An Overview (and Criticism) of Methods to Detect Fake Content Online

5

Antonio Badia
University of Louisville, USA

INTRODUCTION

The Web facilitates the spread of information thanks to its interconnected nature and the ease of publishing on it; but there seems to be lately a drastic increase in content of doubtful veracity (Kumar ans Shah, 2018). Given the use of the Web as a diffusion media (60% of Americans get their news from social media, according to (Allcott et al, 2017)), this has become an important issue. In the Web, it is relatively easy to produce content that maximizes dissemination (achieves 'virality') by using attention-calling techniques (like clickbait) that take advantage of recommendation algorithms. As a result, there is a process of 'algorithmic amplification' of fake content (DiResta, 2018).

This has produced alarm, as false and misleading information is reaching wide audiences and doing it faster than truthful, accurate content (Shao et alia, 2018; Hui et alia, 2018). Therefore, there is much interest in the research community (and society at large) in detecting certain forms of fake content and eliminating (or at least restricting) its diffusion.

In this chapter, we provide an overview of recent algorithmic approaches to detecting fake content. This is a very active and ongoing area of work; we will not offer a comprehensive review of all relevant research, rather offering a representative sample. The thesis of this chapter is that such research is characterized by a very narrow scope and a lack of definition of its target (i.e. what exactly is fake content?). To support this thesis, we first examine the concept and show that it is a multi-faceted, complex phenomena; hence, it is very difficult to agree on an exact definition of what constitutes falsehood. As a result, most research has focused on a narrow subset of the topic (usually called fact checking). Next, we summarize research efforts to detecting fake content, and provide a brief evaluation of the state of the art. Finally, we sketch some suggestions for future research, emphasizing that this is still an open problem and that further work will require a better approach to defining fake content and its various aspects.

We do not cover other, closely related aspects of the problem, like legal, political and criminal approaches to defining, detecting and fighting fake content. Such aspects are very interesting and relevant, but deserve a chapter of their own. Rather, we assume that technological efforts to detect false content are an important tool that can be used by these other approaches, but to do so it must evolve past current attempts.

BACKGROUND: DEFINING FAKE CONTENT

One of the most challenging aspects of the research about fake content is the difficulty of defining the concept. There is much disagreement among authors; while the idea of being 'true' or 'false' has a strong intuitive sense, there is a lack of formal definitions that are widely shared. A considerable amount of work does not formally define 'fake' or 'false' (or, equivalently, 'true' or 'truth'). Thus, ad-hoc defini-

DOI: 10.4018/978-1-5225-9715-5.ch072

tions are used in many cases (Shu et alia, 2017). For instance, a Stanford study on misinformation uses a set of fake content obtained by combining articles from the PolitiFact web site, the Buzzfeed website, and two previous academic articles (Allcot et alia, 2019). classifies as fake news any post from a short list of sites which are 'well known to be providers of false information." To make the issue even more complicated, there are a number of related concepts (fake reviews, clickbait, rumors, hate speech, cognitive hacking) that tend to get confused (that is why this article uses the more neutral label 'fake content' (Tandoc et al. 2017)).

An area that has looked in depth at the problem of true or authentic information is that of Intelligence studies; this area provides a starting point for trying to define false news (Hansen, 2017). Based on this work, we can distinguish the following aspects:

- **'Fake' as 'Non-Factual':** This refers to statements describing events or facts that are considered not to be a faithful representation of what happens in the real world. Many studies implicitly use this idea; for instance, "Fake news is an article that is intentionally and verifiably false." (Shu et al., 2017, p. 23). Note that this approach requires the existence of some ground truth that can be objectively assessed; this is easy in some cases (location of a store, hours it is open) but not so in others. Factual statements may involve several aspects:
  - Factual concrete information, that is, about a particular event or action. The falsehood usually refers to describing an event that did not happen, and denying the occurrence of an event that did happen. Most research in fake content focuses on this aspect of the issue.
  - **Factual General Information:** This refers to general or scientific knowledge. An example of this is how medical knowledge is distorted by many pseudo-scientific theories that proliferate on the Web, like anti-vaccine beliefs . Another example is climate change denial. There is a debate as to whether these platforms actively contribute to maximize the impact of this pseudo-information, due to their algorithms (DiResta, 2018).
- **'Fake' as 'Incomplete/Misleading':** For complex events or actions, an accurate and complete description may involve complex statements. A partial description presenting carefully selected aspects, with each aspect factually true, may create a false impression: "omitted facts or untold stories which, if viewed by the standard of traditional editorial guidelines, would definitely have been considered newsworthy." (Hanson, 2017; p. 21). This is usually achieved by suppressing some relevant aspects and/or highlighting barely meaningful ones, and can be considered a falsehood in the sense that the significance or interpretation of the event or action in a larger context is completely hidden.
- **'Fake' as 'Biased':** The description is done from one perspective only, resulting on a slanted view of the event or action: "the reporting may be so one-sided as to disqualify it. It may not necessarily contain untruths, but it is done less to inform than to leave the news consumer with a certain set of emotions and, ultimately, with certain political preferences." (Hanson, 2017, p. 21). Note that, different from the previous case, where each individual statement was truthful but the collection was not, here the individual statements (whether there is only one or several) are distorted.
- '**Fake' as 'Opinion':** This refers to non-factual information: opinion, commentary on news, and similar. This is the most ambiguous area, since it is assumed that when someone gives an opinion, the speaker is not bound to be objective. This area includes speech which is conflictive for other reasons, like hate speech; however, given its ambiguity, hate speech is usually not considered a target for fake content detection, with a few specific exceptions: in e-commerce, *fake reviews* is