Annotating Credibility: Identifying and Mitigating Bias in Credibility Datasets

Dimitrios Bountouridis
d.bountouridis@tudelft.nl
Delft University of Technology, The Netherlands

Mykola Makhortykh
m.makhortyhk@uva.nl
Amsterdam School of Communication Research, The Netherlands

Emily Sullivan
e.e.sullivan-mumm@tudelft.nl
Delft University of Technology, The Netherlands

Jaron Harambam
j.harambam@uva.nl
Institute for Information Law, University of Amsterdam, The Netherlands

Nava Tintarev
n.tintarev@tudelft.nl
Delft University of Technology, The Netherlands

Claudia Hauff
c.hauff@tudelft.nl
Delft University of Technology, The Netherlands

ABSTRACT

In the current post-truth era, online information is consistently under scrutiny with respect to its credibility (its quality and veracity). Computer science has been prolific in developing automated solutions for relevant tasks such as claim verification or bias estimation. However, the *validity* of such solutions relies heavily on their training and evaluation datasets. Inevitably, systematic and methodological errors (known as data biases) might appear during their compilation. We survey 12 published and freely-available datasets and annotate them for data biases using an established theoretical framework. We employ three expert annotators coming from the disciplines of computer science, philosophy, communication science and show that indeed, all annotated datasets suffer from biases.

1 INTRODUCTION

The internet has changed not only the way that people consume information, but also how that information is created and disseminated. People have virtually unlimited access to information from all kinds of heterogeneous sources and produce many different forms of knowledge, which are easily circulated through various platforms.

Although this free flow of information has enabled a deep democratization of knowledge, in recent years it has also shown its counter side. Disinformation, i.e., “false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit” [6], is rampant online and has become object of much public concern [1, 16]. These developments have placed at the center of public and academic attention the issue of *credibility*, that is the quality and veracity of the information disseminated [18]. Numerous politicians, policy makers, scientists, journalists, and activists have in recent years argued to address credibility by means of exposing the so-called fake news, hoaxes, rumours and political biases [4, 34].

The field of computer science has similarly taken the challenge by developing automated solutions for a series of credibility-related tasks. However, differentiating true from false claims or exposing ideological biases is in itself a difficult job and hardly uncontroversial [12]. Valid computational solutions require proper training and evaluation on data that capture the specifics of the phenomenon in question. This implies formalized definitions of the concepts in focus, and credibility is a complex and multifaceted one [18]. Scaling this process up by means of machine learning techniques makes it even more complex [13, 14]. Nevertheless, this has not stopped computer scientists from compiling and releasing data sets related to credibility tasks.

The main problem is that datasets are commonly seen and used as simply reflecting reality, although they are prone to *data biases*: systematic distortions that compromise their representativeness, and can potentially lead to societal inequalities or other serious consequences [8, 24]. Critical data scholars therefore urge researchers to scrutinize their data sets, so to understand their limitations and consequences [23]. Others call for more transparency regarding the datasets’ background, intended use and ethical concerns [10, 19]. Credibility-related datasets have received little scrutiny, even though their source of data (e.g., journalistic initiatives) and theoretical foundations of the relevant tasks (e.g., fact-checking) have met serious criticism [5, 12]. But who should scrutinize such datasets? The study of credibility is argued to be highly interdisciplinary [18] and every discipline has a different image of the data and therefore, its interpretation [11].

This paper employs experts from three disciplines (computer science, philosophy, and communication science) to survey 12 published datasets and explore to what extent they exhibit data biases and their societal implications. Our major contribution is a critical, multi-perspective look at existing credibility related datasets in order to make computer science aware of biases (both theoretically and concretely), and to offer ways to deal with them.

2 METHODOLOGY

Our methodology comprises two components: first, the gathering of datasets and their associated publications; and secondly, their annotation of exhibited data biases by three experts.

First, we surveyed conference and workshop proceedings by querying Google Scholar[1] with pairwise permutations of the terms

ROME 2019, July 25, 2019, Paris, France

[1] scholar.google.com

"credibility", "bias", "fact-checking", "trust", "veracity", "rumor", "partisan", "fake news", "news" and "annotations" or "datasets". We then performed multiple web searches using the same terms to identify either datasets compiled outside academia or datasets offered by academic competitions (e.g., the Fever[2] competition). The whole process resulted in more than 50 datasets.

We filtered these datasets according to three criteria. First, we only considered those with published data, i.e., works that did not use proprietary and non-public data. Secondly, our focus was on datasets that sought to explicitly address societal concerns related to credibility. As such, datasets that could potentially be used, but are not compiled for credibility tasks, were filtered out (e.g., topic classification datasets). Thirdly, we filtered out any datasets that provide annotations strictly at the source-level (e.g., whether a publisher or website is credible or not). Assuming that the veracity of information should relate to the source rather than the content is a genetic fallacy. The filtering process resulted in 12 datasets (see Table 1).

The framework for recognizing and annotating data biases comes from Olteanu et al. [23]. We consider it the most complete since it is based on the commonalities between existing frameworks in the literature and their systematic organization into a more complete ontology. The details of the framework are presented in the Section 3. A first round of annotations was performed independently by three experts who studied the annotation framework individually; differences were discussed in a group setting during a second round leading to revised annotations.

It has been suggested that a person's positionality i.e., their personal biases emerging from their field of study, occupation and ethnic background etc., affect their perceptions [32]. Our annotators are post-doc researchers in computer science, philosophy and communication science. Two of the annotators identify themselves as "male" and one as "female", while their country or origin includes the U.S.A, Greece and Ukraine. Their annotations for each dataset are explained in Section 4 (and summarized in Table 3), and an overall discussion is presented in Section 5.

3 EVALUATION FRAMEWORK

The framework of Olteanu et al. [23] focuses on social data e.g., posts from social media platforms, and considers a generic data-processing pipeline of four steps: *i)* data acquisition-preparation, *ii)* data processing, *iii)* data analysis, and *iv)* evaluation-interpretation. In this paper, we focus on the first two steps since we are interested in the biases introduced during the compilation process.

The acquisition-preparation part comprises identifying the data requirements, locating the possible data sources, and finally collecting the data from the sources. During those steps, data biases can be incorporated due to *i)* certain challenges related to the behavior, demographics or expertise of the social-platform users, *ii)* the nature of the social platform, such as its interface or implied patterns of behavior, and *iii)* the collection step e.g., the way the platform's API is queried. For instance, imagine a dataset aiming to capture the global social media behavior. By using U.S. Twitter users, the dataset might incorporate population biases (the demographic characteristics of the sampled U.S. population differ from the global target), functional biases (due to the specific size limits of the tweets) and acquisition biases (due to the black-box nature of Twitter's API).

The data-processing part corresponds to biases introduced during manual annotations, cleaning, enrichment, and so on. On the same example, one can imagine biases introduced during an automatic sentiment annotation process using machine learning and trained on tweets from non-U.S. countries. The overall framework, with a more detailed description of the possible data biases along the generic data-processing pipeline, can be seen in Table 2. The same table also provides explanations for the colored labels used throughout the following Section 4. However, due to length constraints and the framework's complexity, the reader is advised to consult the original publication [23] for further information and more bias examples.

In this paper, we are largely interested in datasets *not* pertaining to social media users, that is the original framework's focus. As such, certain terms e.g., "population bias", might refer to authors, journalists or other entities. In the following section, we aim to make those distinctions clear. It should also be mentioned that for the following section, the encoded biases will come with a superscript corresponding to the particular annotator's field: computer science^1, philosophy^2 and communication science^3.

4 CREDIBILITY ESTIMATION

We focus on *credibility*-related tasks: given a claim, or article the goal is to estimate its veracity (the accuracy/truthfulness) or bias (prejudice for/against an idea). Both veracity and bias contribute to shaping the perceived credibility [37] and as such, we consider them to fall under the same umbrella. Verification, in contrast to bias estimation, usually requires supporting the estimate (e.g., *true* or *false*) with an analysis [35] or by cross-referencing the estimate with information from a knowledge base such as Wikipedia. However, the boundaries between veracity and bias estimation are not always well-defined.

In the following sections we present the 12 surveyed datasets (see Table 1) compiled for credibility-related tasks. These are grouped by the granularity of annotation, whether the credibility is assessed with respect to: claims and articles. It should be mentioned that a different grouping could have been used but without affecting the survey's findings.

4.1 Claim level

"Claim" corresponds to an argument, statement or opinion about a certain topic/event/story made usually by a public figure (politician, journalist) or entity (company, governmental agency). Claims can appear during an interview, an article, an editorial contribution to a publication, and in the form of posts in social media.

Fever. The FEVER [33] dataset contains 185k claims manually generated and verified against Wikipedia pages. Claims were initially generated by 25 annotators who manually paraphrased texts from the Wikipedia pages. The annotators were further asked to generate claim variations such that their meaning either remained the same or was altered. Another 25 annotators were assigned to verify the previously created claims (classified as "supported","refuted" or "not enough info"). To minimize the annotators' personal biases the authors specifically instructed them to *not* incorporate their own knowledge or beliefs but did not assess if that was actually

[2] fever.ai

the case (PrcEnrich^{1,2,3} ExtBias^{1,2,3} CntnBias^{3}). At the same time, they instructed them to consider a dictionary of definitions as the only source of world knowledge. This dictionary comprises terms hyper-linked in the original Wikipedia sentence, accompanied with the first sentence from their corresponding page. This practice is potentially problematic considering the distinct norms for displaying and conveying information on Wikipedia that might impact the way claims are constructed (FncBias^{2,3} NormBias^{2,3}). All annotators were native U.S. English speakers and went through an initial training phase by the authors (or by other experienced annotators) that might have induced behavioral expectations (NormBias^{1}).

Liar. The LIAR dataset [36] contains 12.8k short statements pertaining a ten-year period from *politifact.com*; a Pulitzer Prize-winning nonprofit journalistic project aiming to evaluate the accuracy of claims in U.S. politics. We consider LIAR to be a prime example of a biased dataset for a number of reasons. First, for each of the statements, *politifact* provides an analysis of their truthfulness judgment (six-item scale e.g., “true”, “mostly false” or “pants on fire”), accompanied by supporting links and documents. However, for the LIAR dataset, the justification was ignored, as it was not machine-readable [33] (PrcClean^{1} PrcEnrich^{1}). Secondly, in order to verify the editors' judgment, the authors themselves went through a subset of 200 claims and reported a high agreement with the editors' ; however, the details of that process, including details about demographics or political-leanings, are not disclosed (PrcClean^{1,2,3}). At the same time, while the statements and annotations were gathered using *politifact*'s API, the sampling and querying strategies are not reported (not to mention that the API service and documentation seem currently unavailable) (ColAcq^{1,2,3} ColQue^{1,2,3}). Additionally, *politifact*'s definition of worthy of fact-checking statements is not scrutinized (FncBias^{2,3}).

Emergent. The EMERGENT collection [30] is the outcome of a journalistic effort from the Tow Center for Digital Journalism that comprises 300 rumors, called “claims” in the publication, and 2.5k expert-annotated news articles that pertain to the claims. EMERGENT was introduced as an NLP dataset by Ferreira and Vlachos [9] who used it for the task of stance classification: given the headline and body of text (article), the goal is to classify whether the body supports, refutes, is neutral or irrelevant to the headline. Claims were initially collected from various rumour sources including *snopes.com*[3] (similar to *politifact*), Twitter accounts (e.g., *@Hoaxaizer*) and Google alerts. For each claim, journalists browsed and searched for related articles on the web. Each article's veracity (“for”, “against” or “observing” i.e., neutral) was then annotated based on the journalist's opinion. EMERGENT suffers from a number of data biases. First, details about the journalists-annotators, their political leanings, demographics, annotation process and so on are not disclosed; thus, population biases and external biases might have led to biases in the annotations (PopBias^{1,2,3} ExtBias^{1,2,3} PrcEnrich^{1,2,3}). Another problematic aspect of EMERGENT is its reliance on rumour sites without vetting their credibility. The sampling strategy of claim-gathering from these sources is also undisclosed (ColAcq^{1,2,3} ColQue^{1,2,3} ColFil^{1,2,3}), although the authors acknowledge this limitation. This can potentially introduce some temporal biases (TmpVar^{3}), i.e., the dataset may capture a very narrow time period.

Credbank. The CREDBANK dataset [20] was developed for tackling false narratives and rumors in social media. It includes 60 million English-language Twitter posts spanning a three-month period annotated for credibility at the event-level. The dataset compilation process comprised three steps. After filtering out spam tweets were grouped into 45k clusters using Latent Dirichlet Allocation. However, we consider some parts of this process problematic e.g., the removal of tweets with more than three hashtags (PrcClean^{3}). To ensure that the computed clusters correspond to real-world events, a group of Mechanical Turk (MTurk) workers were hired to annotate the clusters of tweets into “event” or “non-event”. It is unclear whether the workers were instructed to consider rumors as events, although that is implied by the authors considering the next step (PrcEnrich^{1,2}). In the second step, each event was annotated by 30 MTurk workers for credibility based on the content of only a subset of the tweets themselves (PrcAggr^{1,2}). The annotation framework is based on linguistic aspects, similar to that of Saurí and Pustejovsky [28] (discussed later). Each event was annotated using a 5-point Likert scale from [-2] “Certainly Inaccurate” to [+2] “Certainly Accurate.” The annotation instructions do not deal with the workers' personal opinions regarding the credibility of an event (ExtBias^{1,2}). In addition, worker demographics are not disclosed, thus population (referring to the workers, not the tweets) biases may be present (PopBias^{1,2}).

FactBank. Saurí and Pustejovsky [28] presented an elaborate framework for annotating the factuality of claims into ten discrete classes (e.g., “fact”, “probable”, and others) based solely on their textual content; an extremely challenging task as the authors acknowledge, considering that factuality can be manifested through the interactions of various linguistic features. The framework aims to detach the annotator's world knowledge from the task by defining factuality as only relevant to the sources at play (e.g., the entities in the text). Under this framework, 9k claims from 208 articles in the TIMEBANK[25] and the AQUAINT TIMEML CORPUS (both contain temporal annotations of news articles/reports[4]) were manually annotated by two linguistic students. The sources of the articles U.S. based and include ABC, CNN, the New York Times and the Wall Street Journal. FACTBANK's main issue relates to the limited amount of annotators and their similar educational background (ExtBias^{1}). At the same time, the detailed annotation framework introduces functional biases (FncBias^{1}) leading to biases in the annotations themselves (PrcEnrich^{1,2}).

Pheme. The PHEME dataset [39] comprises 4.8k tweets (4.5k in English and 300 in German) corresponding to 330 rumour threads associated to nine news-worthy events. Which Twitter thread constituted a rumour was either based on an a-priori selection (PrcAggr^{1,3}) or was decided by journalists that analyzed breaking news in real-time (TmpVar^{2}) using the authors' previous rumor annotation scheme [38]. Tweets were annotated along three dimensions: support or response type (e.g., “supporting” or “denying”), certainty (e.g., “certain” or

[3] www.snopes.com

[4] www.timeml.org, retrieved April 2019

"uncertain") and evidentiality (e.g., "first-hand experience", "employment of reasoning", or "no-evidence"). 230 crowdworkers were employed to perform the annotations, and for each dimension the majority was considered as the final annotation; a possible aggregation bias (PrcAggr^{1,3}). For quality assurance, a subset of the annotations was compared to those of a single journalist with undisclosed background and biases. While the authors make a strong case regarding the annotation schema by supporting it with related studies, a number of things remain unclear. First, the annotation instructions are not disclosed and thus it is possible that ambiguous labels (e.g., "employment of reasoning" or "first-hand experience") were interpreted in multiple ways (the varying inter-annotator agreement supports ExtBias^{1,3}, NormBias^{1}) leading to annotation biases (PrcEnrich^{1,2,3}). Secondly, basic demographics for the crowdworkers are also missing, thus populations biases might be present (PopBias^{1,3}). Finally,Pheme relies extensively on Twitter whose norms for posting, re-tweeting or replying are quite specific (FncBias^{2,3}, PopBias^{2,3}, BhvrBias^{2,3}).

FacebookHoax. The FacebookHoax dataset [31] contains 15k Facebook posts from 32 ("non-hoax") and conspiracy pages ("hoax") pertaining a time period of six months during 2016. The authors assume that any post from a hoax page is "false" and vice versa; therefore, FacebookHoax provides annotations largely at the source level. The dataset has a range of problems. First, the distinction between hoax versus non-hoax pages is static (leading to temporal biases TmpVar^{2,3}) and based on the work of Bessi et al. [2] which lacks justification. This implies population and external biases (where "population" corresponding to the authors) that translate to biases during annotation (PopBias^{1,2,3} ExtBias^{1,2,3} PrcEnrich^{1,2,3}). It should be noted that Bessi et al. acknowledge the limitations of their annotation but FacebookHoax does not. Finally, it should be mentioned that all Facebook pages considered are Italian and thus further population biases (corresponding to the Facebook users) might be present (PopBias^{1,3}).

Décodex. In 2009, Le Monde[5] launched the Décodex project, under which a group of journalists compiled thousands of claims and stories in social media, blogs and news websites grouped by the labels "faux" (fake) or "hoax" . For the "faux" label, the journalists provided a short description and link to a website supporting their decision. According to Venturini et al. [34], the dataset stirred much debate in the French media. The annotation process lacked any transparency i.e., no mention of the annotation framework, the definitions of "faux" and "hoax", the amount of journalists-annotators per item, the background of the journalists and so on. This implies population-, external- and normative biases (PopBias^{1,3} ExtBias^{1,2,3} NormBias^{1,2,3}) leading to annotation biases (PrcEnrich^{1,2,3}). At the same time, the acquisition process of the claims and stories was not disclosed, leading to data collection biases (ColAcq^{1,2,3} ColQue^{1,2,3} ColFil^{1,2,3}).

4.2 Article level

An "article" typically discusses current or recent news pertaining various topics (i.e., politics, sports, finance and others). A news article typically appears on newspapers, magazines and their corresponding websites, while it can be circulated via social media.

SemEval2019 Task 4. A recent dataset comes from Task 4 of the 2019 SemEval[6] series. The task is to decide whether an unknown news article or source (publisher), follows a hyperpartisan argumentation, that is whether "it exhibits blind, prejudiced, or unreasoning allegiance to one party, faction, cause, or person." The dataset comes in two parts; for the first part of 750k articles, the binary annotations at the publisher level (hyperpartisan or not) come from two sources: BuzzFeed News[7] journalists or the *MediaBias-FactCheck.com* project that is similar in nature to *politifact*. Half the publishers and articles are hyperpartisan and these are equally divided between political left and right. The major problem of this part of SemEval's dataset is its reliance on external sources of credibility and political-leaning annotations with undisclosed biases (ExtBias^{1,2,3} leading to PrcEnrich^{1,2,3}). The lack of transparency regarding the acquisition, querying and filtering of the articles in the set is also problematic (ColAcq^{1,2,3} ColQue^{1,2,3} ColFil^{1,2,3}). Not to mention the genetic fallacy which comes with assuming that each article follows the publisher's hyperpartisan-or-not argumentation. The second part of the dataset provides annotations at the article level for 650 items. Crowdworkers were used as annotators and only the articles with high inter-annotator agreement were selected. However, no further details are disclosed regarding the annotators, process (e.g., what is the communicated definition of "hyperpartisanhsip") and overall compilation of this part, as such a number of biases might be present (PopBias^{1,2,3} ExtBias^{1,2,3} ColAcq^{1,2,3} PrcEnrich^{1,2,3}).

NewsTrust. The NewsTrust [21] dataset consists of 47k news articles crawled from 5.7k sources spanning from 1939 to 2014. The articles and 668 of the sources are rated with regard to qualitative aspects like objectivity, correctness of information, bias and credibility, by the *newstrust.net*[8] community members. The ratings between the 6k community members are also available. While this trust-network, wisdom of the crowds approach has its merits, it also has shortcomings. First, NewsTrust lacks a well-defined framework for annotating the concepts of objectivity, bias, etc.; as such, annotators rely on their personal interpretation of those notions and their corresponding rating scale (PrcEnrich^{1,2}). Secondly, there is no mention of quality control: it is possible that annotators rated sources and articles without investigating their content (ExtBias^{1} PrcEnrich^{1}). Thirdly, data regarding the demographics of the *newstrust* members is not disclosed, thus possible population and/or behavioral biases might be present (PopBias^{1,2,3} BhvrBias^{2}). Finally, NewsTrust offers only the final ratings for the news sources, articles and annotators and thus temporal analysis on that data is impossible; the per-year distribution of articles is also unknown (TmpVar^{1,3}).

BuzzFace. The BuzzFace dataset [27] is based on a BuzzFeed report[9] pertaining 2k news articles posted on Facebook during September 2016 by ten U.S. news outlets. The outlets were labeled by the publisher's journalists as political "right" or "left" and the stories

[5] www.lemonde.fr

[6] alt.qcri.org/semeval2019

[7] buzzfeednews.com

[8] Website currently unavailable

[9] www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis

as “mostly true”, “mostly false”, a “mixture of true and false”, and “no factual content”. BuzzFace further extends the annotations by enriching them with hundreds of thousands of Facebook and other comments from similar platforms. BuzzFace suffers from a number of issues. First, to make a case regarding the quality of the veracity annotations and address external biases, the authors argue that Buzzfeed reporters’ careers rely on their ability to judge news reports; a rather weak argument that is not verified (ExtBias^{1} leading to PrcEnrich^{1,2}). The authors mention that the journalists would even change their decision based on feedback; a process not explained thoroughly in the paper (PrcClean^{1} PrcEnrich^{1}). Secondly, the data pertains a very short period of time, thus temporal biases might be present (TmpVar^{2,3}).

FakeNewsNet. Similar to BuzzFace, FakeNewsNet [29] provides “context” in addition to the annotated articles themselves. This translates to 23k articles labelled as “true” or “fake” by *politifact*, *gossipcop.com*[10], and by *E! Online*[11] (considered as “true” by default), accompanied by related tweets e.g., those directly quoting the title of an article, and their metadata. FakeNewsNet suffers from a number of issues. First, similarly to any dataset that bases its credibility annotations on journalistic initiatives or sources, it comes with external biases (referring to journalists/experts) leading to annotation biases (ExtBias^{1,2,3} PrcEnrich^{1,2,3}). Secondly, biases emerge from *politifact*’s, *gossipcop*’s APIs and the heuristics applied during the crawling process for locating articles (ColAcq^{1,2} ColQue^{1,2,3}). For example, negative-sentiment words were excluded during the crawling process without proper justification (ColFil^{2}).

5 DISCUSSION

The overall annotation of biases as seen in Table 3 provides a number of insights. First, all studied credibility datasets display bias, with some types of bias more frequent than others: *Enrichment* biases, as part of the process, were found in all 12 datasets (7 for which all annotators agreed); *External* bias, as part of the source, for 11 datasets (6 complete agreement); *Population* bias, as a general challenge, for 7 datasets (5 complete agreement). Secondly, while a number of biases were independently annotated by all three experts, others were identified by one or two experts. This partially supports our decision for using a multidisciplinary team of annotators. Thirdly, certain biases e.g., population, and external, seem to *not* be mutually exclusive. Discussion with the annotators revealed that the definitions of certain biases (provided by Olteanu et al. [23]) can be ambiguous and overlap with each other. We now discuss the most prominent biases in Table 3.

External and population biases . The most prominent biases are external and population biases. According to our annotations, this relates to the datasets being dependent on fact-checking services (*politifact.com* or *snopes.com*), journalists and crowdworkers.

Indeed, some fact-checking services have been shown to be unbiased at least in terms of language usage between political parties [4]. At the same time, many follow a code of principles defined by the International Fact-Checking Network (IFCN)[12]. However, Brandtzaeg et al. [3] showed that both journalists and social media users treat such services with positive views but also, with a mix of skepticism. The authors speculate that this largely relates to the overall lack of transparency concerning which claims are chosen to be fact-checked, and how their credibility is determined. Most of the datasets presented in this paper did not reflect on these considerations.

Considering the worryingly low levels of trust in public institutions and journalism related to polarisation and undue political influence [22] it is surprising that a number of datasets use annotations only from journalists, whether these are part of fact-checking services or employed as annotators for the particular study. Journalists are indeed experts, but typically research works fail to disclose crucial information i.e., the journalists’ political biases, their selection process and their demographics.

Alternatively, a number of works use crowdsourcing services to annotate their data. Similarly to the journalists; important information about the workers’ background, demographics and so on, is rarely disclosed. Furthermore, most works aim for a high inter-annotator agreement as a measure of annotation quality. However, annotator agreement might not even be desirable: Drosou et al. [7] argue that each person has a personal interpretation of the facts that should probably be included if aiming to capture the “wisdom of the crowds.”

Enrichment biases . The nature of the veracity classification is typically of low granularity i.e., it allows for a claim/article to be either “true” or “false”. While a convenient scheme for automatic evaluation, this does not allow annotators to express a shade of uncertainty. In addition, most works do not include in their corresponding publication the explanations of high level concepts provided to the annotators. For valid computational solutions, the notions of veracity, bias, or credibility should be formalized and clearly defined. These concepts are prone to subjective interpretation, but most dataset papers make little effort to minimize its effects.

Finally, annotation instructions are often not disclosed with the publication. This lack of transparency of the process is particularly relevant for datasets/works aiming to solve issues of high societal importance.

Normative and functional biases. Our discussions so far had a clear focus on delineating the biases that human judgement incorporates. However, there are also biases reflected in the claims (or articles) themselves. And although these were not frequently annotated, they are worthy of a more elaborate discussion. For example, speech acts take on different meanings depending on the platform where the claims were made, embodying functional and normative biases. For example, Twitter’s character limit gives rise to different linguistic features compared to other platforms, such as Facebook [17, 26]. Furthermore, depending on the intended function of a claim, the same words can have different truth evaluative markers e.g., a political slogan, satire, or a claim about the world.

5.1 Remedies

No representation of reality, such as data sets, can ever be neutral or non-biased; thus, removing data biases altogether is an impossible goal. Instead, the task is to assess whether a bias has an empirical or theoretical foundation Cazalens et al. [5]. Graves [13] argues that

[10] www.gossipcop.com
[11] www.eonline.com
[12] ifcncodeofprinciples.poynter.org

fact-checking in particular requires human judgement and sensitivity to context that is hard to automate. Similarly, Wathen and Burkell [37] identify a range of factors that influence the credibility of online information, such as trustworthiness, plausibility or similarity to the reader's beliefs. These concerns relate to on-going philosophical debates over truth conditions, including the relationship between normative judgments and empirical facts.

Considering that a non-biased data set cannot exist, we propose two general strategies to deal nevertheless with undesired and unknown data biases: diversity and transparency.

Diversity: Reducing data bias (where possible). Alleviating the nefarious effects of data biases is firstly possible by focusing on the goal of *diversity* [15]. In our context, this translates into two practices: collecting/using data from a wider range of sources and places and having a more diverse range of annotators. Since credibility related datasets depends heavily on the quality of people who label the data, diversifying them in terms of class, ethnicity, ideology and so on is important. Obtaining a variety of perspectives, means getting one closer to a truthful, less biased, reality. Note that diversity in annotation also implies multi-perspective point of view from the authors' side. Multiple disciplines and backgrounds should be involved in all steps of the data set compilation process: from its initial conceptual design, data acquisition, to the final post-processing, and actual usage.

Transparency: Acknowledging data bias (where impossible to reduce). The second remedy we offer is in line with various other approaches for making data sets collectors and users more aware of the characteristics and limitations of the data sets [10, 19, 23]. This implies a *continuous* focus on the notions of *intended use/scope*, the *epistemic goals* and *limitations* of the data sets.

Intended usage encapsulates the intended goals and values the dataset aims to capture. Credibility datasets have implicit *epistemic goals*: promoting truth or objectivity over misleading information or hyper-partisan information. Making explicit the intended epistemic goals, exposes the values behind the data. This allows critical reflection on whether a given data set is best suited to capture the intended value, and the academic community can discuss whether the intended value is actually worth designing for.

Intended scope encapsulates the extent of the generalization ability of the dataset; and thus, its *limitations*. Unfortunately, while an intuitive, scientific practice, a majority of the datasets do not clarify the scope and seem to imply generalization across platforms and contexts. Acknowledging that Facebook users are not representative of all social media users, or that journalists are experts with biases, are examples that can help us understand the dataset's scope and limitations.

Considering that it is impossible to remove all biases, it is important to acknowledge that de-biasing is never finished. Researchers must always be attentive to possible data distortions and reflect on when these significantly impact validity and when they do not. This remedy is a commitment to *continuously* engage in the above remedies and to look for other worthwhile approaches.

Toward the goal of transparency and accountability for the use of data in machine-learning tasks, Gebru et al. [10] recently proposed that every dataset should be accompanied with a document ("datasheet") explaining its intended use (i.e., task gap it aims to fill), funding (i.e., who paid for its creation), creation and composition (i.e., data gathering, filtering and annotation), and other properties. Similarly, Mitchell et al. [19] proposed an approach called "model cards" for machine learning and artificial-intelligence models. Such solutions should be extended to address the particularities of the credibility concept e.g., by placing a greater effort on how the concepts of bias/truth/hoax are communicated to the annotators, or under which criteria certain journalists are employed.

6 CONCLUSIONS

Considering the societal importance and public interest in the credibility assessment of online information this paper adopted the framework of Olteanu et al. [23] to analyze data biases in 12 datasets used for relevant tasks. For the sake of a multi-perspective view, we employed three experts in varying fields to perform the task independently.

Our study found that data biases are prevalent in credibility datasets. In particular, external, population and enrichment biases are frequent. Nevertheless, our findings should *not* discourage researchers to continue working on credibility solutions. Datasets can never be neutral or unbiased; like any other representation of the world around us, they are always produced by certain people, with a certain worldview, in a certain time, making certain methodological choices. Since datasets are not neutral or unbiased, future researches should not assume them to be such and should rather focus on *diversity* and *transparency*. For example, addressing the limitations of the data, adopting novel transparency frameworks (e.g., datasheets for datasets) or embracing multi-disciplinarity, that is the inclusion of experts from varying fields in all the steps of the dataset compilation process.

Our study also suggests further domain-specific investigations of dataset biases. Credibility assessment is only one of the many societal concerns pertaining information distribution online. News media organizations increasingly use computational systems for automatizing content production (e.g., robot journalism) or news distribution (e.g., personalized recommendations). Future studies can look into biases in these areas to understand their effect on the validity of the datasets and the works that use them.

REFERENCES

[1] Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

[2] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one* 10, 2 (2015), e0118093.

[3] Petter Bae Brandtzaeg, Asbjørn Følstad, and María Ángeles Chaparro Domínguez. 2018. How Journalists and Social Media Users Perceive Online Fact-Checking and Verification Services. *Journalism Practice* 12, 9 (2018), 1109–1129.

[4] Dallas Card, Lucy H Lin, and Noah A Smith. 2018. Politifact Language Audit. (2018).

[5] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A Content Management Perspective on Fact-Checking. In *Proceedings of the Web Conference 2018-alternate paper tracks" Journalism, Misinformation and Fact Checking"*. 1–10.

[6] Madeleine de Cock Buning, Richard Allen, A Bargaoanu, Anja Bechmann, N Curran, D Dimitrov, G Dzinich, D Frau-Meigs, F Fubini, K Gniffke, et al. 2018. Final report of the High Level Expert Group on Fake News and Online Disinformation. (2018).

[7] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in big data: A review. *Big data* 5, 2 (2017), 73–84.

[8] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[9] William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 1163–1168.

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[11] Lisa Gitelman. 2011. Notes for the Upcoming Collection âĂŸRaw DataâĂŹis an Oxymoron. (2011).

[12] Lucas Graves. 2016. *Deciding whatâĂŹs true: The rise of political fact-checking in American journalism*. Columbia University Press.

[13] Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. *Factsheet* 2 (2018), 2018–02.

[14] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1803–1812.

[15] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2 (2018), 191–207.

[16] Sheila Jasanoff and Hilton R Simmet. 2017. No funeral bells: Public reason in a âĂŸpost-truthâĂŹage. *Social studies of science* 47, 5 (2017), 751–770.

[17] Han Lin and Lin Qiu. 2013. Two sites, two voices: Linguistic differences between Facebook status updates and tweets. In *Proceedings of the international Conference on Cross-Cultural Design*. Springer, 432–440.

[18] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the Association for Information Science and Technology* 58, 13 (2007), 2078–2091.

[19] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2018. Model cards for model reporting. *arXiv preprint arXiv:1810.03993* (2018).

[20] Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*.

[21] Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 353–362.

[22] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2018. *Reuters Institute Digital News Report 2018*. Reuters Institute for the Study of Journalism.

[23] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[24] Cathy OâĂŹNeill. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. *Nueva York, NY: Crown Publishing Group* (2016).

[25] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, Vol. 2003. Lancaster, UK., 40.

[26] Matthew Rowe and Harith Alani. 2014. Mining and comparing engagement dynamics across multiple social media platforms. In *Proceedings of the ACM conference on Web science*. ACM, 229–238.

[27] Giovanni C Santia and Jake Ryland Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

[28] Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation* 43, 3 (2009), 227.

[29] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286* (2018).

[30] Craig Silverman. 2015. Lies, damn lies and viral content. (2015).

[31] Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506* (2017).

[32] David Takacs. 2003. How Does Your Positionality Bias Your Epistemology?. *Thought & Action* 19, 1 (2003), 27–38.

[33] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355* (2018).

[34] Tommaso Venturini, Mathieu Jacomy, Liliana Bounegru, and Jonathan Gray. 2017. Visual Network Exploration for Data Journalists. *The Routledge Handbook to Developments in Digital Journalism Studies* (2017).

[35] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*. 18–22.

[36] William Yang Wang. 2017. Liar, liar pants on fire: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[37] C Nadine Wathen and Jacquelyn Burkell. 2002. Believe it or not: Factors influencing credibility on the Web. *Journal of the Association for Information Science and Technology* 53, 2 (2002), 134–144.

[38] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 347–353.

[39] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.

7 SUPPLEMENTAL MATERIAL

Table 1: The datasets in our survey related to credibility estimation tasks ordered by year of release. Numbers are rounded up for the sake of brevity.

Dataset	Year	Focus level	#Items	Annotators	Claims Source	Articles Source
SemEval2019 Task 4	2018	Article + Source	750k articles, 600 sources	Journalists + crowdworkers	-	Various
Fever [33]	2018	Claim	185k claims	Undisclosed	*Wikipedia*	-
BuzzFace [27]	2018	Article	2k articles	Journalists	-	Various
FakeNewsNet [29]	2018	Article	23k articles	Journalists	-	Various
Liar [36]	2017	Claim	12.8k claims	Journalists	*politifact.com*	-
Décodex	2017	Claim + Article	271 claims, 10k articles	Journalists	Various	-
FacebookHoax[31]	2017	Claim + Source	15k posts, 32 sources	Authors	Facebook	-
Pheme[39]	2016	Claim	4.8k claims	Crowdworkers	Twitter	-
Emergent[9]	2016	Article + Claim	300 headlines, 2.5k articles	Journalists	Various	Various
NewsTrust [21]	2015	Article + Source	47k articles, 670 sources	Community	-	Various
CredBank [20]	2015	Claim + events	60m tweets	Crowdworkers	*Twitter*	-
FactBank [28]	2009	Claim	9k claims, 200 articles	Students	Various U.S.	Various

Table 2: Data biases and their definition along an idealized data processing pipeline [23]. For this paper, each bias comes with a colored code that is used as a reference when annotating the credibility datasets in Section 4.

Acquisition and Preparation			Processing
General Challenges	**Source**	**Collect**	
Population bias PopBias Biases due to differences in demographics or other user characteristics between a population of users represented in a dataset and a target population.	Functional biases FncBias Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment.	Acquisition ColAcq Biases that may occur during data collection e.g., due to programmatic access to platforms, limited access or opaque sampling strategies.	Cleaning PrcClean Biases that may occur during detecting and correcting errors and inconsistencies (via normalization, substitution of missing values) in the data.
Behavioral bias BhvrBias Biases due to differences in user behavior across platforms or contexts, or across users represented in different datasets.	Normative biases NormBias Biases that are a result of written norms or expectations about unwritten norms describing acceptable patterns of behavior on a given platform.	Querying ColQue Biases due to data access through an API that involves communicating a set of criteria for selecting, ranking, and returning the data being requested.	Enrichment PrcEnrich Biases that may occur while adding manual and/or automatic annotations to the data.
Content bias CntnBias Biases that are expressed as lexical, syntactic, semantic, and structural differences in the contents generated by users.	External biases ExtBias Biases resulting from factors outside the social platform, including considerations of socioeconomic status, ideological/religious/political leaning, education, social pressure, privacy concerns, topical interests, language, personality, and culture.	Filtering ColFil Biases due to unintentionally filtering out data items that might not be relevant for a study.	Aggregation PrcAggr Biases that may occur during data aggregation: structuring, organizing, representing or transforming the data.
Linking bias LinkBias Biases that are expressed as differences in the attributes of networks obtained from user connections, interactions or activity.	Non-individuals NonIndvl Interactions on social platforms that are not produced by individuals, but by accounts representing various types of organizations, or by automated agents.		
Temp. variations TmpVar Biases due to differences in populations or behaviors over time.			
Redundancy Redndcy Single data items that appear in multiple copies, which can be duplicates or near duplicates.			

Table 3: Full table of data biases as annotated by our group of multidisciplinary experts. Identified biases are shown with a tick-mark "✓". Each bias comes with superscript(s) corresponding to the annotators' expertise: computer science[1], philosophy[2] and communication science[3].

		FEVER	LIAR	CREDBANK	PHEME	FACEBOOKHOAX	DÉCODEX	SEMEVAL2019 TASK 4	EMERGENT	NEWSTRUST	BUZZFACE	FAKENEWSNET	FACTBANK
Process	Aggregation			✓^{1,2}	✓^{1,3}								
	Enrichment	✓^{1,2,3}	✓^{1}	✓^{1,2}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2}	✓^{1,2}	✓^{1,2,3}	✓^{1,2}
	Cleaning		✓^{1,2,3}	✓^{3}							✓^{1}		
Collect	Filtering						✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}			✓^{2}	
	Quering		✓^{1,2,3}				✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}			✓^{1,2,3}	
	Acquisition		✓^{1,2,3}				✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}		✓^{1}	✓^{1,2}	
Source	Non-individuals												
	External	✓^{1,2,3}		✓^{1,2}	✓^{1,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1}	✓^{1}	✓^{1,2,3}	✓^{1}
	Normative	✓^{1,2,3}			✓^{1}		✓^{1,2,3}						
	Functional	✓^{2,3}	✓^{2,3}		✓^{2,3}								✓^{1}
General challenges	Redundancy												
	Temporal				✓^{2}	✓^{2,3}			✓^{3}	✓^{1,3}	✓^{2,3}		
	Linking												
	Content	✓^{3}											
	Behavioral				✓^{2,3}					✓^{2}			
	Population			✓^{1,2}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,3}	✓^{1,2,3}	✓^{1,2,3}	✓^{1,2,3}			