SciLens: Evaluating the Quality of Scientific News Articles Using Social Media and Scientific Literature Indicators

Panayiotis Smeros
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
panayiotis.smeros@epfl.ch

Carlos Castillo
Universitat Pompeu Fabra (UPF)
Barcelona, Catalunya, Spain
carlos.castillo@upf.edu

Karl Aberer
École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
karl.aberer@epfl.ch

ABSTRACT

This paper describes, develops, and validates SciLens, a method to evaluate the quality of scientific news articles. The starting point for our work are structured methodologies that define a series of quality aspects for manually evaluating news. Based on these aspects, we describe a series of indicators of news quality. According to our experiments, these indicators help non-experts evaluate more accurately the quality of a scientific news article, compared to non-experts that do not have access to these indicators. Furthermore, SciLens can also be used to produce a completely automated quality score for an article, which agrees more with expert evaluators than manual evaluations done by non-experts. One of the main elements of SciLens is the focus on both content and context of articles, where context is provided by (1) explicit and implicit references on the article to scientific literature, and (2) reactions in social media referencing the article. We show that both contextual elements can be valuable sources of information for determining article quality. The validation of SciLens, done through a combination of expert and non-expert annotation, demonstrates its effectiveness for both semi-automatic and automatic quality evaluation of scientific news.

KEYWORDS

ACM Reference Format:
Panayiotis Smeros, Carlos Castillo, and Karl Aberer. 2019. SciLens: Evaluating the Quality of Scientific News Articles Using Social Media and Scientific Literature Indicators. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313657

1 INTRODUCTION

Scientific literacy is broadly defined as a knowledge of basic scientific facts and methods. Deficits in scientific literacy are endemic in many societies, which is why understanding, measuring, and furthering the public understanding of science is important to many scientists [6].

Mass media can be a potential ally in fighting scientific illiteracy. Reading scientific content has been shown to help align public knowledge of scientific topics with the scientific consensus, although in highly politicized topics it can also reinforce pre-existing biases [27]. There are many ways in which mass media approaches science, and even within the journalistic practice there are several sub-genres. Scientific news portals, for instance, include most of the categories of articles appearing traditionally in newspapers [21] such as *editorial*, *op-ed*, and (less frequently) *letters to the editor*. The main category of articles, however, are scientific *news* articles, where journalists describe scientific advances.

Scientific news articles have many common characteristics with other classes of news articles; for instance, they follow the well-known *inverted pyramid* style, where the most relevant elements are presented at the beginning of the text. However, they also differ in important ways. Scientific news are often based on findings reported in scientific journals, books, and talks, which are highly specialized. The task of the journalist is then to *translate* these findings to make them understandable to a non-specialized, broad audience. By necessity, this involves negotiating several trade-offs between desirable goals that sometimes enter into conflict, including appealing to the public and using accessible language, while at the same time accurately representing research findings, methods, and limitations [46].

The resulting portrayal of science in news varies widely in quality. For example, the website "Kill or Cure?"[1] has reviewed over 1,200 news stories published by The Daily Mail (a UK-based tabloid) finding headlines pointing to 140 substances or factors that cause cancer (including obesity, but also Worcestershire sauce), 113 that prevent it (including garlic and green tea), and 56 that both cause and prevent cancer (including rice). Evidently, news coverage of cancer research that merely seeks to classify every inanimate object into something that either causes or prevents cancer does not help to communicate effectively scientific knowledge on this subject.

Our contribution. In this paper we describe *SciLens,* a method for evaluating the quality of scientific news articles. The technical contributions we describe are the following:

- a framework, depicted in Figure 1, for semi-automatic and automatic article quality evaluation (§3);
- a method for contextual data collection that captures the contents of an article, its relationship with the scientific literature, and the reactions it generates in social media (§4);
- a series of automatically-computed *quality indicators* describing:
  - the content of a news article, where we introduce a method to use quotes appearing on it as quality indicators (§5.1),
  - the relationship of a news article with the scientific literature, where we introduce content-based and graph-based similarity methods (§5.2), and

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313657

[1] http://kill-or-cure.herokuapp.com

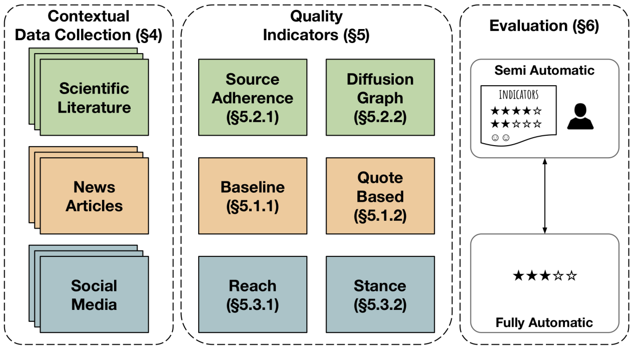


Figure 1: Overview of SciLens, including contextual data collection, quality indicators, and evaluation.

  - the social media reactions to the article, where we introduce a method to interpret their stance (supporting, commenting, contradicting, or questioning) as quality signals (§5.3);
- an experimental evaluation of our methods involving experts and non-experts (§6).

2 RELATED WORK

In this section, we present background information that frames our research (§2.1), previous work on evaluating news quality (§2.2), and methods to extract quality indicators from news articles (§2.3). This is a broad research area where results are scattered through multiple disciplines and venues; our coverage is by no means complete.

2.1 Background on Scientific News

A starting point for understanding communication of science has historically been the "deficit model," in which the public is assumed to have a deficit in scientific information that is addressed by science communication (see, e.g., Gross [25]). In a simplified manner, scientific journalism, as practiced by professional journalists as well as science communicators and bloggers from various backgrounds, can be seen as a translation from a discourse inside scientific institutions to a discourse outside them. However, there are many nuances that make this process much more than a simple translation. For instance, Myers [44], among others, notes that (i) in many cases the gulf between experts and the public is not as large as it may seem, as many people may have some information on the topic; (ii) there is a continuum of popularization through different genres, i.e., science popularization is a matter of degree; and (iii) the scientific discourse is intertwined with other discourses, including the discussion of political and economic issues.

Producing a high-quality article presenting scientific findings to the general public is unquestionably a challenging task, and often there is much to criticize about the outcome. In the process of writing an article, "information not only changes textual form, but is simplified, distorted, hyped up, and dumbed down" [44]. Misrepresentation of scientific knowledge by journalists has been attributed to several factors, including "a tendency to sensationalize news, a lack of analysis and perspective when handling scientific issues, excessive reliance on certain professional journals for the selection of news, lack of criticism of powerful sources, and lack of criteria for evaluating information" [13].

In many cases, these issues can be traced to journalists adhering to journalistic rather than scientific norms. According to Dunwoody [15], this includes (i) a tendency to favor conflict, novelty, and similar news values; (ii) a compromise of accuracy by lack of details that might be relevant to scientists, but that journalists consider uninteresting and/or hard to understand for the public; and (iii) a pursuit of "balance" that mistakenly gives similar coverage to consensus viewpoints and fringe theories. Journalists tend to focus on events or episodic developments rather than long-term processes, which results in preferential coverage to initial findings even if they are later contradicted, and little coverage if results are disconfirmed or shown to be wrong [14]. Furthermore, news articles typically do not include caveat/hedging/tentative language, i.e., they tend to report scientific findings using a language expressing certainty, which may actually have the opposite effect from what is sought, as tentative language makes scientific reporting more credible to readers [32].

2.2 Evaluation of Quality of News

There are many approaches for evaluating the quality of articles on the web; we summarize some of these approaches in Table 1.

Manual Evaluation. The simplest approach for evaluating news article quality relies on the manual work of domain experts. This is a highly subjective task, given that quality aspects such as credibility are to a large extent perceived qualities, made of many dimensions [20]. In the health domain, evaluations of news article quality have been undertaken for both general health topics [53] and specific health topics such as Pancreatic Cancer [57].

Independent, non-partisan *fact-checking portals* perform manual content verification at large scale, typically employing a mixture of professional and volunteer staff. They can cover news articles on general topics (e.g., Snopes.com) or specific topics such as politics (e.g., PolitiFact.com). In the case of science news, ClimateFeedback.org is maintained by a team of experts on climate change with the explicit goal of helping non-expert readers evaluate the quality of news articles reporting on climate change. Each evaluated article is accompanied by a brief review and an overall quality score. Reviews and credibility scores from fact-checking portals have been recently integrated with search results [36] and social media posts [40] to help people find accurate information. Furthermore, they are frequently used as a ground truth to build systems for rumor tracking [54], claim assessment [50], and fake multimedia detection [8]. Articles considered by fact-checking portals as misinformation have been used as "seeds" for diffusion-based methods studying the spread of misinformation [56].

Our approach differs from previous work because it is completely automated and does not need to be initialized with labels from expert- or crowd-curated knowledge bases. For instance, in the *diffusion graph*, which is the graph we construct during contextual data collection (§4) from social media posts and scientific papers, we do not need prior knowledge on the quality of different nodes.

Automatic and Semi-Automatic Evaluation. Recent work has demonstrated methods to automate the extraction of signals or indicators of article quality. These indicators are either expressed at a

Table 1: Summary of selected references describing techniques for evaluating news article quality.

	Fact-checking portals	Shao et al. [54]	Boididou et al. [8]	Popat et al. [50]	Tambuscio et al. [56]	Ciampaglia et al. [11]	Urban and Schweiger [58]	Zhang et al. [62]	Shu et al. [55]	Kumar et al. [37]	Sbaffi and Rowley [53]	Taylor et al. [57]	**SciLens**
Automatic assessment	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓
No ground-truth needed	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✓	✓
Uses article content	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓
Uses reactions from social media	✗	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓
Uses referenced scientific literature	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓
Domain-agnostic	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗	✗	✓
Web-scale	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✗	✓

conceptual level [58] (e.g, *balance of view points*, *respect of personal rights*) or operationalized as features that can be computed from an article [62] (e.g., *expert quotes* or *citations*). Shu et al. [55] describe an approach for detecting fake news on social media based on social and content indicators. Kumar et al. [37] describe a framework for finding hoax Wikipedia pages mainly based on the author's behavior and social circle, while Ciampaglia et al. [11] use Wikipedia as ground truth for testing the validity of dubious claims. Baly et al. [5] describe site-level indicators that evaluate an entire website instead of individual pages.

Our work differs from these by being, to the best of our knowledge, the first work that analyzes the quality of a news article on the web combining its own content with context that includes social media reactions and referenced scientific literature. We provide a framework, *SciLens*, that is scalable and generally applicable to any technical/scientific context at any granularity (from a broad topic such as "health and nutrition" to more specific topics such as "gene editing techniques").

2.3 Indicator Extraction Techniques

Our method relies on a series of indicators that can be computed automatically, and intersects previous literature that describes related indicators used to evaluate article quality or for other purposes.

Quote Extraction and Attribution. The most basic approach to quote extraction is to consider that a quote is a "block of text within a paragraph falling between quotation marks" [16, 51]. Simple regular expressions for detecting quotes can be constructed [45, 52]. Pavllo et al. [48] leverages the redundancy of popular quotes in large news corpora (e.g., highly controversial statements from politicians that are intensely discussed in the press) for building unsupervised bootstrapping models, while Pareti et al. [47] and Jurafsky et al. [34] train supervised machine learning models using corpora of political and literary quotes (Wikiquote, https://www.wikiquote.org, is such a corpus that contains general quotes).

Our work does not rely on simple regular expressions, such as syntactic patterns combined with quotations marks, which in our preliminary experiments performed poorly in quote extraction from science news; instead we use regular expressions based on classes of words. We also do not use a supervised approach as there is currently no annotated corpus for scientific quote extraction. For the research reported on this paper, we built an information extraction model specifically for scientific quotes from scratch, i.e., a "bootstrapping" model, which is based on word embeddings. This is a commonly used technique for information extraction when there is no training data and we can manually define a few high-precision extraction patterns [33].

Semantic Text Similarity. One of the indicators of quality that we use is the extent to which the content of a news article represents the scientific paper(s) it is reporting about. The Semantic Text Similarity task in Natural Language Processing (NLP) determines the extent to which two pieces of text are semantically equivalent. This is a popular task in the *International Workshop on Semantic Evaluation* (SemEval). Three approaches that are part of many proposed methods over the last few years include: (i) *surface-level similarity* (e.g., similarity between sets or sequences of words or named entities in the two documents); (ii) *context similarity* (e.g., similarity between document representations); and (iii) *topical similarity* [26, 38].

In this paper, we adopt these three types of similarity, which we compute at the document, paragraph, and sentence level. The results we present suggest that combining different similarity metrics at different granularities results in notable improvements over using only one metric or only one granularity.

Social Media Stance Classification. Our analysis of social media postings to obtain quality indicators considers their *stance*, i.e., the way in which posting authors position themselves with respect to the article they are posting about. Stance can be binary ("for" or "against"), or be described by more fine-grained types (supporting, contradicting, questioning, or commenting) [28], which is what we employ in this work. Stance classification of social media postings has been studied mostly in the context of online marketing [35] and political discourse and rumors [63].

In our work, we build a new stance classifier based on textual and contextual features of social media postings and replies, annotated by crowdsourcing workers. To the best of our knowledge, there is no currently available corpus covering the scientific domain. As part of our work, we release such corpus.

3 SCILENS OVERVIEW

The goal of SciLens is to help evaluate the quality of scientific news articles. As Figure 1 shows, we consider a contextual data collection, a computation of quality indicators, and an evaluation of the results.

Contextual Data Collection (§4). First, we consider a set of keywords that are representative of a scientific/technical domain; for this paper, we have considered a number of key words and phrases related to health and nutrition. Second, we extract from a social media platform (in this case, Twitter), all postings matching these keywords, as well as public replies to these postings. Third, we follow all links from the postings to web pages, and download such pages; while the majority of them are news sites and blogs of various kinds, we do not restrict the collection by type of site at this point. Fourth, we follow all links from the web pages to URLs in a series of pre-defined domain names from scientific repositories, academic portals and libraries, and universities. Fifth, we clean-up the collection by applying a series of heuristics to de-duplicate articles and remove noisy entries.

Quality Indicators (§5). We compute a series of quality indicators from the content of articles, and from their referencing social media postings and referenced scientific literature.

Regarding the content of the articles, we begin by computing several content-based features described by previous work. Next, we perform an analysis of quotes in articles, which are a part of journalistic practices in general and are quite prevalent in the case of scientific news. Given that attributed quotes are more telling of high quality than unattributed or “weasel” quotes, we also carefully seek to attribute each quote to a named entity which is often a scientist, but can also be an institution.

Regarding the scientific literature, we would like to know the strength of the connection of articles to scientific papers. For this, we consider two groups of indicators: content-based and graph-based. Content-based indicators are built upon various metrics of text similarity between the content of an article and the content of scientific papers, considering that the technical vocabulary is unlikely to be preserved as-is in articles written for the general public. Graph-based indicators are based on a diffusion graph in which scientific papers and web pages in academic portals are nodes connected by links. High-quality articles are expected to be connected through many short paths to academic sources in this graph.

Regarding social media postings, we measure two aspects: reach and stance. Reach is measured through various proxies for attention, that seek to quantify the impact that an article has in social media. The stance is the positioning of posting authors with respect to an article, which can be positive (supporting, or commenting on an article without expressing doubts), or negative (questioning an article, or directly contradicting what the article is saying).

Evaluation (§6). We evaluate the extent to which the indicators computed in SciLens are useful for determining the quality of a scientific news article. We consider that these indicators can be useful in two ways. First, in a semi-automatic setting, we can show the indicators to end-users and ask them to evaluate the quality of a scientific news article; if users that see these indicators are better at this task that users that do not see them, we could claim that the indicators are useful. Second, in a fully automatic setting, we can train a model based on all the indicators that we computed. In both cases, the ground truth for evaluation is provided by experts in communication and science.

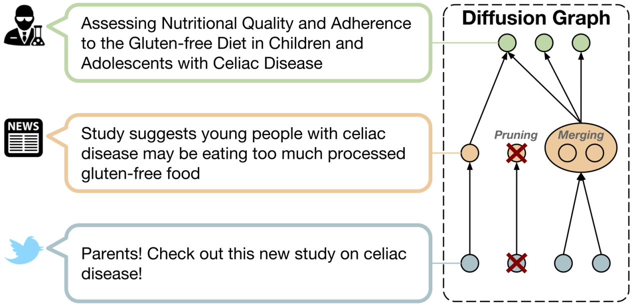

Assessing Nutritional Quality and Adherence to the Gluten-free Diet in Children and Adolescents with Celiac Disease
NEWS
Study suggests young people with celiac disease may be eating too much processed gluten-free food
Parents! Check out this new study on celiac disease!
Diffusion Graph
Pruning
Merging

Figure 2: Contextual data collection, including social media postings, which reference a series of news articles, which cite one or more scientific papers. In our diffusion graph, paths that do not end up in a scientific paper or paths that contain unparsable nodes (e.g., malformed HTML pages) are *pruned*, and articles with the same content in two different outlets (e.g., produced by the same news agency) are *merged*.

4 CONTEXTUAL DATA COLLECTION

The contextual data collection in our work seeks to capture all relevant content for evaluating news article quality, including referenced scientific papers and reactions in social media. This methodology can be applied to any specialized or technical domain covered in the news, as long as: (i) media coverage in the domain involves “translating” from primary technical sources, (ii) such technical sources can be characterized by known host/domain names on the web, and (iii) social media reactions can be characterized by the presence of certain query terms. Examples where this type of contextual data collection could be applied beyond scientific news include news coverage of specialized topics such as law or finance.

We consider two phases: a crawling phase, which starts from social media and then collects news articles and primary sources (§4.1), and a pruning/merging phase, which starts from primary sources and prunes/de-duplicates articles and postings (§4.2). This process is depicted in Figure 2 and explained next.

4.1 Crawling of Postings, Articles, and Papers

The crawling phase starts with social media postings, which are identified as candidates for inclusion based on the presence of certain topic-related keywords in them. In the case of this study, we selected “health and nutrition” as our main topic: this is among the most frequent topics in scientific news reporting, which is known to have a medical/health orientation [4, 15, 61]. The initial set of keywords was obtained from Nutrition Facts (https://nutritionfacts.org/topics), a non-commercial and non-profit website that provides scientific information on healthy eating. The list contains over 2,800 keywords and key phrases such as “HDL cholesterol,” “polyphenols” and the names of hundreds of foods from “algae” to “zucchini”. We further expanded this list with popular synonyms from WordNet [42].

We harvest social media postings from DataStreamer.io (formerly known as Spinn3r.com), covering a 5-year period from June 2013 through June 2018. In this collection, we find $2.5M$ candidate postings matching at least one of our query terms from which we discard postings without URLs.

Next, we crawl the pages pointed to by each URL found in the remaining postings. These pages are hosted in a wide variety of domains, the majority being news outlets and blogging platforms. We scan these pages for links to scientific papers, which we do identify by domain names. We use a predefined list of the top-1000 universities as indicated by CWUR.org plus a manually curated list of open-access publishers and academic databases obtained from Wikipedia[2] and expanded using the "also visited websites" functionality of SimilarWeb.com. Overall, we obtained a diffusion graph of **2.4M** nodes and **3.7M** edges.

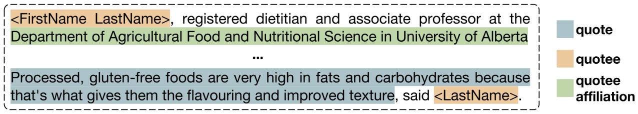
4.2 Pruning and Merging

The initial data collection described in §4.1 is recall-oriented. Now, we make it more precise by pruning and merging items.

Pruning. During the pruning phase, we keep in our collection only documents that we managed to successfully download and parse (e.g., we discard malformed HTML pages and PDFs). We also prune paths that do not end up in a scientific paper i.e., articles that do not have references and all the tweets that point to these articles. This phase helps us eliminate most of the noisy nodes of the diffusion graph that were introduced due to the generic set of seed keywords that we used in the crawling phase (§4.1).

Merging. We notice a large number of duplicate articles across news outlets, which we identify by text similarity i.e, by cosine similarity of more than 90% between the bag-of-words vectors representing the articles. This happens when one outlet re-posts an article originally published in another outlet, or when both syndicate from the same news agency. Once we find such duplicates or near-duplicates, we keep only one of them (the one having more out-links, breaking ties arbitrarily) and remove the redundant ones. Social media postings that point to the duplicates are re-wired to connect to the one that survived after merging, hence we do not lose a potentially important signal of article quality.

The resulting collection is large and mostly composed of elements that are closely related to the topic of health and nutrition: **49K** social media postings, **12K** articles (most of them in news sites and blogs), and **24K** scientific links (most of them peer-reviewed or grey-literature papers). Even after pruning, our collection is possibly more comprehensive than the ones used by systems used to appraise the impact of scientific papers. For instance, when compared to Altmetric.com [1] we find that our collection has more links to scientific papers than what Altmetric counts. In their case, referencing articles seem to be restricted to a controlled list of mainstream news sources, while in our case we often find these references plus multiple references from less known news sources, blogs, and other websites.



<FirstName LastName>, registered dietitian and associate professor at the Department of Agricultural Food and Nutritional Science in University of Alberta

...

Processed, gluten-free foods are very high in fats and carbohydrates because that's what gives them the flavouring and improved texture, said <LastName>.

quote
quotee
quotee affiliation

Figure 3: Example of quote extraction and attribution (best seen in color). Quotee has been anonymized.

5 QUALITY INDICATORS

We compute indicators from the content of news articles (§5.1), from the scientific literature referenced in these articles (§5.2), and from the social media postings referencing them (§5.3). The full list of indicators is presented on Table 2.

5.1 News Article Indicators

These indicators are based on the textual content of a news article.

5.1.1 Baseline Indicators. As a starting point, we adopt a large set of content-based quality indicators described by previous work. These indicators are: (i) title deceptiveness and sentiment: we consider if the title is "clickbait" that oversells the contents of an article in order to pique interest [41, 60]; (ii) article readability: indicator of the level of education someone would need to easily read and understand the article [19]; and (iii) article length and presence of author byline [62].

5.1.2 Quote-Based Indicators. Quotes are a common and important element of many scientific news articles. While selected by journalists, they provide an opportunity for experts to directly present their viewpoints in their own words [12]. However, identifying quotes in general is challenging, as noted by previous work (§2.3). In the specific case of our corpus, we observe that they are seldom contained in quotation marks in contrast to other kinds of quotes (e.g., political quotes [51]). We also note that each expert quoted tends to be quoted once, which makes the problem of *attributing* a quote challenging as well.

Quote Extraction Model. To extract quotes we start by addressing a classification problem at the level of a sentence: we want to distinguish between quote-containing and non-containing sentences. To achieve this, we first select a random sample from our dataset, then manually identify quote patterns, and finally, we generalize automatically these patterns to cover the full dataset. As we describe in the related work section (§2.3), this is a "bootstrapping" model built from high-precision patterns, as follows.

The usage of *reporting verbs* is a typical element of quote extraction models [47]. Along with common verbs that are used to quote others (e.g., "say," "claim") we used verbs that are common in scientific contexts, such as "prove" or "analyze." First, we manually create a seed set of such verbs. Next, we extend it with their nearest neighbors in a word embedding space; the word embeddings we use are the *GloVe* embeddings, which are trained on a corpus of Wikipedia articles [49]. We follow a similar approach for nouns related to studies (e.g., "survey," "analysis") and nouns related to scientists (e.g., "researcher," "analyst"). Syntactically, quotes are usually expressed using indirect speech. Thus, we also obtain part-of-speech tags from the candidate quote-containing sentences.

[2] https://en.wikipedia.org/wiki/List_of_academic_databases_and_search_engines

Using this information, we construct a series of regular expressions over *classes* of words ("reporting verbs," "study-related noun," and part-of-speech tags) which we evaluate in §6.1.

Quote Attribution. For the purposes of evaluating article quality, it is fundamental to know not only that an article has quotes, but also their provenance: *who* or *what* is being quoted. After extracting all the candidate quote-containing sentences, we categorize them according to the information available about their quotee.

A quotee can be an *unnamed scientist* or an *unnamed study* if the person or article being quoted is not disclosed (e.g., "researchers believe," "most scientists think" and other so-called "weasel" words). Sources that are not specifically attributed such as these ones are as a general rule considered less credible than sources in which the quotee is named [62].

A quotee can also be a *named entity* identifying a specific person or organization. In this case, we apply several heuristics for quote attribution. If the quotee is a *named person*, if she/he is referred with her/his last or first name, we search within the article for the full name. When the full name is not present in the article, we map the partial name to the most common full name that contains it within our corpus of news articles. We also locate sentences within the article that mention this person together with a named organization. This search is performed from the beginning of the article as we assume they follow an *inverted pyramid* style. In case there are several, the most co-mentioned organization is considered as the affiliation of the quotee.

If the quotee is an *organization*, then it can be either mentioned in full or using an acronym. We map acronyms to full names of organizations when possible (e.g., we map "WHO" to "World Health Organization"). If the full name is not present in an article, we follow a similar procedure as the one used to determine the affiliation of a researcher, scanning all the articles for co-mentions of the acronym and a named organization.

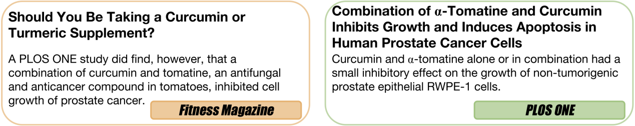
An illustrative example of the extraction and the attribution phase can be shown in Figure 3.

Scientific Mentions. News articles tend to follow journalistic conventions rather than scientific ones [15]; regarding citation practices, this implies they seldom include formal references in the manner in which one would find them in a scientific paper. Often there is no explicit link: journalists may consider that the primary source is too complex or inaccessible to readers to be of any value, or may find that the scientific paper is located in a "pay-walled" or otherwise inaccessible repository. However, even when there is no explicit link to the paper(s) on which an article is based, good journalistic practices still require to identify the information source (institution, laboratory, or researcher).

Mentions of academic sources are partially obtained during the quote extraction process (§5.1.2), and complemented with a second pass that specifically looks for them. During the second pass, we use the list of universities and scientific portals that we used during the *crawling phase* of the data collection (§4.1).

5.2 Scientific Literature Indicators

In this section, we describe content- and graph-based indicators measuring how articles are related to the scientific literature.



Should You Be Taking a Curcumin or Turmeric Supplement?

A PLOS ONE study did find, however, that a combination of curcumin and tomatine, an antifungal and anticancer compound in tomatoes, inhibited cell growth of prostate cancer.

Fitness Magazine

Combination of α-Tomatine and Curcumin Inhibits Growth and Induces Apoptosis in Human Prostate Cancer Cells

Curcumin and α-tomatine alone or in combination had a small inhibitory effect on the growth of non-tumorigenic prostate epithelial RWPE-1 cells.

Figure 4: A news article (left) and a scientific paper (right) with Semantic Text Similarity of 87.9%. Indicatively, two passages from these documents, whose conceptual similarity is captured by our method, are presented. In these two passages we can see the effort of the journalist on translating from an academic to a less formal language, without misrepresenting the results from the paper.

5.2.1 Source Adherence Indicators. When there is an explicit link from a news article to the URL where a scientific paper is hosted, we can measure the extent to which these two documents convey the same information. This is essentially a computation of the *Semantic Text Similarity* (STS) between the news article and its source(s).

Supervised Learning for STS. We construct an STS model using supervised learning. The **features** that we use as input to the model consist of the following text similarity metrics: (i) the Jaccard similarity between the sets of named entities (persons and organizations), dates, numbers and percentages of the two texts; (ii) the cosine similarity between the *GloVe* embeddings of the two texts; (iii) the Hellinger similarity [30] between topic vectors of the two texts (obtained by applying LDA [7]); and (iv) the relative difference between the length in words of the two texts. Each of them is computed three times: (1) considering the entire contents of the article and the paper; (2) considering one paragraph at a time, and then computing the average similarity between a paragraph in one document and a paragraph in the other; and (3) considering one sentence at a time, and then computing the average similarity between a sentence in one document and a sentence in the other. In other words, in (2) and (3) we compute the average of each similarity between the Cartesian product of the passages.

The **training data** that we use is automatically created from pairs of documents consisting of a news article and a scientific paper. Whenever a news article has exactly one link to a scientific paper, we add the article and the paper to training data in the positive class. For the negative class, we sample random pairs of news articles and papers. The **learning schemes** used are Support Vector Machine, Random Forests and Neural Networks. Details regarding the evaluation of these schemes are provided in §6.1.2. An example of a highly related pair of documents, as determined by this method, is shown in Figure 4.

Handling Multi-Sourced Articles. When an article has a single link to a scientific paper, we use the STS of them as an indicator of quality. When an article has multiple links to scientific papers, we select the one that has the maximum score according to the STS model we just described. We remark that this is just an indicator of article quality and we do not expect that by itself it is enough to appraise the quality of the article. Deviations from the content of the scientific paper are not always wrong, and indeed a good journalist might consult multiple sources and summarize them in a way that re-phrases content from the papers used as sources.

5.2.2 Diffusion Graph Indicators. We also consider that referencing scientific sources, or referencing pages that reference scientific sources, are good indicators of quality. Figure 2 showing a graph from scientific papers to articles, and from articles to social media postings and from them to their reactions, suggests this can be done using graph-based indicators. We consider the following:

(1) personalized PageRank [29] on the graph having scientific articles and universities as root nodes and news articles as leaf nodes; and
(2) betweenness and degree on the full diffusion graph [22, 23].

Additionally, we consider the traffic score computed by Alexa.com for the website in which each article is hosted, which estimates the total number of visitors to a website.

5.3 Social Media Indicators

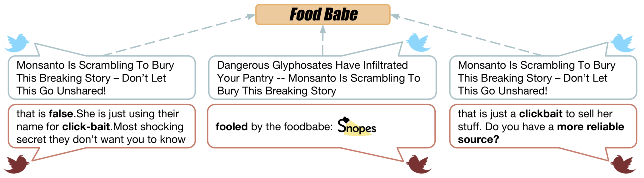
We extract signals describing the quantity and characteristics of social media postings referencing each article. Quantifying the amount of reactions in various ways might give us signals about the interest in different articles (§5.3.1). However, this might be insufficient or even misleading, if we consider that false news may reach a larger audience and propagate faster than actual news [59]. Hence, we also need to analyze the content of these postings (§5.3.2).

5.3.1 Social Media Reach. Not every social media user posting the URL of a scientific news article agrees with the article's content, and not all users have sufficient expertise to properly appraise its contents. Indeed, sharing articles and reading articles are often driven by different mechanisms [2]. However, and similarly to citation analysis and to link-based ranking, the volume of social media reactions to an article might be a signal of its quality, although the same caveats apply.

Given that we do not have access to the number of times a social media posting is shown to users, we extract several proxies of the *reach* of such postings. First, we consider the total number of postings including a URL and the number of times those postings are "liked" in their platform. Second, we consider the number of followers and followees of posting users in the social graph. Third, we consider a proxy for international news coverage, which we operationalize as the number of different countries (declared by users themselves) from which users posted about an article.

Additionally, we assume that a level of attention that is sustained can be translated to a larger exposure and may indicate long-standing interest on a topic. Hence, we consider the temporal coverage i.e., the length of the time window during which postings in social media are observed. To exclude outliers, we compute this period for 90% of the postings, i.e., the article's "shelf life" [9].

5.3.2 Social Media Stance. We consider the stance or positioning of social media postings with respect to the article they link to, as well as the stance of the responses (replies) to those postings. According to what we observe in this corpus, repliers sometimes ask for (additional) sources, express doubts about the quality of an article, and in some cases post links to fact-checking portals that contradict the claims of the article. These repliers are, indeed, acting as "social media fact-checkers," as the example in Figure 5 shows. Following a classification used for analyzing ideological debates [28], we consider four possible stances: supporting, commenting, contradicting, and questioning.



Monsanto Is Scrambling To Bury This Breaking Story – Don't Let This Go Unshared!

Dangerous Glyphosates Have Infiltrated Your Pantry -- Monsanto Is Scrambling To Bury This Breaking Story

Monsanto Is Scrambling To Bury This Breaking Story – Don't Let This Go Unshared!

that is **false**.She is just using their name for **click-bait**.Most shocking secret they don't want you to know

fooled by the foodbabe: Snopes

that is just a **clickbait** to sell her stuff. Do you have a **more reliable source?**

Figure 5: Example in which the stance of social media replies (bottom row) indicates the poor quality of an article promoted through a series of postings (top row).

Retrieving replies. Twitter's API does not provide a programmatic method to retrieve all the replies to a tweet. Thus, we use a web scraper that retrieves the text of the replies of a tweet from the page in which each tweet is shown on the web. The design of this web scraper is straightforward and allows us to retrieve all the *first-level* replies of a tweet.

Classifying replies. To train our stance classifier, we use: (i) a general purpose dataset provided in the context of *SemEval 2016* [43], and (ii) a set of 300 tweets from our corpus which were annotated by crowdsourcing workers. From the first dataset we discard tweets that are not relevant to our corpus (e.g., debates on *Atheism*), thus we keep only debates on *Abortion* and *Climate Change*. The second set of annotated tweets is divided into 97 contradicting, 42 questioning, 80 commenting and 71 supporting tweets. We also have 10 tweets that were marked as "not-related" by the annotators and thus we exclude them from our training process. The combined dataset contains 1,140 annotated tweets. The **learning scheme** we use is a Random Forest classifier based on **features** including the number of: (i) total words, (ii) positive/negative words (using the Opinion Lexicon [31]), (iii) negation words, (iv) URLs, and (v) question/exclamation marks. We also computed the similarity between the replies and the tweet being replied to (using cosine similarity on *GloVe* vectors [49]), and the sentiment of the reply and the original tweet [39]. Details regarding the evaluation are provided in § 6.1.3.

6 EXPERIMENTAL EVALUATION

We begin the experimental evaluation by studying the performance of the methods we have described to extract quality indicators (§6.1). Then, we evaluate if these indicators correlate with scientific news quality. First, we determine if publications that have a good (bad) reputation or track record of rigor in scientific news reporting have higher (lower) scores according to our indicators (§6.2). Second, we use labels from experts (§6.3) to compare quality evaluations done by non-experts with and without access to our indicators (§6.4).

6.1 Evaluation of Indicator Extraction Methods

6.1.1 Quote Extraction and Attribution. The evaluation of our quote extraction and attribution method (§5.1.2) is based on a manually-annotated sample of articles from our corpus. A native English speaker performed an annotation finding 104 quotes (37 quotes attributed to persons, 33 scientific mentions and 34 "weasel" or unattributed quotes) in a random sample of 20 articles.

Table 2: Summary of all the quality indicators provided by the framework SciLens.

Context	Type	Indicator
Article	Baseline	Title [Clickbait, Subjectivity, Polarity], Article Readability, Article Word Count, Article Bylined
	Quote-Based	#Total Quotes, #Person Quotes, #Scientific Mentions, #Weasel Quotes
Sci. literature	Source Adherence	Semantic Textual Similarity
	Diffusion Graph	Personalized PageRank, Betweenness, [In, Out] Degree, Alexa Rank
Social media	Reach	#Likes, #Retweets, #Replies, #Followers, #Followees, [International News, Temporal] Coverage
	Stance	Tweets/Replies [Stance, Subjectivity, Polarity]

We compare three algorithms: (i) a baseline approach based on regular expressions searching for content enclosed in quote marks, which is usually the baseline for this type of task; (ii) our quote extraction method without the quote attribution phase, and (iii) the quote extraction and attribution method, where we consider a quote as correctly extracted if there is no ambiguity regarding the quotee (e.g., if the quotee is fully identified in the article but the attribution finds only the last name, we count it as incorrect).

As we observed, although the baseline approach has the optimal precision, it is unable to deal with cases where quotes are not within quote marks, which are the majority (**100%** precision, **8.3%** recall). Thus, our approach, without the quote attribution phase, improves significantly in terms of recall (**81.8%** precision, **45.0%** recall). Remarkably, the heuristics we use for quote attribution work well in practice and serve to increase both precision and recall (**90.9%** precision, **50.0%** recall). The resulting performance is comparable to state-of-the-art approaches in other domains (e.g., Pavllo et al. [48] obtain **90%** precision, **40%** recall).

6.1.2 Source Adherence. We use the supervised learning method described on §5.2.1 to measure Semantic Text Similarity (STS). We test three different learning models: Support Vector Machine, Random Forests and Neural Networks. The three classifiers use similarities computed at the document, sentence, and paragraph level, and combining all features from the three levels. Overall, the best accuracy (**93.5%**) was achieved by using a Random Forests classifier and all the features from the three levels of granularity, combined.

6.1.3 Social Media Stance. We evaluate the stance classifier described in §5.3.2 by performing 5-fold cross validation over our dataset. When we consider all four possible categories for the stance (supporting, commenting, contradicting and questioning), the accuracy of the classifier is **59.42%**. This is mainly due to confusion between postings expressing a mild support for the article and postings just commenting on the article, which also tend to elicit disagreement between annotators. Hence, we merge these categories into a "supporting or commenting" category comprising postings that do not express doubts about an article. Similarly, we consider "contradicting or questioning" as a category of postings expressing doubts about an article; previous work has observed that indeed false information in social media tends to be questioned more often (e.g., [10]). The problem is then reduced to binary classification.

To aggregate the stance of different postings that may refer to the same article, we compute their weighed average stance considering supporting or commenting as +1 (a "positive" stance) and contradicting or questioning as −1 (a "negative" stance). As weights we consider the popularity indicators of the postings (i.e., the number of likes and retweets). This is essentially a text quantification task [24], and the usage of a classification approach for a quantification task is justified because our classifier has nearly identical pairs of true positive and true negative rates (**80.65%** and **80.49%** respectively), and false positive and false negative rates (**19.51%** and **19.35%** respectively).

6.2 Correlation of Indicators among Portals of Diverse Reputability

We use two lists that classify news portals into different categories by reputability. The first list, by the American Council on Science and Health [3] comprises 50 websites sorted along two axes: whether they produce evidence-based or ideologically-based reporting, and whether their science content is compelling. The second list, by Climate Feedback [17], comprises 20 websites hosting 25 highly-shared stories on climate change, categorized into five groups by scientific credibility, from very high to very low.

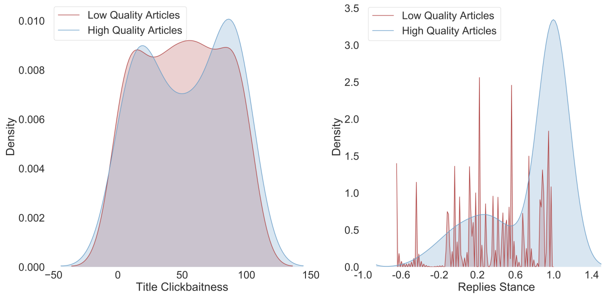
We sample a few sources according to reputability scores among the sources given consistent scores in both lists: high reputability (The Atlantic), medium reputability (New York Times), and low reputability (The Daily Mail). Next, we compare all of our indicators in the sets of articles in our collection belonging to these sources. Two example features are compared in Figure 6. We perform ANOVA [18] tests to select discriminating features. The results are shown on Table 3. Traffic rankings by Alexa.com, scientific mentions, and quotes, are among some of the most discriminating features.

6.3 Expert Evaluation

We ask a set of four external experts to evaluate the quality of a set of articles. Two of them evaluated a random sample of 20 articles about the gene editing technique CRISPR, which is a specialized topic being discussed relatively recently in mass media. The other two experts evaluated a random sample of 20 articles on the effects of Alcohol, Tobacco, and Caffeine (the "ATC" set in the following), which are frequently discussed in science news.

Experts were shown a set of guidelines for article quality based on previous work (§2). Then, they read each article and gave it a score in a 5-point scale, from very low quality to very high quality. Each expert annotated the 20 articles independently, and was given afterwards a chance to cross-check the ratings by the other expert and revise her/his own ratings if deemed appropriate.

The agreement between experts is distributed as follows: (i) *strong agreement*, when the rates after cross-checking are the same (7/20 in ATC, 6/20 in CRISPR); (ii) *weak agreement*, when the rates differ



(a) (b)

Figure 6: Kernel Density Estimation (KDE) of a traditional quality indicator (*Title Clickbaitness* on the left) and our proposal quality indicator (*Replies Stance* on the right). We observe that for both high and low quality articles the distribution of *Title Clickbaitness* is similar, thus the indicator is non-informative. However, most of the high quality articles have *Replies Stance* close to 1.0 which represents the *Supporting/Commenting* class of replies, whereas low quality articles span a wider spectrum of values and often have smaller or negative values representing the *Contradicting/Questioning* class of replies. Best seen in color.

by one point (12/20 in ATC, 10/20 in CRISPR), and (iii) *disagreement*, when the rates differ by two or more points (1/20 in ATC, 4/20 in CRISPR). Annotation results are show on Figure 7, and compared to non-expert evaluations, which are described next.

6.4 Expert vs Non-Expert Evaluation

We perform a comparison of quality evaluations by experts and non-experts. Non-experts are workers in a crowdsourcing platform. We ask for five non-expert labels per article, and employ what our crowdsourcing provider, *Figure Eight* (figure-eight.com), calls tier-3 workers, which are the most experienced and accurate. As a further quality assurance method, we use the agreement among workers to disregard annotators producing consistently annotations that are significantly different from the rest of the crowd. This is done at the worker level, and as a result we remove on average about one outlier judgment per article.

Table 3: Top five discriminating indicators for articles sampled from pairs of outlets having different levels of reputability (p-value: < 0.005 *, < 0.01 **, < 0.05 *).**

The Atlantic vs. Daily Mail (very high vs. very low)	NY Times vs. Daily Mail (medium vs. very low)
Alexa Rank***	Alexa Rank***
#Scientific Mentions***	Article Bylined***
Article Readability**	#Scientific Mentions***
#Total Quotes*	Article Readability***
Title Polarity	#Total Quotes**
The Atlantic vs. NY Times (very high vs. medium)	All Outlets (from very high to very low)
Alexa Rank***	Alexa Rank***
Article Bylined***	Article Bylined***
Article Word Count*	Article Word Count***
#Replies*	#Scientific Mentions***
#Followers	Article Readability***

We consider two experimental conditions. On the first condition (**non-expert without indicators**), non-experts are shown the exact same evaluation interface as experts. On the second condition (**non-expert with indicators**), non-experts are shown 7 of the quality indicators we produced, which are selected according to Table 3. Each indicator (except the last two) is shown with stars, with ★☆☆☆☆ indicating that the article is in the lowest quintile according to that metric, and ★★★★★ indicating the article is in the highest quintile. The following legend is provided to non-experts to interpret the indicators:

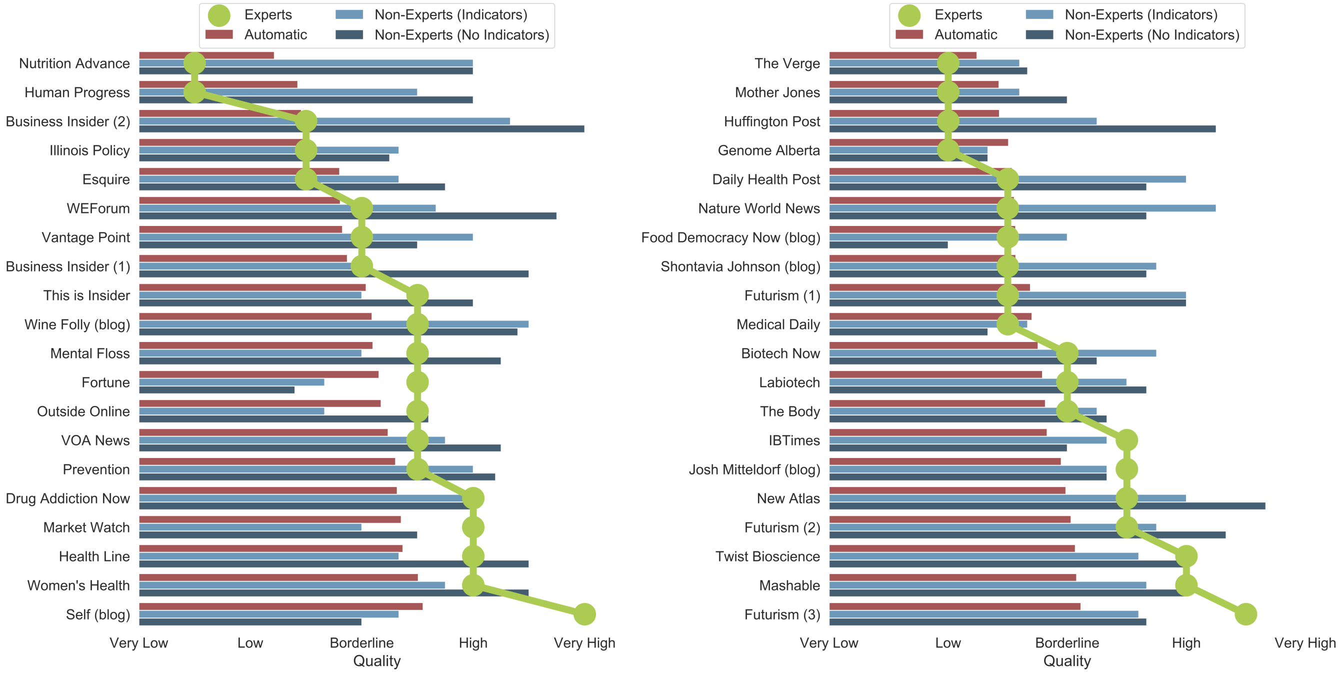
Visitors per day of this news website (more visitors = more stars)
Mentions of universities and scientific portals (more mentions = more stars)
Length of the article (longer article = more stars)
Number of quotes in the article (more quotes = more stars)
Number of replies to tweets about this article) (more replies = more stars)
Article signed by its author (✓= signed,✗= not signed)
Sentiment of the article's title (☺☺= most positive, ☹☹= most negative)

Results of comparing the evaluation of experts and non-experts in the two conditions we have described are summarized in Figure 7. In the figure, the 20 articles in each set are sorted by increasing expert rating; assessments by non-experts differ from expert ratings, but this difference tends to be reduced when non-experts have access to quality indicators.

In Table 4 we show how displaying indicators leads to a decrease in these differences, meaning that non-expert evaluations become closer to the average evaluation of experts, particularly when experts agree. In the ATC set the improvement is small, but in CRISPR

Table 4: Differences among expert evaluations, evaluations provided by non-experts and fully automatic evaluations provided by the SciLens framework, measured using RMSE (lower is better). ATC and CRISPR are two sets of 20 articles each. Strong agreement indicates cases where experts fully agree, weak agreement when they differed by one point, and disagreement when they differed by two or more points. No-Ind. is the first experimental condition for non-experts, in which no indicators are shown. Ind. is the second experimental condition, in which indicators are shown.

	Experts		**Non-Experts**		**Fully**
	by agreement	#	No ind.	Ind.	**automated**
ATC	Strong agreement	7	0.80	**0.45**	1.41
	Weak agreement	12	1.28	1.18	**0.76**
	Disagreement	1	0.40	1.30	**0.00**
	All articles	20	1.10	**1.00**	**1.00**
CRISPR	Strong agreement	6	1.40	1.17	**1.00**
	Weak agreement	10	0.86	0.76	**0.67**
	Disagreement	4	**0.96**	1.22	1.03
	All articles	20	1.96	0.96	**0.85**



(a) Articles on Alcohol, Tobacco, and Caffeine

(b) Articles on CRISPR

Figure 7: Evaluation of two sets of 20 scientific articles. The line corresponds to expert evaluation, while the bars indicate fully automatic evaluation (red), assisted evaluation by non-experts (light blue), and manual evaluation by non-experts (dark blue). Best seen in color.

it is large, bringing non-expert scores about 1 point (out of 5) closer to expert scores.

Table 4 and Figure 7 also includes a fully automated quality evaluation, built using a weakly supervised classifier over all the features we extracted. As weak supervision, we used the lists of sites in different tiers of reputability (§6.2) and considered that *all articles* on each site had the same quality score as the reputation of the site. Then, we used this classifier to annotate the 20 articles in each of the two sets. Results show that this achieves the **lowest** error with respect to expert annotations.

7 CONCLUSIONS

We have described a method for evaluating the quality of scientific news articles. This method, SciLens, requires to collect news articles, papers referenced in them, and social media postings referencing them. We have introduced new quality indicators that consider quotes in the articles, the similarity and relationship of articles with the scientific literature, and the volume and stance of social media reactions. The approach is general and can be applied to any specialized domain where there are primary sources in technical language that are "translated" by journalists and bloggers into accessible language.

In the course of this work, we developed several quality indicators that can be computed automatically, and demonstrated their suitability for this task through multiple experiments. First, we showed several of them are applicable at the site level, to distinguish among different tiers of quality with respect to scientific news. Second, we showed that they can be used by non-experts to improve their evaluations of quality of scientific articles, bringing them more in line with expert evaluations. Third, we showed how these indicators can be combined to produce fully automated scores using weak supervision, namely data annotated at the site level.

Limitations. Our methodology requires access to the content of scientific papers and social media postings. Regarding the latter, given the limitations of the data scrapers we have used only replies to postings and not replies-to-replies. We have also used a single data source for social media postings. Furthermore, we consider a broad definition of "news" to build our corpus, covering mainstream media as well as other sites, including fringe publications. Finally, our methodology is currently applicable only on English corpora.

Reproducibility. Our code uses the following `Python` libraries: `Pandas` and `Spark` for data management, `NetworkX` for graph processing, `scikit-learn` and `PyTorch` for ML, and `SpaCy`, `Beautiful Soup`, `Newspaper`, `TextSTAT` and `TextBlob` for NLP. All the data, code as well as the expert and crowd annotations used in this paper are available for research purposes in ***http://scilens.epfl.ch***.

ACKNOWLEDGMENTS

We would like to thank the external experts who helped us on the evaluation of our framework: **Aina Crosas** (Communication

Officer at the Barcelona Biomedical Research Park), **Andreu Prados** (Pharmacist and Registered Dietitian Nutritionist specialized in Health Communication), **Jose F. Rizo** (Biologist and President of the Fundación Ñamku), and **Dimitra Synefiaridou** (PhD Student in Microbiology at the Veening Lab). This work is partially supported by the Open Science Fund of EPFL (http://www.epfl.ch/research/initiatives/open-science-fund) and the La Caixa project (LCF/PR/PR16/11110009).

REFERENCES

[1] Euan A. Adie and William Roe. 2013. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing* 26, 1 (2013), 11–17. https://doi.org/10.1087/20130103

[2] Deepak Agarwal, Bee-Chung Chen, and Xuanhui Wang. 2012. Multi-faceted ranking of news articles using post-read actions. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, Xue-wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki (Eds.). ACM, 694–703. https://doi.org/10.1145/2396761.2396850

[3] Alex Berezow. March 5, 2017. Infographic: The Best and Worst Science News Sites. *American Council on Science and Health* (March 5, 2017). https://acsh.org/news/2017/03/05/infographic-best-and-worst-science-news-sites-10948

[4] Franziska Badenschier and Holger Wormer. 2012. Issue selection in science journalism: Towards a special theory of news values for science news? In *The sciences' media connection–public communication and its repercussions*. Springer, 59–85.

[5] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 3528–3539. https://aclanthology.info/papers/D18-1389/d18-1389

[6] Martin W Bauer, Nick Allum, and Steve Miller. 2007. What can we learn from 25 years of PUS survey research? Liberating and expanding the agenda. *Public understanding of science* 16, 1 (2007), 79–95.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022. http://www.jmlr.org/papers/v3/blei03a.html

[8] Christina Boididou, Symeon Papadopoulos, Lazaros Apostolidis, and Yiannis Kompatsiaris. 2017. Learning to Detect Misleading Content on Twitter. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, Bogdan Ionescu, Nicu Sebe, Jiashi Feng, Martha Larson, Rainer Lienhart, and Cees Snoek (Eds.). ACM, 278–286. https://doi.org/10.1145/3078971.3078979

[9] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the life cycle of online news stories using social media reactions. In *Computer Supported Cooperative Work, CSCW '14, Baltimore, MD, USA, February 15-19, 2014*, Susan R. Fussell, Wayne G. Lutters, Meredith Ringel Morris, and Madhu Reddy (Eds.). ACM, 211–223. https://doi.org/10.1145/2531602.2531623

[10] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2013. Predicting information credibility in time-sensitive social media. *Internet Research* 23, 5 (2013), 560–588. https://doi.org/10.1108/IntR-05-2012-0095

[11] 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 (jun 2015), e0128193. https://doi.org/10.1371/journal.pone.0128193

[12] Peter Conrad. 1999. Uses of expertise: Sources, quotes, and voice in the reporting of genetics in the news. *Public Understanding of Science* 8 (1999).

[13] Vladimir De Semir. 2000. Scientific journalism: problems and perspectives. *International Microbiology* 3, 2 (2000), 125–128.

[14] Estelle Dumas-Mallet, Andy Smith, Thomas Boraud, and François Gonon. 2017. Poor replication validity of biomedical association studies reported by newspapers. *PLOS ONE* 12, 2 (02 2017), 1–15. https://doi.org/10.1371/journal.pone.0172650

[15] Sharon Dunwoody. 2014. Science journalism: prospects in the digital age. In *Routledge handbook of public communication of science and technology*. Routledge, 43–55.

[16] David K. Elson and Kathleen R. McKeown. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, Maria Fox and David Poole (Eds.). AAAI Press. http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1945

[17] Emmanuel M. Vincent. 17 Jan 2018. Most popular climate change stories of 2017 reviewed by scientists. *Climate Feedback* (17 Jan 2018). https://climatefeedback.org/most-popular-climate-change-stories-2017-reviewed-scientists/

[18] Ronald Aylmer Fisher. 2006. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

[19] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[20] B. J. Fogg and Hsiang Tseng. 1999. The Elements of Computer Credibility. In *Proceeding of the CHI '99 Conference on Human Factors in Computing Systems: The CHI is the Limit, Pittsburgh, PA, USA, May 15-20, 1999.*, Marian G. Williams and Mark W. Altom (Eds.). ACM, 80–87. https://doi.org/10.1145/302979.303001

[21] Jim Foust. 2017. *Online journalism: principles and practices of news for the Web*. Taylor & Francis.

[22] Linton C Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry* (1977), 35–41.

[23] Linton C Freeman. 1978. Centrality in social networks conceptual clarification. *Social networks* 1, 3 (1978), 215–239.

[24] Wei Gao and Fabrizio Sebastiani. 2016. From classification to quantification in tweet sentiment analysis. *Social Netw. Analys. Mining* 6, 1 (2016), 19:1–19:22. https://doi.org/10.1007/s13278-016-0327-z

[25] Alan G. Gross. 1994. The roles of rhetoric in the public understanding of science. *Public Understanding of Science* 3, 1 (1994), 3–23. https://doi.org/10.1088/0963-6625/3/1/001 arXiv:https://doi.org/10.1088/0963-6625/3/1/001

[26] Lushan Han, Justin Martineau, Doreen Cheng, and Christopher Thomas. 2015. Samsung: Align-and-Differentiate Approach to Semantic Textual Similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). The Association for Computer Linguistics, 172–177. http://aclweb.org/anthology/S/S15/S15-2031.pdf

[27] P Sol Hart, Erik C Nisbet, and Teresa A Myers. 2015. Public attention to science and political news and support for climate change mitigation. *Nature Climate Change* 5, 6 (2015), 541.

[28] Kazi Saidul Hasan and Vincent Ng. 2013. Stance Classification of Ideological Debates: Data, Models, Features, and Constraints. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*. Asian Federation of Natural Language Processing / ACL, 1348–1356. http://aclweb.org/anthology/I/I13/I13-1191.pdf

[29] Taher H. Haveliwala. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Trans. Knowl. Data Eng.* 15, 4 (2003), 784–796. https://doi.org/10.1109/TKDE.2003.1208999

[30] Ernst Hellinger. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik* 136 (1909), 210–271.

[31] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel (Eds.). ACM, 168–177. https://doi.org/10.1145/1014052.1014073

[32] Jakob D Jensen. 2008. Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human communication research* 34, 3 (2008), 347–369.

[33] Wei Jin, Hung Hay Ho, and Rohini K. Srihari. 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki (Eds.). ACM, 1195–1204. https://doi.org/10.1145/1557019.1557148

[34] Dan Jurafsky, Angel X. Chang, Grace Muzny, and Michael Fang. 2017. A Two-stage Sieve Approach for Quote Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, 460–470. https://aclanthology.info/papers/E17-1044/e17-1044

[35] Anand Konjengbam, Subrata Ghosh, Nagendra Kumar, and Manish Singh. 2018. Debate Stance Classification Using Word Embeddings. In *Big Data Analytics and Knowledge Discovery - 20th International Conference, DaWaK 2018, Regensburg, Germany, September 3-6, 2018, Proceedings (Lecture Notes in Computer Science)*, Carlos Ordonez and Ladjel Bellatreche (Eds.), Vol. 11031. Springer, 382–395. https://doi.org/10.1007/978-3-319-98539-8_29

[36] Justin Kosslyn and Cong Yu. April 7, 2017. Fact Check now available in Google Search and News around the world. *Google* (April 7, 2017). http://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world

[37] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 591–602. https://doi.org/10.1145/2872427.2883085

[38] Matthias Liebeck, Philipp Pollack, Pashutan Modaresi, and Stefan Conrad. 2016. HHU at SemEval-2016 Task 1: Multiple Approaches to Measuring Semantic

Textual Similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). The Association for Computer Linguistics, 595–601. http://aclweb.org/anthology/S/S16/S16-1090.pdf

[39] Steven Loria. 2018. Sentiment Analysis. (2018). http://textblob.readthedocs.io

[40] Tessa Lyons. May 23, 2018. Hard Questions: What's Facebook's Strategy for Stopping False News? *Facebook* (May 23, 2018). http://newsroom.fb.com/news/2018/05/hard-questions-false-news

[41] Saurabh Mathur. 2017. Clickbait Detector. (2017). http://github.com/saurabhmathur96/clickbait-detector

[42] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41. https://doi.org/10.1145/219717.219748

[43] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *ACM Trans. Internet Techn.* 17, 3 (2017), 26:1–26:23. https://doi.org/10.1145/3003433

[44] Greg Myers. 2003. Discourse studies of scientific popularization: Questioning the boundaries. *Discourse studies* 5, 2 (2003), 265–279.

[45] Timothy O'Keefe, Silvia Pareti, James R. Curran, Irena Koprinska, and Matthew Honnibal. 2012. A Sequence Labelling Approach to Quote Attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, Jun'ichi Tsujii, James Henderson, and Marius Pasca (Eds.). ACL, 790–799. http://www.aclweb.org/anthology/D12-1072

[46] Chiara Palmerini. 2007. Science reporting as negotiation. In *Journalism, Science and Society*. Chapter 11, 113–122.

[47] Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 989–999. http://aclweb.org/anthology/D/D13/D13-1101.pdf

[48] Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable Unsupervised Extraction of Quotation-Speaker Pairs from Large News Corpora via Bootstrapping. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018.* AAAI Press, 231–240. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17827

[49] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. http://aclweb.org/anthology/D/D14/D14-1162.pdf

[50] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 1003–1012. https://doi.org/10.1145/3041021.3055133

[51] Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *Proceedings of Recent Advances in Natural Language Processing*. 487–492.

[52] Andrew Salway, Paul Meurer, Knut Hofland, and Øystein Reigem. 2017. Quote Extraction and Attribution from Norwegian Newspapers. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA 2017, Gothenburg, Sweden, May 22-24, 2017*, Jörg Tiedemann and Nina Tahmasebi (Eds.). Association for Computational Linguistics, 293–297. https://aclanthology.info/papers/W17-0241/w17-0241

[53] Laura Sbaffi and Jennifer Rowley. 2017. Trust and Credibility in Web-Based Health Information: A Review and Agenda for Future Research. *Journal of medical Internet research* 19, 6 (jun 2017), e218. https://doi.org/10.2196/jmir.7579

[54] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 745–750. https://doi.org/10.1145/2872518.2890098

[55] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explorations* 19, 1 (2017), 22–36. https://doi.org/10.1145/3137597.3137600

[56] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. 2015. Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, 977–982. https://doi.org/10.1145/2740908.2742572

[57] Joseph W. Taylor, Marie Long, Elizabeth Ashley, Alex Denning, Beatrice Gout, Kayleigh Hansen, Thomas Huws, Leifa Jennings, Sinead Quinn, Patrick Sarkies, Alex Wojtowicz, and Philip M. Newton. 2015. When Medical News Comes from Press ReleasesâĂŤA Case Study of Pancreatic Cancer and Processed Meat. *PLOS ONE* 10, 6 (06 2015), 1–13. https://doi.org/10.1371/journal.pone.0127848

[58] Juliane Urban and Wolfgang Schweiger. 2014. News Quality from the Recipients' Perspective. *Journalism Studies* 15, 6 (2014), 821–840. https://doi.org/10.1080/1461670X.2013.856670 arXiv:https://doi.org/10.1080/1461670X.2013.856670

[59] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559 arXiv:http://science.sciencemag.org/content/359/6380/1146.full.pdf

[60] Wei Wei and Xiaojun Wan. 2017. Learning to Identify Ambiguous and Misleading News Headlines. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, Carles Sierra (Ed.). ijcai.org, 4172–4178. https://doi.org/10.24963/ijcai.2017/583

[61] Emma Weitkamp. 2003. British newspapers privilege health and medicine topics over other science news. *Public Relations Review* 29, 3 (2003), 321–333.

[62] Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David R. Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 603–612. https://doi.org/10.1145/3184558.3188731

[63] Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf. Process. Manage.* 54, 2 (2018), 273–290. https://doi.org/10.1016/j.ipm.2017.11.009