A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles

Amy X. Zhang
MIT CSAIL
Cambridge, MA, USA
axz@mit.edu

Aditya Ranganathan
Berkeley Institute for Data Science
Berkeley, CA, USA
adityarn@berkeley.edu

Sarah Emlen Metz
Berkeley Institute for Data Science
Berkeley, CA, USA
emlen.metz@berkeley.edu

Scott Appling
Georgia Institute of Technology
Atlanta, GA, USA
scott.appling@gtri.gatech.edu

Connie Moon Sehat
Global Voices
London, UK
connie@globalvoices.org

Norman Gilmore
Berkeley Institute for Data Science
Berkeley, CA, USA
norman@virtualnorman.com

Nick B. Adams
Berkeley Institute for Data Science
Berkeley, CA, USA
nickbadams@berkeley.edu

Emmanuel Vincent
Climate Feedback
University of California, Merced
Merced, CA, USA
emvincent@climatefeedback.org

Jennifer 8. Lee
Hacks/Hackers
San Francisco, CA, USA
jenny@hackshackers.com

Martin Robbins
Factmata
London, UK
martin.robbins@factmata.com

Ed Bice
Meedan
San Francisco, CA, USA
ed@meedan.com

Sandro Hawke
W3C
Cambridge, MA, USA
sandro@w3.org

David Karger
MIT CSAIL
Cambridge, MA, USA
karger@mit.edu

An Xiao Mina
Meedan
San Francisco, CA, USA
an@meedan.com

ABSTRACT

The proliferation of misinformation in online news and its amplification by platforms are a growing concern, leading to numerous efforts to improve the detection of and response to misinformation. Given the variety of approaches, collective agreement on the indicators that signify credible content could allow for greater collaboration and data-sharing across initiatives. In this paper, we present an initial set of indicators for article credibility defined by a diverse coalition of experts. These indicators originate from both within an article's text as well as from external sources or article metadata. As a proof-of-concept, we present a dataset of 40 articles of varying credibility annotated with our indicators by 6 trained annotators using specialized platforms. We discuss future steps including expanding annotation, broadening the set of indicators, and considering their use by platforms and the public, towards the development of interoperable standards for content credibility.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3188731

KEYWORDS

misinformation, disinformation, information disorder, credibility, news, journalism, media literacy, web standards

ACM Reference Format:
Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer 8. Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3184558.3188731

1 INTRODUCTION

While the propagation of false information existed well before the internet [11], recent changes to our information ecosystem [24] have created new challenges for distinguishing misinformation from credible content. Misinformation, or information that is false or misleading, can quickly reach thousands or millions of readers, helped by inattentive or malicious sharers and algorithms optimized for engagement. Many solutions to remedy the propagation of misinformation have been proposed—from initiatives for publishers to signal their credibility[1], to technologies for automatically labeling misinformation and scoring content credibility [6, 29, 35, 42], to the

[1]The Trust Project: https://thetrustproject.org

engagement of professional fact-checkers or experts [13], to campaigns to improve literacy [19] or crowdsource annotations [31].

While all these initiatives are valuable, the problem is so multifaceted that each provides only partial alleviation. Instead, a holistic approach, with reputation systems, fact-checking, media literacy campaigns, revenue models, and public feedback all contributing, could collectively work towards improving the health of the information ecosystem. To foster this cooperation, we propose a *shared vocabulary for representing credibility*. However, credibility is not a Boolean flag: there are many indicators, both human- and machine-generated, that can feed into an assessment of article credibility, and differing preferences for what indicators to emphasize or display. Instead of an opaque score or flag, a more transparent and customizable approach would be to allow publishers, platforms, and the public to both understand and communicate what aspects of an article contribute to its credibility and why.

In this work, we describe a set of initial indicators for article credibility, grouped into *content* signals, that can be determined by only considering the text or content of an article, as well as *context* signals, that can be determined through consulting external sources or article metadata. These indicators were iteratively developed through consultations with journalists, researchers, platform representatives, and others during a series of conferences, workshops, and online working sessions. While there are many signals of credibility, we focus on article indicators that do not need a domain expert but require human judgment and training. This focus differentiates our work from efforts targeting purely computational or expert-driven indicators, towards broadening participation in credibility annotation and improving media literacy. To validate the indicators and examine how they get annotated, we gathered a dataset of 40 highly shared articles focused on two topics possessing a high degree of misinformation in popular media: public health [49] and climate science [21]. These articles were each annotated with credibility indicators by 6 annotators with training in journalism and logic and reasoning. Their rich annotations help us understand the consistency of the different indicators across annotators and how well they align with domain expert evaluations of credibility. We are releasing the data publicly[2], and will host an expanded dataset in service to the research community and public.

The process outlined in this paper serves as a template for creating a standardized set of indicators for evaluating content credibility. With broad consensus, these indicators could then support an ecosystem of varied annotators and consumers of credibility data. By focusing on indicators, we leave open the question of who or what performs annotation. Indeed, the presence of a standard permits flexibility on the part of users and platforms to determine whose annotations to surface based on who they consider trustworthy. Our approach also leaves open the question of how annotations are generated, from the use of partial or full automation, to annotations by experts or publishers, to crowdsourced or friendsourced methods. However, results from our initial study suggest certain methods may be more or less fruitful for different indicators.

Aligned with the goals of groups such as the W3C Credentials Community Group[3], using our indicators, any interested party could contribute annotations using open standards developed during this work, while any system for displaying or sharing news could make their own decisions about how to aggregate, weight, filter and display credibility information. For instance, systems such as web browsers, search engines, or social platforms could surface information about a news article to benefit readers, much like how nutrition labels for food and browser security labels for webpages provide context in the moment. Readers could also verify an article by building on the annotations left by others or even interrogate the indicators that a particular publisher or other party provides. Finally, the data may be helpful to researchers and industry watchdogs seeking to monitor the ecosystem as a whole.

[2]Credibility Coalition: http://credibilitycoalition.org
[3]W3C Credentials Community Group: https://www.w3.org/community/credentials/

2 RELATED WORK

In recent years, researchers have sought to better define and characterize misinformation and its place in the larger information ecosystem. Some researchers have chosen to eschew the popularized term “fake news", calling it overloaded [50]. Instead, they have opted for terms such as “information pollution" and “information disorder" [50], to focus not only on the authenticity of the content itself, but also the motivations and actions of creators, including disinformation agents [47], readers, media companies [20] and their advertising models [16], platforms, and sharers. Accordingly, our approach covers a broad range of indicators developed by experts representing a range of disciplines and industries.

An important aspect of characterizing misinformation is understanding how people perceive the credibility of information. Reviews of the credibility research literature [24, 41] describe various aspects of credibility attribution, including judgments about the credibility of a particular source or a broader platform (e.g., a blog versus social media) [43], as well as message characteristics that impact perceptions of credibility of the message or source [30]. Studies have pointed out the differences in perceived credibility that can occur based on differences in personal relevance [53], individual online usage [10], and the co-orientation of reader and writer views [25], among others. This prior work suggests that a one-size-fits-all approach or an approach that provides an opaque “credibility score" will not be able to adapt to individual needs.

However, research has also found that readers can be swayed by superficial qualities that may be manipulated, such as a user’s avatar on social media [27] or number of sources quoted in an article [48], demonstrating the need for greater media literacy. Another study found that fact-checkers correctly determine credibility using lateral searching, while non-experts fall victim to convincing logos [51]. In response, researchers have considered how interfaces could provide better context for gauging information quality, in areas such as Wikipedia [33], related articles in social media [4], and flags on disputed content [32]. By surfacing more nuanced signals about the credibility of an article, we hope to provide greater context to readers and platforms to make informed judgments.

There exists a significant amount of related work on computational models related to information credibility [29, 45]. Many models focus on aspects of language that can be a signal of low credibility [37], such as hedging [29] or biased language [39]. Researchers have also studied the linguistic characteristics of deceptively written content [54] and their relation to credibility [5], as

well as misleading headlines [8]. As social media is increasingly a space where misinformation is propagated, researchers have studied how rumors [3] as well as corrections [2] spread on social media. Building on this work, researchers have built models to predict credibility of social media posts [6, 17, 22], as well as tools for investigating rumors or claims [23, 38, 40]. In addition, researchers have focused on the credibility of individual claims or assertions within text [34]. In the area of computational fact-checking, researchers evaluate the truthfulness of claims by comparing concepts within knowledge graphs [9, 52]. Though the focus of this study is article indicators, all of these signals contribute to assessments of information credibility more broadly, and this prior work suggests some credibility indicators that might be automatable in the future.

Finally, our research is related to prior work on human annotation of credibility, such as annotations of social media [6, 26] or of television and newspaper content [14]. In contrast to this earlier work, we chose to use trained annotators as opposed to a random sampled population or Amazon Mechanical Turk workers, and we collected annotations about specific indicators instead of just overall credibility. These decisions allowed us to capture a richer and more informed characterization of credibility.

3 TOWARDS STRUCTURED CREDIBILITY INDICATORS FOR ONLINE JOURNALISM

The need for a common vocabulary around credibility became apparent at the first MisinfoCon[4], a conference dedicated to misinformation that saw many projects to define and classify misinformation and credibility but no easy way to communicate findings and data across projects. From an initial meeting at the conference, workshops in San Francisco and New York were convened, with over 40 representatives from journalism and fact-checking groups, research labs, social and annotation platforms, web standards, and more. A broader alliance called the *Credibility Coalition* emerged among these participants, with weekly remote working sessions. From these sessions, participants drafted over 100 indicator suggestions, taking example from existing credibility initiatives, such as Climate Feedback[5] and the Trust Project. As outside input is crucial for the success of this project, representatives presented to communities such as the Mozilla Festival[6] and the International Press Telecommunications Council Meeting[7], to publicize the work and host workshops for gathering feedback.

Over time, the indicators coalesced into 12 major categories, including reader behavior, revenue models, publication metadata, and inbound and outbound references. From this collection, 16 indicators were chosen for annotation. We chose article-level indicators that require human annotation from trained annotators but no domain expertise. Thus, we did not consider automated indicators for this study or ones that require significant expertise, such as domain knowledge of the subject matter, offline investigation, or data gathering requiring technical knowledge or access to proprietary data. As our current focus is articles, we chose to ignore indicators related to publishers, authors, or any multimedia content.

[4]MisinfoCon: https://misinfocon.com
[5]Climate Feedback process: https://climatefeedback.org/process/
[6]Mozilla Festival (MozFest), London, Oct 2017: https://mozillafestival.org
[7]IPTC, Barcelona, Nov 2017: https://iptc.org

3.1 Content Indicators

Content indicators are those that can be determined by analyzing the title and text of the article without consulting outside sources or metadata. We present the following 8 content indicators, their definitions, and what we asked of annotators.

Title Representativeness: Article titles can be misleading or opaque about the topic, claims, or conclusions of the content. Annotators were asked to rate the representativeness of the article title. If it was found unrepresentative, they were asked to clarify how the title was unrepresentative; for instance, by being off-topic, carrying little information, or overstating or understating claims.

"Clickbait" Title: "Clickbait" is defined as "a certain kind of web content...that is designed to entice its readers into clicking an accompanying link" [35]. Annotators were asked to rate the degree to which a headline was clickbait. If annotators rated a title as clickbait, they were asked to clarify the form of clickbait in a follow-up question, such as a "listicle" or a cliffhanger.

Quotes from Outside Experts: Articles often seek outside feedback from independent experts in the field. This additional validation provides support for the conclusions drawn and reveals a level of journalistic rigor [48]. For this indicator, we asked annotators to highlight where experts were quoted in the article.

Citation of Organizations and Studies: Journalists can also cite or quote from a range of organizations or scientific studies to add context or support to the article and enhance its credibility. We asked annotators to highlight where any scientific studies or any organizations were cited, as well as indicate whether the article was primarily about a single study.

Calibration of Confidence: The use of tentative propositions in writing, often quantified, allows readers to assess claims with appropriate confidence. We asked annotators to mark whether authors used appropriate language to show confidence in their claims, and to highlight sections of an article where authors acknowledge their level of uncertainty (e.g. hedging, tentative, assertive language).

Logical Fallacies: Logical fallacies often mislead readers, as both writer and reader fall prey to poor but tempting arguments. Indeed, studies have shown that people find them more convincing than is rational [12]. We asked our annotators to look for the *straw man fallacy* (presenting a counterargument as a more obviously wrong version of existing counterarguments), *false dilemma fallacy* (treating an issue as binary when it is not), *slippery slope fallacy* (assuming one small change will lead to a major change), *appeal to fear fallacy* (exaggerating the dangers of a situation), and the *naturalistic fallacy* (assuming that what is natural must be good).

Tone: Readers can be misled by the emotional tone of articles. Such language is common in opinion pieces, which readers may parse as straight news. We asked our annotators to look for exaggerated claims or emotionally charged sections, especially for expressions of contempt, outrage, spite, or disgust.

Inference: *Correlation* and *causation* are often conflated, and the implications can be dramatic, for example in medical trials. There is also the more subtle conflation between singular causation ("the drunk driver caused *that* accident") and general causation ("drinking and driving causes accidents"). For this indicator, we asked annotators to determine what type of causality—correlation,

singular causation, or general causation—was at play, and whether there is convincing evidence for the claims expressed.

3.2 Context Indicators

In total, there were 8 context indicators collected by annotators. Context indicators require annotators to look outside of the article text and research external sources or examine the metadata surrounding the article text, such as advertising and layout.

Originality: Republishing text is a common practice in online news. Reasons include licensing agreements from a wire service such as Reuters, or the article can simply be stolen or reworded without attribution. We asked annotators to find whether the article was an original piece of writing or duplicated elsewhere, and if so, to check whether attribution was given.

Fact-checked: We asked annotators to determine whether the central claim of the article, if one exists, was fact-checked by an approved organization, as well as the outcome of the check. While many organizations conduct fact-checking [15], we limited our consideration to organizations vetted by a verified signatory of Poynter's International Fact-Checking Network (IFCN)[8]. Because many IFCN members utilize schema.org's ClaimReview schema, there is potential to automate this process in the future [9].

Representative Citations: Journalists are expected to accurately represent any sources that they cite or quote, such as articles, interviews, or other external materials. As an article may have many sources, we asked annotators to check the representation of only the first three sources mentioned in the article. Annotators were asked to find the original content and rate how accurately the description in the article represented the original content.

Reputation of Citations: Without domain experts, it is difficult to systematically evaluate the validity or credibility of a cited source. However, for scientific studies, there are at least some existing public measures such as impact factor, despite their documented issues [44]. Thus, we asked annotators to find the impact factor of the publication of any scientific study cited.

Number of Ads: Most publications depend on ad content and recommendation engines as a core part of their business model. Per a recent Facebook strategy, a very high number of ads relative to content may be an indicator of a financially-motivated misinformation site [1]. We asked annotators to count the number of display ads, content recommendation engines, such as Taboola or Outbrain, as well as recommended sponsored content.

Number of Social Calls: Most publications depend on social networks and viral content to drive traffic to their site. That said, a high number of exhortations to share content on social media, email the article, or join a mailing list can be an indicator of financially-motivated misinformation. We ask annotators to count the number of calls to share on social media, email, or join a mailing list.

"Spammy" Ads: As well as quantity, the ads on the page may be of a "spammy" nature, such as containing disturbing or titillating imagery, or celebrities, or clickbait titles. Thus, we asked annotators to rate the "spammyness" of the ads.

Placement of Ads and Social Calls: Finally, the placement of ads and social calls may be an indicator, for instance by appearing in pop-up windows, covering up article content, or distracting through additional animation and audio. We ask annotators to rate the aggressiveness of the placement of ads and social calls.

[8]International Fact-Checking Network (IFCN): https://www.poynter.org/international-fact-checking-network-fact-checkers-code-principles

4 DATA COLLECTION

This section describes our process for gathering articles, finding annotators, and selecting platforms for credibility annotation.

4.1 Articles

We focused on the topics of climate science and public health, where misinformation is prevalent despite a high degree of stable knowledge and expert consensus. Articles were selected using BuzzSumo[9], a service that surfaces the most shared articles on social media for any search term. Terms we searched included "climate change" and "global warming" for climate science, and "health", "vaccines", and "disease" for public health. Articles returned from the year 2017 were collected into one list and sorted by most overall shares, so as to prioritize high impact articles with broad appeal. We removed 2 articles that were too long for annotation (3,500+ words), 2 that were primarily images, and one suspended article. Finally, the 40 most shared articles from the list were selected. In total, there were 22 articles related to public health, 10 related to climate science, 7 related to diseases, and 4 related to vaccines. The most shared article was about vaccines by a publisher called "Earth. We Are One" and shared 1.9 million times in 2017, according to BuzzSumo. To ensure that article content would not change or disappear during the study, they were archived using Archive.is[10].

4.2 Annotators

Six annotators were recruited for this task, with 3 focused on content indicators and 3 marking context indicators, as content annotation requires different prior knowledge and training than context. The 3 content annotators were recruited from the teaching staff of a UC Berkeley course on scientific-style critical thinking called Sense and Sensibility and Science (SSS)[11] The content annotators were selected because of their exemplary performance in the course and their skills in scientific critical thinking. The 3 context annotators were recruited by Meedan[12] from a number of journalism schools. Annotators were either journalism students or recent graduates. Annotators were paid $150 for the entire task, or $3.75 per article. For a $15 wage per hour, this amounts to around 15 minutes spent per article, which we sought to target when devising annotations.

The average age of the annotators was 22.1, and 5 annotators were female, while 1 was male. We sought to diversify our population in terms of political orientation to mitigate issues with bias. Asked about their political affiliation, 3 stated Democrat, 1 Republican, 1 Independent, and one stated none. On economic issues, 2 named themselves as very liberal, 1 moderately liberal, 1 moderate, 1 moderately conservative, and 1 as very conservative. On social issues, 4 considered themselves very liberal while 2 considered themselves moderate. When asked what publications they read regularly, 4 annotators mentioned The New York Times, while 2

[9]BuzzSumo: http://buzzsumo.com
[10]Webpage archiving tool: http://archive.is
[11]SSS course at Berkeley: http://sensesensibilityscience.com
[12]Meedan: https://meedan.com

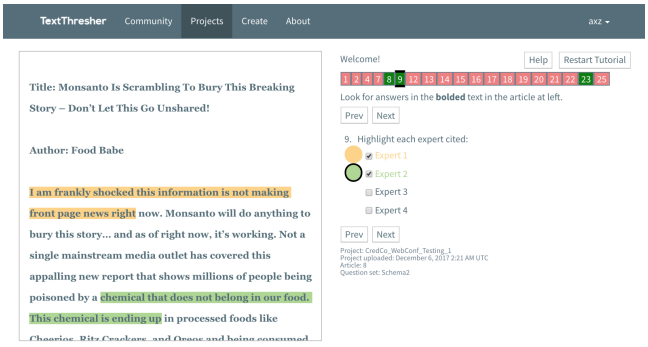

Figure 1: Screenshot of TextThresher platform used for content indicator annotation.

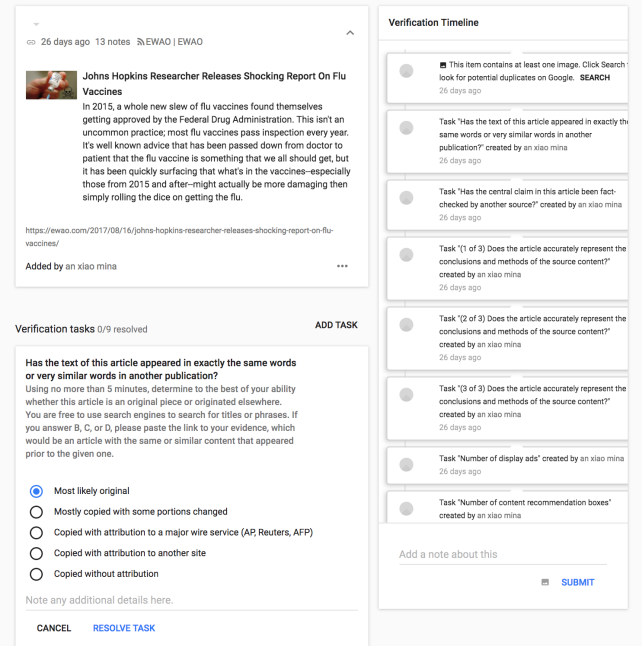

Figure 2: Screenshot of Check platform used for context indicator annotation.

mentioned CNN. The remaining 22 publications were mentioned only once. Our future work will aim towards greater diversity among annotators as we grow our pool.

Finally, we collected a credibility score for each article from domain experts to serve as "gold standard" credibility scores. We determined the general topic of each article, such as climate science, psychology, or public health. Then we reached out on different platforms to find domain experts in those areas, such as scientists or industry practitioners, to score the article on a 5 point scale and leave notes. For articles dedicated to breaking news events, we had a journalist score them for credibility. Overall there were 5 domain expert annotators who volunteered their time.

4.3 Annotation Platforms

The three annotators tasked with content indicators used a collaborative annotation software called TextThresher[13], which is also in use by the citizen-science misinformation and media literacy platform PublicEditor designed at the UC Berkeley Institute for Data Science [36]. Figure 1 shows the tool guiding contributors to answer a series of questions about article text, highlighting the portions of text that justify their answers.

Because TextThresher supports the annotation of plain text files, the tool was useful for our approach to content indicator evaluation, which seeks to reduce annotator bias by removing text from its original context. Previous workshops we conducted revealed that participants' assessments were strongly influenced by the name of the publication and its layout. With TextThresher, we only show annotators the article's title, any image captions, and the main text of each article. TextThresher also guides users to label specific words and phrases within articles, which enables the precise identification of specific phenomena. Users' labels can then be displayed to news readers to improve their media literacy, and in high-traction supervised machine learning scenarios.

The three annotators tasked with context indicators used a tool called Check[14] built by Meedan, a nonprofit software company that builds digital tools for journalists. Using Check, we showed a preview of the article with a link to see the article in context. Because these indicators involved looking at the information around the article as well as conducting research on external information, it was no longer possible to obfuscate the publisher or other information. Each annotator could see all their own annotation tasks related to an article on the page and mark each as complete when done. They could also keep track of their progress and go back to articles to edit or resume annotation.

[13]For details on the TextThresher software: http://www.goodlylabs.org/research/
[14]Check: https://meedan.com/en/check/

5 DATASET ANALYSIS

We next turn to analysis of the annotation data. Here, we focus on two measures: (1) how much annotators agreed with one another when identifying indicators, and (2) how much the annotators' assessments of overall article credibility agreed with domain experts' assessments. We calculate inter-rater reliability (IRR) using Krippendorff's alpha, as it can be used for more than 2 scorers and can be adapted to many different data types, including nominal, ordinal, and interval scores, all of which are present in the data we collected.

When we aggregate annotations to then correlate with domain expert scores, in the case of ordinal and interval data, we compute an average across annotators, while in the case of nominal data, we use the category most chosen, if it exists. To determine correlation, in the case of ordinal and interval indicator data, we use the Spearman rank correlation as it allows for ordinal data, and relationships need not be linear. For nominal input data, as there is no concept of correlation, we convert the categories into binary variables and perform a multiple linear regression. We report the coefficient of determination (R^2), which reports what percentage of the variance in domain expert scores is explained by the model.

Content Indicator	Data Type	IRR	Relation to Experts
Title Representativeness	ordinal	0.367	ρ=0.234
"Clickbait" Title	ordinal	0.581	ρ=-0.709***
Quotes from Outside Experts	interval	0.673	ρ=0.327
Citation of Organizations	interval	0.283	ρ=0.145
Citation of Studies	interval	0.763	ρ=0.107
Single Study Article	nominal	0.877	R^2=0.031
Confidence - Extent Claims Justified	ordinal	-0.093	ρ=0.690***
Confidence - Acknowledge Uncertainty	ordinal	0.534	ρ=-0.247
Logical Fallacies - Straw Man	ordinal	-0.096	ρ=-0.402*
Logical Fallacies - False Dilemma	ordinal	0.102	ρ=-0.303
Logical Fallacies - Slippery Slope	ordinal	0.478	ρ=0.374*
Logical Fallacies - Appeal to Fear	ordinal	0.314	ρ=-0.424*
Logical Fallacies - Naturalistic	ordinal	0.377	ρ=-0.533**
Tone - Emotionally Charged	ordinal	0.098	ρ=0.611***
Tone - Exaggerated Claims	ordinal	0.235	ρ=0.606***
Inference - Type of Claims	nominal	0.154	R^2=0.029
Inference - Convincing Evidence	ordinal	0.540	ρ=0.764***

Table 1: Inter-rater reliability for content indicators and their relationship to expert scores of credibility. (*p<0.05, **p<0.01, ***p<0.001)

Context Indicator	Data Type	IRR	Relation to Experts
Originality	nominal	0.346	R^2=0.068
Fact-checked	nominal	0.303	R^2=0.309*
Representative Citations	ordinal	0.312	ρ=0.612***
Reputation of Citations	interval	0.852	ρ=-0.026
Number of Ads	interval	0.535	ρ=-0.135
Number of Content Recommendation	interval	-0.088	ρ=0.144
Number of Sponsored Content	interval	0.422	ρ=-0.196
Number of Social Calls	interval	0.564	ρ=0.179
Number of Mailing List or Email Calls	interval	0.375	ρ=0.453**
"Spammy" Ads	ordinal	0.554	ρ=-0.309
Placement of Ads and Social Calls	ordinal	0.326	ρ=-0.417*

Table 2: Inter-rater reliability for context indicators and their relationship to expert scores of credibility. (*p<0.05, **p<0.01, ***p<0.001)

5.1 Content Indicators

In Table 1, we show both the IRR between the three annotators as well as the relationship to domain expert scores for each of the questions for content indicators. Some indicators had moderate to strong IRR, such as the ones involving highlighting number of citations, quotes, or scientific studies. Other indicators with moderate reliability included clickbait, which also had high correlation to expert ratings. The correlation with expert ratings combined with high IRR suggests that these types of indicators may be a useful signal of credibility to target for further annotation.

Annotators had weaker agreement on questions related to logical fallacies generally due to scarcity and inadequate training. While at first glance, "false dillemma" showed up in 16.1% of articles and "straw man" applied to 37.7%, further analysis reveals that all these annotations were due to a single annotator. As the annotators were not given explicit definitions of the fallacies, differences in interpretation could lead to low IRR. Future annotation of logical fallacy indicators could include more training and examples.

Indicators referencing claims (Confidence–extent claims justified; Scientific Inference–types of claims) also had low IRR. Some prior evidence suggests this was expected: of the 200+ students taking the Sense & Sensibility & Science course at UC Berkeley, less than half answered questions about the type of scientific inference in a claim correctly. However, annotators for this study had high IRR and high correlation for follow up questions (Scientific Inference–convincing evidence, etc.) This suggests that non-expert annotators may find it difficult to classify claims, but once a claim is classified, annotators can be used for further evaluation.

We notice some indicators that had a moderate to strong correlation with domain experts, such as the perception of convincing evidence, that would be expected. Other indicators had a more unexpected relationship to expert credibility. For instance, the presence of slippery slope logical fallacies actually had a weak positive correlation with expert perception of credibility. Interestingly, some of the indicators such as "Tone–emotionally charged" that had low agreement between annotators still had strong correlation with experts, demonstrating that individual scores may have been calibrated to different levels but still moved similarly. However, normalizing scores for each annotator did not significantly alter IRR. We also found that there were several content indicators that were auto-correlated. Some were highly correlated within an indicator, such as the two questions related to Tone (ρ=0.736, p <0.001), suggesting that the number of questions could be reduced or that annotators could have been biased in one direction across questions. Future analysis, perhaps comparing with a gold standard set of annotations, is necessary. From discussions with annotators, some annotators did in fact say that some questions felt redundant. Several annotators also remarked that as they annotated more and more articles, their initial read of the article (before answering any of the questions) was already punctuated with a mental checklist.

5.2 Context Indicators

In Table 2, we show the IRR and correlation with domain experts for context indicators. Most indicators showed moderate to strong agreement between annotators. One major exception is the indicator asking annotators to count content recommendation boxes

Credibility Rating	IRR	Avg (SD)	Relation to Experts
Content Pre-Annotation	0.695	2.61 (0.98)	ρ=0.630***
Content Post-Annotation	0.665	2.60 (0.97)	ρ=0.748***
Context Pre-Annotation	0.715	2.81 (1.18)	ρ=0.783***
Context Post-Annotation	0.616	2.70 (1.16)	ρ=0.793***
Domain Experts	-	2.29 (1.38)	-

Table 3: Inter-rater reliability for assessments of credibility by annotators before doing annotation as well as after, as well as their relationship to expert scores of credibility. (*p<0.05, **p<0.01, ***p<0.001)

had no agreement between annotators. We found this was because some annotators counted every content recommendation article shown, while others counted an entire box of articles as a single entity. While there was weak agreement between annotators on the question of proper characterization of sources, this may be lower partially because we noticed annotators did not always choose to annotate the same three sources due to disagreement on what constitutes a source. However, the question about source characterization had the strongest correlation with expert credibility.

When examining the advertising and social sharing indicators, it was interesting to note that counting the quantity of the different types of supplementary content was not significant except for the case of mailing lists and email. This suggests that both low and high credibility publications may be using similar monetization techniques. Likewise, if they are using similar advertising networks, they may both be serving up similarly "spammy" ads, as echoed by advertising industry experts [28]. One difference however is in the aggressiveness of ad placement. Future work could consider signals determined by standards set by the Coalition for Better Ads[15].

Finally, we notice a lack of correlation between the impact factor of scientific citations and credibility. In total, only 25% of articles had any annotations of impact factor. As there is no single structured database for journal impact factors, annotators went through a manual process of searching. In the future, a structured database of impact factors and other publication quality signals, such as number of citations, could make machine assessments easier.

5.3 Comparing Credibility Scores

We asked annotators to mark their overall impression of credibility of the article on a 5-point scale both before and after each article annotation. As seen in Table 3, for both sets of annotators, the IRR dropped from before conducting annotation to after. While there was also slight differences between before and after mean scores, these were not significant. We notice that the correlation to domain expert scores increases from before annotation to after. This suggests that annotators became more aligned with domain experts after investigating our indicators.

Finally, we also notice differences between the different sets of annotators and the domain experts. Overall, context annotators had a stronger correlation to experts than content annotators. These differences may partially be due to the information to which the annotators had access. In the case of content, annotators did not have information about the source or the presentation of the article webpage, which can be strong indicators of credibility. As both context annotators and domain experts had access to this context, their scores may be more aligned. However, context annotators rated articles significantly higher than experts both before and after annotation (paired t-test, p <0.05), while this difference was not significant between content annotators and experts.

[15]Coalition for Better Ads: https://www.betterads.org/standards/

5.4 Predicting Credibility from Indicators

To understand the predictive value provided by both sets of indicators, two backward stepwise multiple regression models were ran to regress domain expert scores onto each set of indicators. As some indicators had high auto-correlation, we removed variables within the same indicator that were highly correlated with each other, leaving a single variable to represent that indicator. For the content model, R^2 for the overall model was 0.482 and adjusted R^2 was 0.448. After model convergence, two variables remained: Clickbait Title and Logical Fallacies–slippery slope. This model was found to significantly predict credibility ($F = 13.972, p < 0.001$). For the context model, R^2 for the overall model was 0.750 and adjusted R^2 was 0.692. After model convergence, 6 variables remained: Fact-checked–reported false, Fact-checked–reported mixed results, Number of Social Calls, Number of Mailing List Calls, and Placement of Ads and Social Calls. Together, they were also found to significantly predict credibility ($F = 12.986, p < 0.001$). The context regression model overall had a better fit than the content model which, along with issues of collinearity among the content indicators, lead us to believe more work is needed to differentiate the phenomena within specific indicators in the content category.

6 DISCUSSION AND FUTURE WORK

In this work, we outline a process for defining indicators of credibility and validating them through the collection of annotation data. The ability to quickly test new indicator definitions for both reliable annotation and correlation with expert-defined credibility will be important as we continue to scale to more indicators, more articles and other content, and more diverse annotators.

From going through this process with a focused set of indicators and articles, we found several indicators that show reliability and correlation with domain expert scores of credibility, such as the presence of a clickbait title or the accurate representation of sources cited in the article. We also obtained findings that suggest certain indicators are less useful, such as the presence or spammy nature of advertising and social calls. Finally, we received feedback on the importance of indicators towards modeling expert credibility, which helps determine indicators that may be redundant or more or less predictive. This initial foray additionally allowed us to examine the distinction between content versus context indicators and the training and annotation interfaces required for accurate assessment of each. We found areas that may be out of reach for non-expert annotators, such as inferring types of claims, or that require more training or technical tools for lateral searching, such as assessing the reputation of citations. Looking forward, one can imagine different annotation strategies for different indicators based on these findings, with some fully or partially automatically captured, some annotated

by experts or publishers, and some surfaced by the crowd or one's immediate trust network.

In terms of immediate future steps, we aim to scale up our work to 5,000 to 10,000 annotated articles across a range of topics, styles, and publications, and work with researchers and web platform representatives to put this data to use towards building models of credibility that are both interpretable and robust to manipulation. In order to ensure the sustainability and inclusivity of our work as we continue to define credibility standards, we have formed the W3C Credible Web Community Group[16], first introduced as a session at W3C's Technical Plenary/Advisory Committee meeting[17] in 2017.

Our efforts are also aimed at improving media literacy and shrinking gaps of understanding between domain expertise and public knowledge. Through our work and outreach, we aim to convey how credibility is a negotiation among communicants [25], where publishers and authors seek to convey credibility while readers and platforms seek to ascertain it. Greater, richer communication, understanding of dependencies between communicants, and tools to improve the transfer of information are necessary towards reducing the spread of misinformation. Along these lines, our work raises more long-term research questions that we aim to explore.

Indicator Resilience. Analogous to anti-spam efforts, the usefulness of automated credibility assessments may vary dramatically depending on the motivation and resources of the misinformation propagators. A nation-state actor with a geopolitical strategy may be harder to dissuade than financially motivated "fake news" creators. On the other hand, some indicators, such as raw number of ads or fact-checks from IFCN-verified signatories, may be more difficult to manipulate. Ultimately, we believe that the ability to compare the resilience of indicators is important in the context of increasingly machine-driven information landscapes.

Journalistic Practice. Also key is that the indicators are not just a tool for detecting misinformation but also the quality of information itself. For instance, recent efforts by Facebook to limit content with a high degree of clickbait suggests simple ways to improve quality [46]. Annotators looking at content indicators found logical fallacies and incorrect use of causal claims even among some highly reputable news sources. We believe there is potential for the indicators to help improve standards for mainstream journalism, whether through custom tools or as a training methodology.

Media Literacy. Recent work from the Pew Research Center shows that more than half of adults think that "training in the digital realm would help them when it comes to accessing information that can aid in making decisions" [18]. In our study, we found that annotators changed how they approached new articles as the process went on, and we also saw changes in their credibility scores after annotation that aligned better with experts. Indeed, many of our context indicators are designed to map to existing processes for fact checking, such as reading laterally [51] and going "upstream" to find the source of a claim [7]. Likewise, the ability to employ critical thinking or pick up on misleading language allows readers to reject misinformation that takes advantage of psychological biases. Going forward, we aim to develop a set of training materials so that anyone can get involved in annotation. We also will display all collected annotations on our website using Hypothes.is[18], a web annotation platform, for the public to be able to inspect the annotations in our dataset in context of the articles.

Freedom of Expression. How can attempts to detect and curb misinformation online meaningfully differ from efforts to censor the internet? The weaponization of "fake news" by autocratic countries already demonstrates the risks here: in a context where political leaders aim to centralize their control over the truth, determining blanket falsehood becomes a strategy of state control. In this regard, transparency presents a double-edged sword. As described earlier, it creates incentives for innovations in manipulation by agents of disinformation. At the same, transparency helps reveal how annotators arrived at their conclusions about these indicators. We seek to study how greater transparency in indicators, by enabling the ability to share clear processes and findings, can help strike a balance between improving the health of our information ecosystem while preserving basic principles of free speech.

7 LIMITATIONS

There are limits to the potential effects of these indicators, and understanding their applicability is important. Even if we are successful in curbing some of the psychological foundations of misinformation, such as frequency of exposure, more work is needed to fully address the many social and identity-related motivations for believing misinformation. These indicators were developed in the US and UK contexts and may not be applicable to other languages and parts of the world. As mentioned earlier, we focus on articles for this work but aim to expand to images, video, and other digital multimedia in the future. Additionally, digital initiatives will need to also consider the wider information ecosystem that includes television and talk radio.

8 CONCLUSION

In this work, we presented a set of 16 indicators of article credibility, focused on article content as well as external sources and article metadata, refined over several months by a diverse coalition of media experts. We also presented a process for gathering annotations of these credibility indicators, including platform design and annotator recruitment, as well as an initial dataset of 40 articles annotated by 6 trained annotators and scored by domain experts. From analyzing our data, we isolated indicators that are reliably annotated across articles and that correlate with domain experts. Finally, we described the broader initiative of creating a set of standards around content credibility, of which this project is a part, as well as future directions for research.

9 ACKNOWLEDGEMENTS

This paper would not be possible without the valuable support and feedback of members of the Credibility Coalition, who have joined in-person meetings, weekly calls, and daily Slack chats to generously contribute their time, effort, and thinking to this project. There are too many to thank in the space we have, and we have included acknowledgments at www.credibilitycoalition.org.

[16] https://www.w3.org/community/credibility/
[17] W3C TPAC: https://www.w3.org/2017/11/TPAC/
[18] Hypothes.is: https://web.hypothes.is/

REFERENCES

[1] Jason Abbruzzese. 2017. Facebook is going to do something about those terrible ads on your website. (May 2017). http://mashable.com/2017/05/10/facebook-crackdown-bad-ads-news-feed/

[2] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 155–168.

[3] Ahmer Arif, Kelley Shanahan, Fang-Ju Chou, Yoanna Dosouto, Kate Starbird, and Emma S Spiro. 2016. How information snowballs: Exploring the role of exposure in online rumor propagation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 466–477.

[4] Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65, 4 (2015), 619–638.

[5] David B. Buller and Judee K. Burgoon. 1996. Interpersonal Deception Theory. *Communication Theory* 6, 3 (1996), 203–242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 675–684.

[7] Michael A. Caulfield. 2017. Go Upstream to the Find the Source. (Jan 2017). https://webliteracy.pressbooks.com/chapter/go-upstream-to-find-the-source/

[8] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 15–19.

[9] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational Fact Checking from Knowledge Networks. *PLOS ONE* 10, 6 (06 2015), 1–13. https://doi.org/10.1371/journal.pone.0128193

[10] Andrew J Flanagin and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 515–540.

[11] William B. Frakes. 1986. Information and misinformation: An investigation of the notions of information, misinformation, informing, and misinforming. *Journal of the American Society for Information Science* 37, 1 (1986), 48–49. https://doi.org/10.1002/(SICI)1097-4571(198601)37:1<48::AID-ASI10>3.0.CO;2-3

[12] William K. Frankena. 1939. The naturalistic fallacy. *Mind* 48, 192, Article 4 (1939), 13 pages. https://doi.org/10.1093/mind/XLVIII.192.464

[13] Daniel Funke. 2017. It's been a year since Facebook partnered with fact-checkers. How's it going? (Dec. 2017). Retrieved January 5, 2018 from https://www.poynter.org/news/its-been-year-facebook-partnered-fact-checkers-hows-it-going

[14] Cecilie Gaziano and Kristin McGrath. 1986. Measuring the concept of credibility. *Journalism quarterly* 63, 3 (1986), 451–462.

[15] lucas graves and Tom glaisyer. 2012. *The Fact-Checking Universe in Spring 2012: An Overview*. The New America Foundation, Washington, DC, USA. https://www.issuelab.org/resource/the-fact-checking-universe-in-spring-2012-an-overview.html

[16] Jennifer D. Greer. 2003. Evaluating the Credibility of Online Information: A Test of Source and Advertising Influence. *Mass Communication and Society* 6, 1 (2003), 11–28. https://doi.org/10.1207/S15327825MCS0601$_3$

[17] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. SIAM, SIAM, 3600 Market Street, 6th Floor | Philadelphia, PA 19104-2688 USA, 153–164.

[18] John B. Horrigan and John Gramlich. 2017. Many Americans, especially blacks and Hispanics, are hungry for help as they sort through information. (Nov 2017). http://www.pewresearch.org/fact-tank/2017/11/29/many-americans-especially-blacks-and-hispanics-are-hungry-for-help-as-they-sort-through-information/

[19] IREX.org. 2017. Ukrainians' self-defense against disinformation: What we learned from Learn to Discern. (June 2017). Retrieved January 5, 2018 from https://www.irex.org/insight/ukrainians-self-defense-against-disinformation-what-we-learned-learn-discern

[20] Alice Marwick and Rebecca Lewis. 2017. *Media Manipulation and Disinformation Online*. Report. Data & Society Research Institute. https://edoc.coe.int/en/media-freedom/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

[21] Aaron M. McCright and Riley E. Dunlap. 2011. THE POLITICIZATION OF CLIMATE CHANGE AND POLARIZATION IN THE AMERICAN PUBLIC'S VIEWS OF GLOBAL WARMING, 2001âĂŞ2010. *Sociological Quarterly* 52, 2 (2011), 155–194. https://doi.org/10.1111/j.1533-8525.2011.01198.x

[22] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. Twitter Under Crisis: Can we trust what we RT?. In *Proceedings of the first workshop on social media analytics*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 71–79.

[23] Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. 2015. Using twittertrails.com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 69–72.

[24] Miriam J. Metzger, Andrew J. Flanagin, Keren Eyal, Daisy R. Lemus, and Robert M. Mccann. 2003. Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. *Annals of the International Communication Association* 27, 1 (2003), 293–335. https://doi.org/10.1080/23808985.2003.11679029 arXiv:https://doi.org/10.1080/23808985.2003.11679029

[25] Hans K Meyer, Doreen Marchionni, and Esther Thorson. 2010. The journalist behind the news: credibility of straight, collaborative, opinionated, and blogged âĂIJnewsâĂİ. *American Behavioral Scientist* 54, 2 (2010), 100–119.

[26] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *Ninth International AAAI Conference on Web and Social Media*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., Article 10582, 10 pages.

[27] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 441–450.

[28] Lucia Moses. 2016. 'The underbelly of the internet': How content ad networks fund fake news. (Nov. 2016). Retrieved January 5, 2018 from https://digiday.com/media/underbelly-internet-fake-news-gets-funded/

[29] Ryosuke Nagura, Yohei Seki, Noriko Kando, and Masaki Aono. 2006. A Method of Rating the Credibility of News Documents on the Web. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, Article 1148316, 2 pages. https://doi.org/10.1145/1148170.1148316

[30] Daniel J O'keefe. 2002. *Persuasion: Theory and research*. Vol. 2. Sage, Los Angeles, CA.

[31] Will Oremus. 2016. Only You Can Stop the Spread of Fake News. http://www.slate.com. (December 2016).

[32] Gordon Pennycook and David G Rand. 2017. *Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy*. Technical Report. SSRN. http://dx.doi.org/10.2139/ssrn.3035384

[33] Peter Pirolli, Evelin Wollny, and Bongwon Suh. 2009. So you know you're getting the best possible information: a tool that increases Wikipedia credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 1505–1508.

[34] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 2173–2178.

[35] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European Conference on Information Retrieval*. Springer, Springer International Publishing, Gewerbestrasse 11, 6330 Cham, Switzerland, 810–817.

[36] Aditya Ranganathan, Daniel Kim, Nick Adams, and Saul Perlmutter et al. 2017. *Crowdsourcing Credibility: A Citizen-Science Approach to NewsLiteracy via Public Editor*. Technical Report. University of Berkeley. https://northwestern.app.box.com/s/77ekftnfp0w8ixxkivkgodqubwhaumyv

[37] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. 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*. ACL, 209 N. Eighth Street, Stroudsburg PA 18360, USA, 2921–2927.

[38] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. 2011. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 249–252.

[39] Marta Reacasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *51st Annual Meeting of the Association for Computational Linguistics*. ACL, ACL, 209 N. Eighth Street, Stroudsburg PA 18360, USA, 1650–1659.

[40] Paul Resnick, Samuel Carton, Souneil Park, Yuncheng Shen, and Nicole Zeffer. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Proc. Computational Journalism Conference*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 5.

[41] Soo Young Rieh and David R. Danielson. 2007. Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology* 41, 1 (2007),

307–364. https://doi.org/10.1002/aris.2007.1440410114

[42] Christine Schmidt. 2017. This project aims to "de-flatten" digital publishing by matching the best content with premium ads. (Nov 2017). http://www.niemanlab.org/2017/11/this-project-aims-to-de-flatten-digital-publishing-by-matching-the-best-content-with-premium-ads/

[43] Mike Schmierbach and Anne Oeldorf-Hirsch. 2012. A little bird told me, so I didn't believe it: Twitter, credibility, and issue perceptions. *Communication Quarterly* 60, 3 (2012), 317–337.

[44] Per O Seglen. 1997. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal* 314, 7079 (1997), 498.

[45] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19, 1, Article 3137600 (Sept. 2017), 15 pages. https://doi.org/10.1145/3137597.3137600

[46] Henry Silverman and Lin Huang. 2017. News Feed FYI: Fighting Engagement Bait on Facebook. (Dec 2017). https://newsroom.fb.com/news/2017/12/news-feed-fyi-fighting-engagement-bait-on-facebook/

[47] Kate Starbird. 2017. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter.. In *ICWSM*. ACM, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701., 230–239.

[48] S Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly* 75, 1 (1998), 55–68.

[49] Lauren Vogel. 2017. Viral misinformation threatens public health. *Canadian Medical Association Journal* 189, 50, Article E1567 (Dec 2017), 1 pages. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5738254/

[50] Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policy making.* Coucil of Europe Report DGI(2017)09. Council of Europe. https://edoc.coe.int/en/media-freedom/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

[51] Sam Wineburg and Sarah McGrew. 2017. *Lateral Reading: Reading Less and Learning More When Evaluating Digital Information.* Technical Report Working Paper No. 2017-A1. Stanford History Education Group. http://dx.doi.org/10.2139/ssrn.3048994

[52] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2017. Computational Fact Checking through Query Perturbations. *ACM Transactions on Database Systems (TODS)* 42, 1 (2017), 4.

[53] Kenneth C.C. Yang. 2007. Factors influencing Internet usersâĂŹ perceived credibility of news-related blogs in Taiwan. *Telematics and Informatics* 24, 2 (2007), 69 – 85. https://doi.org/10.1016/j.tele.2006.04.001

[54] Wenlin Yao, Zeyu Dai, Ruihong Huang, and James Caverlee. 2017. Online Deception Detection Refueled by Real World Data Collection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., Varna, Bulgaria, 793–802. https://doi.org/10.26615/978-954-452-049-6$_1$02