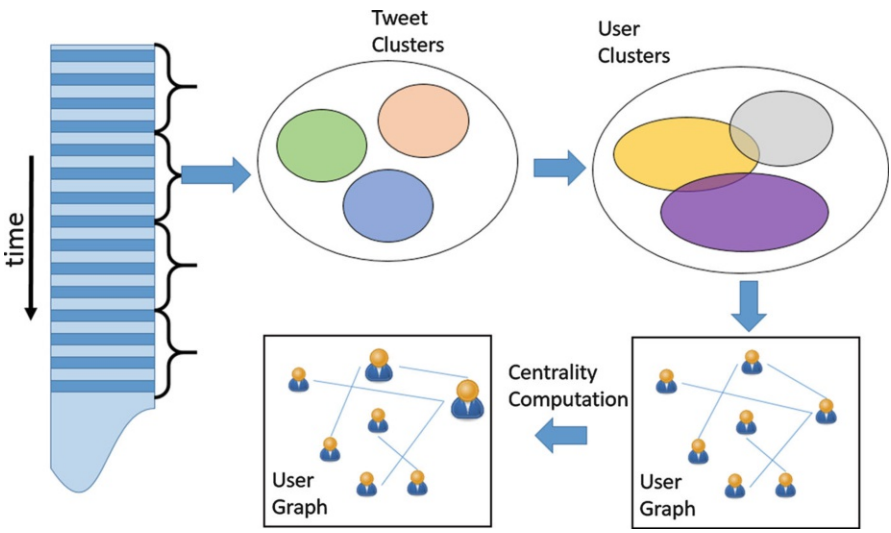

Fig. 4 Propagandist misinformation detection pipeline [17]

part of assumption #2. The tweets within the clusters are replaced with the userids of the authors, thus converting a disjoint clustering of tweets to an overlapping clustering of users; this is so since a user may have authored tweets that fall into disjoint clusters under the tweet clustering. This is inspired by both assumptions #1 and #2. The user clusters are converted into a user graph with cluster colocation being the criterion for edge induction. This user graph is then subjected to centrality detection to identify key users. In parallel, not shown in the diagram, there are two additional steps:

- A topic analysis over tweets to identify political topics in accordance with assumptions #5 and #7.
- Identification of user groups by application of a priori algorithm over user clusters. This addresses mostly assumption #4 and aligns with certain others.

The other assumptions, among the seven listed, are used by the human process. The authors do not perform a large empirical evaluation in the absence of labelled information but indicate the validity of the results from the method through manual vetting.

It may be seen that the manual steps in the process severely limit the applicability of the method in a large-scale manner. Additionally, given the lack of empirical validation over a labelled dataset, the recall (i.e., quantifying what has been missed) is not clear either. However, this presents a first effort in using inter-user behavioral dynamics within misinformation detection pipeline.

3.4 Inter-user Dynamics

We now come to the most recent work [8], one that uses inter-user behavioral dynamics in fake news detection using graph-based methods. GTUT, the method, relies on identifying temporally and textually synchronous behavior among users, as the key bootstrapping heuristic for identifying misinformation. This is enabled through a graph-based approach outlined below.

The graph employed by GTUT is a biclique, containing two kinds of nodes, *users* and *articles*. There exists an edge between a user and an article if the user has tweeted mentioning the article. In fact, a specific user may have tweeted about an article multiple times, leading to multiple tweets. Thus, an edge may *contain* multiple tweets. The first phase in the three-phase GTUT starts by identifying bicliques, a combination of a set of users and a set of articles such that each user–article pair in the combination is connected. One such biclique is illustrated in Fig. 5. Once such bicliques are identified, they are scored based on their *temporal* and *textual* coherence.

$$TTScore(B) = \lambda \times Temporal(B) + (1 - \lambda) \times Textual(B) \tag{1}$$

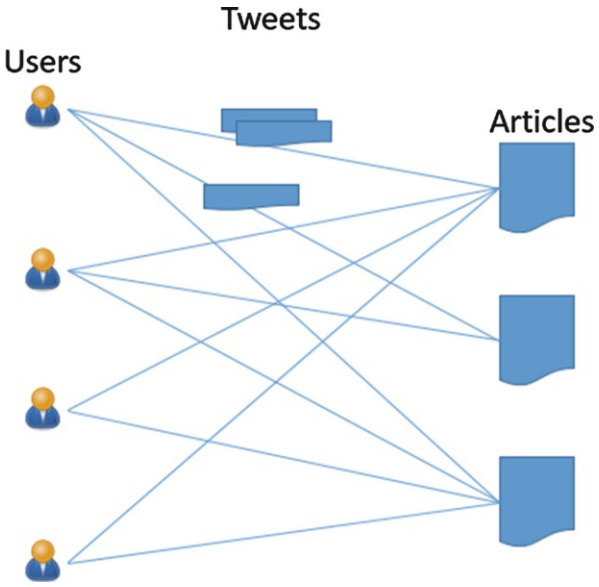

Fig. 5 Illustration of a biclique from [8]

These biclique level scores are transferred to each article A as follows:

$$TTScore(A) = \frac{\sum_{B \in BiCliques(A)} TTScore(B)}{|BiCliques(A)|} \tag{2}$$

where $BiCliques(A)$ indicates the bicliques that article A is part of. In other words, the score of an article is simply the average of the scores of bicliques that contain it. The 5% of articles with the highest coherence scores (indicating highly synchronous posting activity) are labelled as a core set of *fake* articles, with the analogous set at the other end being labelled as a core set of *legitimate* articles. This completes the first phase in GTUT.

The second phase propagates the fake and legitimate labels from the core set to all articles contained across the bicliques. The label propagation uses a graph structure with nodes being articles and edges being weighted as a weighted sum of *biclique similarity*, *user similarity*, and *textual similarity*.

$$\begin{aligned} E(A, A') = \alpha \times Jacc(BiCliques(A), BiCliques(A')) \\ +\beta \times Jacc(Users(A), Users(A')) + (1 - \alpha - \beta) \times Sim(A, A') \end{aligned} \tag{3}$$

where $Jacc(.,.)$ indicates the Jaccard similarity and $Users(A)$ are the set of users who shared the article A, and $Sim(.,.)$ is a textual similarity measure. At the end of this phase, each article contained in a biclique is labelled as either *fake* or *legitimate*.

The third phase propagates the labelling from within the bicliques to articles outside the bicliques; this uses the same structure as in the second phase, employing

label spreading. However, being outside the bicliques, there are only two factors in determining edge weights, which are user similarity and textual similarity. This completes the labelling process for all articles.

The methodology outlined above starts with identifying a core set of fake and legitimate articles and spreads the labels progressively outward to eventually cover all articles. This serial order of labelling imposes a high dependency on the initial core set finding; inaccurate finding of core sets of fake and legitimate articles could potentially lead the next two phases wayward. While the authors illustrate good empirical accuracies over two large-scale datasets, more studies could be used to assert the generalizability of the initialization heuristic.

4 The Road Ahead for Unsupervised Fake News Detection

We now outline some pathways in which research on unsupervised fake news detection could progress, in order to advance the state of the art. This is purely based on opinions that are in turn based on observations in the field, and an understanding of the fake news domain developed through engaging in research in the field and need to be taken with abundant caution.

4.1 Specialist Domains and Authoritative Sources

Of particular concern in 2020, as this chapter is being written, is that of COVID-19[7] fake news. These have been peddled by authoritative sources such as heads of state.[8] The debunking of such news, in the offline world, often happens through specialists considering the claim in the light of scholarly evidence and assessing whether the claim is tenable. A natural approach to automate fake news detection in such specialist domains is to similarly make use of authoritative knowledge sources. This would require a significant effort in developing bespoke techniques depending on the structure and nature of reliable knowledge sources in each domain. For example, while NHS[9] and CDC[10] provide information for the layperson in the form of semi-structured articles, other sources such as PubMed[11] provide access to scholarly articles. Other sources, such as TRIP,[12] reside somewhere midway in the spectrum, while providing reliable and trustworthy information. We presume

[7] https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

[8] https://www.bbc.co.uk/news/technology-52632909.

[9] https://www.nhs.uk/.

[10] https://www.cdc.gov/.

[11] https://pubmed.ncbi.nlm.nih.gov/.

[12] https://www.tripdatabase.com/.

a similar landscape might characterize other domains such as scientific domains (e.g., fake news around climate change) and history (e.g., painting a nonfactual picture of historical events). We have found scanty usage of authoritative knowledge sources even among supervised methods for fake news detection; a notable work is MedFact [23], which targets to adopt principles from evidence-based medicine, albeit superficially, using information retrieval methods, in the process of medical fake news verification.

4.2 Statistical Data for Fake News Detection

Consider a particular fake claim that was made by Gerard Batten, a British politician, in the context around Brexit. The claim, illustrated in Fig. 6, says that there are only approximately 100 lorries that cross the border between the UK and Ireland in the island of Ireland. This was promptly debunked by various fact-checking agencies and media in the UK and Ireland, including TheJournal[13] and FactCheckNI.[14] Both of them pointed to a reference from a UK parliamentary report[15] that indicated that there were 177k *heavy goods crossings* across the Irish border each month, which equated to 5.9k such crossings each day on an average. There are two key aspects to this fact-checking effort: identifying that lorries refer to heavy goods vehicles and that the daily crossings can be computed from aggregate numbers. While the former is a task that relies on NLP and domain knowledge, the latter involves mathematical calculations, an elementary one, that of division, in this case. Such statistical claims appear all the time in the political domain, and those may involve population statistics of religious groups (heavily employed by the right wing in India) among others. Debunking these often involves the following steps:

- **Identification:** Identifying the pertinent statistic from an authoritative source, along with information about it.
- **Data Processing:** Normalization, interpolation, or extrapolation to enable direct comparison with the statistic in the claim.
- **Domain Conditioning:** Conditioning the processed statistic on well-understood patterns in the domain. For example, a high population growth rate is often correlated with low economic conditions and thus needs to be conditioned on the latter, to enable comparison across different cohorts.
- **Comparison and Veracity Assessment:** Once the data is processed and the comparable statistic identified, it may then be compared with the statistic in the claim and the veracity assessed in the backdrop of the knowledge of the domain patterns.

[13] https://www.thejournal.ie/factcheck-lorries-4469494-Feb2019/.

[14] https://factcheckni.org/fact-checks/is-border-trade-0-5-of-uk-eu-trade/.

[15] https://publications.parliament.uk/pa/cm201719/cmselect/cmniaf/329/32906.htm.

Fig. 6 Brexit fake news example

Based on informal conversations with a UK-based fact-checking agency, we learnt that a significant number of fake claims that they perform manual fact checking on involve statistical analysis and number crunching. This might also be seen as a fake news detection problem that is hard to be analyzed from within the supervised learning framework due to the very nature of the task, making bespoke UFND likely the best mode of attack for the task.

4.3 Early Detection

A number of existing supervised learning methods for fake news detection make abundant use of propagation information in order to identify fake news. These, due to their design, are incapable of addressing emerging fake news accurately,

since the dense feature footprint would need to accumulate before accurate veracity computation can be performed. Thus, unsupervised methods may be the only resort until the news has time to pass through the network enough to amass a significant digital footprint. This demarcates a niche space for unsupervised fake news detection.

4.4 Miscellaneous

We now outline a few other promising directions for unsupervised fake news detection.

Maligning Brands Through Fake Information

Within the space of e-commerce, there has been an increasing trend of using fake information to malign particular brands [33] or particular stores. These could differ on the basis of the kind of narrative used, and one in which historical labelled information may be of limited utility, making this a fertile area for unsupervised fake news detection. These also include reviews about brands posted on trading websites as well as maps; recent studies have established the utility of emotion and sentiment information in fake review detection [15].

Explainability in UFND

There is an increasing appreciation that any ML algorithm should not just provide its decision but also a rationale supporting the decision. Facilitating user engagement was also highlighted in the EU High-Level Expert Group report on disinformation [3], in the interest of ensuring that democratic practices be upheld and the diversity of the media ecosystem be preserved. This makes explainability or other forms of enhancing interpretability an interesting area for fake news detection in general. In fact, unlike supervised methods, unsupervised methods (and their designers) cannot relegate the decision to historical labelled data and hold more liability for the decision.

5 Conclusions

In this chapter, we provided a bird's-eye view of work in unsupervised fake news detection. In what we designed as a unique perspective, we endeavored to provide a critical analysis that is accessible to an informed layperson (rather than just the machine learning specialist). We started off by situating unsupervised methods

among the plethora of paradigms in machine learning and outlined the specific challenges that are of high importance in unsupervised learning for fake news detection. This was followed by a conceptual analysis of UFND methods, a critical analysis of such conceptual foundations, and a listing of possible conceptual building blocks that may enhance both the existing UFND methods as well as provide a platform to design newer UFND methods. This was followed by a methodological analysis of UFND methods, along with a critical perspective outlining their limitations and strengths. We then concluded the chapter with a set of possible interesting directions to advance the frontier in unsupervised fake news detection.

References

1. Acerbi, A.: Cognitive attraction and online misinformation. Palgrave Commun. **5**(1), 1–7 (2019)
2. Anoop, K., Deepak, P., Lajish, L.V.: Emotion cognizance improves fake news identification. CoRR, abs/1906.10365 (2019). http://arxiv.org/abs/1906.10365
3. Buning, M.d.C., et al.: A multidimensional approach to disinformation. In: EU Expert Group Reports (2018)
4. Conroy, N.K., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015)
5. Dutta, H.S., Chakraborty, T.: Blackmarket-driven collusion among retweeters—analysis, detection, and characterization. IEEE Trans. Inf. Forensics Secur. **15**, 1935–1944 (2019)
6. Fisch, A.: Trump, JK Rowling, and confirmation bias: an experiential lesson in fake news. Radical Teach. **111**, 103–108 (2018)
7. Fu, K.S., Mui, J.: A survey on image segmentation. Pattern Recogn. **13**(1), 3–16 (1981)
8. Gangireddy, S.C., Deepak, P., Long, C., Chakraborty, T.: Unsupervised fake news detection: a graph-based approach. In: ACM Hypertext and Social Media (2020)
9. Guess, A., Nagler, J., Tucker, J.: Less than you think: prevalence and predictors of fake news dissemination on Facebook. Sci. Adv. **5**(1), eaau4586 (2019)
10. Herath, S., Harandi, M., Porikli, F.: Going deeper into action recognition: a survey. Image Vis. Comput. **60**, 4–21 (2017)
11. Jamsheela, O., Raju, G.: Frequent itemset mining algorithms: a literature survey. In: Proceedings of the 2015 IEEE International Advance Computing Conference (IACC), pp. 1099–1104. IEEE, New York (2015)
12. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: a survey. J. Artif. Intell. Res. **4**, 237–285 (1996)
13. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)
14. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM Sigkdd Explor. Newsl. **17**(2), 1–16 (2016)
15. Melleng, A., Jurek-Loughrey, A., Deepak, P.: Sentiment and emotion based representations for fake reviews detection. In: Mitkov, R., Angelova, G. (eds.) Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP) 2019, Varna, Bulgaria, 2–4 September 2019, pp. 750–757. INCOMA Ltd., New York (2019). https://doi.org/10.26615/978-954-452-056-4_087
16. Murungi, D., Yates, D., Purao, S., Yu, J., Zhan, R.: Factual or believable? negotiating the boundaries of confirmation bias in online news stories. In: Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)

17. Orlov, M., Litvak, M.: Using behavior and text analysis to detect propagandists and misinformers on twitter. In: Annual International Symposium on Information Management and Big Data, pp. 67–74. Springer, Berlin (2018)
18. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
19. Pennycook, G., Rand, D.G.: Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. J. Pers. **88**(2), 185–200 (2020)
20. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: 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, pp. 2931–2937 (2017)
21. Richardson, R., Schultz, J.M., Crawford, K.: Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. NYUL Rev. Online **94**, 15 (2019)
22. Rubin, V.L., Conroy, N., Yimin, C.: Towards news verification: deception detection methods for news discourse. In: Hawaii International Conference on System Sciences (2015)
23. Samuel, H., Zaiane, O.: Medfact: towards improving veracity of medical information in social media using applied machine learning. In: Canadian Conference on Artificial Intelligence, pp. 108–120. Springer, Berlin (2018)
24. Settles, B.: Active learning literature survey. In: Technical Report University of Wisconsin-Madison Department of Computer Sciences (2009)
25. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: a survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol. (TIST) **10**(3), 1–42 (2019)
26. Shu, K., Wang, S., Liu, H.: Exploiting tri-relationship for fake news detection, vol. 8 (2017). arXiv preprint:1712.07709
27. Shu, K., Bernard, H.R., Liu, H.: Studying fake news via network analysis: detection and mitigation. In: Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining, pp. 43–65. Springer, Berlin (2019)
28. Singh, I., Deepak, P., Anoop, K.: On the coherence of fake news articles. CoRR abs/1906.11126 (2019). http://arxiv.org/abs/1906.11126
29. Smith, G.D., Ebrahim, S.: Data Dredging, Bias, or Confounding: they can all get you into the BMJ and the Friday Papers (2002)
30. Strapparava, C.: Emotions and NLP: future directions. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (2016)
31. Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S., de Alfaro, L.: Some like it HOAX: automated fake news detection in social networks. arXiv preprint:1704.07506 (2017)
32. Tsymbal, A.: The problem of concept drift: definitions and related work. Comput. Sci. Dep. Trinity Coll. Dublin **106**(2), 58 (2004)
33. Visentin, M., Pizzi, G., Pichierri, M.: Fake news, real problems for brands: the impact of content truthfulness and source credibility on consumers' behavioral intentions toward the advertised brands. J. Interact. Mark. **45**, 99–112 (2019)
34. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
35. Wang, W.Y.: "liar, liar pants on fire": a new benchmark dataset for fake news detection. arXiv preprint:1705.00648 (2017)
36. Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., Liu, H.: Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 5644–5651 (2019)
37. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. IEEE Trans. Knowl. Data Eng. **20**(6), 796–808 (2008)
38. Zarocostas, J.: How to fight an infodemic. Lancet **395**(10225), 676 (2020)
39. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Synth. Lect. Artif. Intell. Mach. Learn. **3**(1), 1–130 (2009)