SciLens News Platform: A System for Real-Time Evaluation of News Articles

Angelika Romanou[†] Panayiotis Smeros[†] Carlos Castillo[‡] Karl Aberer[†]

[†]École Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
firstname.lastname@epfl.ch

[‡]Universitat Pompeu Fabra (UPF)
Barcelona, Spain
carlos.castillo@upf.edu

ABSTRACT

We demonstrate the SciLens News Platform, a novel system for evaluating the quality of news articles. The SciLens News Platform automatically collects contextual information about news articles in real-time and provides quality indicators about their validity and trustworthiness. These quality indicators derive from i) social media discussions regarding news articles, showcasing the reach and stance towards these articles, and ii) their content and their referenced sources, showcasing the journalistic foundations of these articles. Furthermore, the platform enables domain-experts to review articles and rate the quality of news sources. This augmented view of news articles, which combines automatically extracted indicators and domain-expert reviews, has provably helped the platform users to have a better consensus about the quality of the underlying articles. The platform is built in a distributed and robust fashion and runs operationally handling daily thousands of news articles. We evaluate the SciLens News Platform on the emerging topic of *COVID-19* where we highlight the discrepancies between low and high-quality news outlets based on three axes, namely their newsroom activity, evidence seeking and social engagement. A live demonstration of the platform can be found here: *http://scilens.epfl.ch*.

PVLDB Reference Format:
Angelika Romanou, Panayiotis Smeros, Carlos Castillo, Karl Aberer. SciLens News Platform: A System for Real-Time Evaluation of News Articles. *PVLDB*, 13(12): 2969-2972, 2020.
DOI: https://doi.org/10.14778/3415478.3415521

1. INTRODUCTION

In the age of information inflation, the news is not always produced and consumed in a centralized fashion. Although media companies with specialized journalists are typically responsible for discovering and communicating news to the people, there are intricate ways in which this information is diffused towards the public, mostly via social media [4]. The news landscape has changed radically, mainly because of: i) the instantaneous rate at which individuals publish news-worthy content, ii) the vast reachability of this content by broad audiences, and iii) the lack of regulation and quality control. The modern news landscape consists of mainstream news outlets which are supported, complemented and often criticized by independent or alternative media channels.

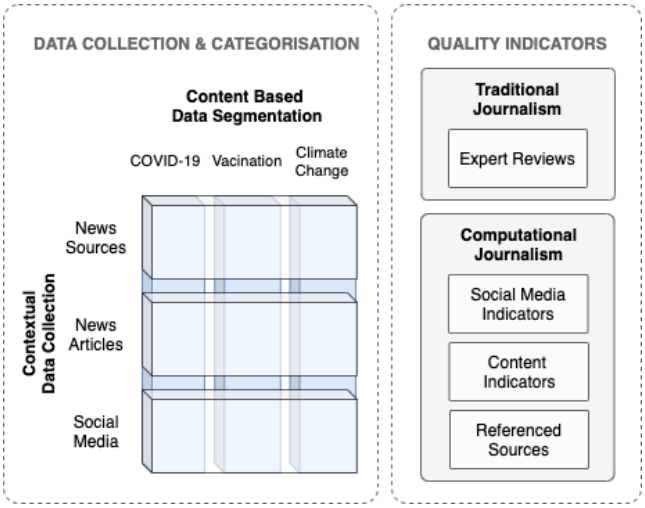

Figure 1: Overview of the SciLens News Platform, a system that collects contextual information and provides a wide range of heterogeneous quality indicators regarding news articles.

However, this plethora of news sources has inevitably led to a burst of misinformation [5]. Fake or fabricated stories are continuously spread among large groups of people, while credible information is often indistinguishable from misleading content. This misleading content is a result of either detrimental and malicious behaviours (e.g., in politics) or poor interpretation and lack of domain knowledge (e.g., in science). In the meanwhile, news companies that want to keep their users engaged, provide personalized news palettes, reinforcing preexisting media biases and empowering the influence of "echo chambers".

There is a wide-ranging amount of endeavours against misinformation. These endeavours span the area of traditional journalism, with independent fact-checking initiatives (e.g., Snopes.com), as well as the upcoming area of data journalism [2], with computational methodologies against misinformation. However, a modern news platform should be able to combine these two methodologies by i) having a robust way of processing the vast amount of information that is continuously produced, and ii) combine automatically-extracted and domain-agnostic quality indicators with the reliable but small-scale expert reviews.

In this paper, we demonstrate the SciLens News Platform, a novel system for evaluating the quality of news articles that bridges

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3415478.3415521

the gap between traditional and computational journalism. Our platform automatically extracts, stores, and displays heterogeneous quality indicators for scientific news, that were first introduced by Smeros et al. [9]. These indicators derive from i) social media discussions regarding news articles, showcasing the reach and stance towards these articles, and ii) their content and their referenced sources, showcasing the journalistic foundations of these articles. The quality indicators are combined with expert reviews in a unified environment (Figure 1). The SciLens News Platform is build in a distributed and robust fashion and runs operationally handling daily thousands of news articles. In the following sections, we present an overview of the system and a tangible use-case on the emerging topic of *COVID-19*.

2. RELATED WORK

The area of computational journalism is a broad research area where results are scattered through multiple disciplines. An extended survey on the related work of this paper is presented by Smeros et al. [9].

As mentioned above, the traditional approach for evaluating news article quality relies on the manual work of domain experts. Independent, non-partisan fact-checking portals perform manual content verification either on general topics (e.g., Snopes.com) or specific topics such as politics (e.g., PolitiFact.com) and science (e.g., ScienceFeedback.co).

Recent work demonstrates methods to automate the extraction of quality indicators. Indicatively, Zhang et al. [10] compile a detailed list of such quality indicators, while Shu et al. [8] introduce an approach for detecting fake news on social media based on these indicators. Ciampaglia et al. [3] and Hassan et al. [6] use well-known knowledge bases like Wikipedia as ground truth for testing the validity of dubious claims, while Popat et al. [7] describe a system that explains the news articles' stance towards such claims.

3. SYSTEM OVERVIEW

The SciLens Data Platform incorporates automated quality indicators for news articles (§3.1), as well as a systematic way of acquiring expert reviews (§3.2). The overall architecture of the system is presented in §3.3.

3.1 Automated Quality Indicators

We compute three heterogeneous sets of quality indicators, namely, content, news context, and social media indicators. Regarding the content of a news article, we consider various well-established metrics for the quality of news such as the clickbaitness of its title, the subjectivity, and readability of its body and whether it is by-lined by its author.

As for the news context of an article, we investigate the strength of the connection between this article and its primary sources of information. Thus, we consider three types of references: i) internal references within the same news outlet; many news outlets, in order to increase their user engagement, introduce such references either in "see also" sections or in the main body of their articles, ii) external references to potential primary sources of information (e.g., references from nation-wide news outlets to local news outlets), and iii) particularly for the case of scientific news, scientific references, i.e., references to a predefined list of academic repositories, grey-literature and peer-reviewed journals and institutional websites; as we see in our use-case (§3.2), articles from high-quality scientific outlets are expected to have more references pointing to academic sources than articles from low-quality scientific outlets.

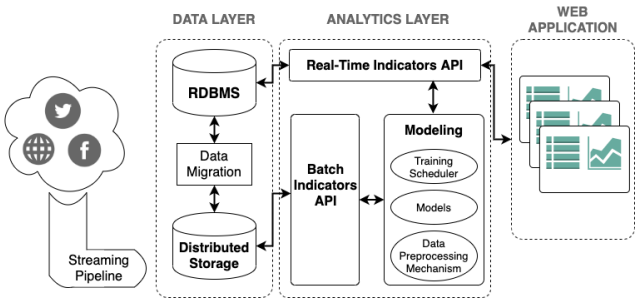

Figure 2: SciLens News Platform architecture. First, a streaming pipeline acts as the entry point of data collection. Then, a data layer, comprised by an RDBMS and a Distributed Storage, stores the incoming data. Lastly, the analytics layer manages the data, trains the Machine Learning models and serves the extracted indicators to the web application.

Finally, regarding the social media context, we measure two aspects, specifically the reach and stance towards a news article. Reach is measured through the proxy of social media popularity, which quantifies the impact of an article in a social media platform. Stance, on the other hand, is the positioning of social media platform users towards an article. Stance can be positive (i.e, users support or comment on an article without expressing doubts), or negative (i.e., users question the quality of an article, or directly contradict what the article is saying).

According to a thorough experimental evaluation which is presented by Smeros et al. [9], the aforementioned indicators help non-expert users evaluate more accurately the quality of news articles, compared to non-experts that do not have access to these indicators.

3.2 Expert Reviews

Along with the set of automated quality indicators, the system allows experts to annotate any article based on seven criteria: 1) Factual accuracy, 2) Scientific understanding, 3) Logic/Reasoning, 4) Precision/Clarity, 5) Sources quality, 6) Fairness, and 7) Clickbaitness on a *Likert Scale*[1], from very low quality to very high quality. These are common criteria used in state-of-the-art fact-checking portals like ScienceFeedback.co.

Based on these evaluation scores, the system computes a weighted, time-sensitive average and displays a final score of the criteria for each article. Optionally, expert users can provide with extensive free-text reviews about the news articles, which are also displayed to the non-expert users.

3.3 Architecture

The architecture of the Scilens News Platform consists of three components which are responsible for the collection, storage and segmentation of data as well as for the models training and the indicators serving to the web application. Bellow we describe these components which are also shown in Figure 2.

Data Collection and Storage. The SciLens News Platform uses a hybrid data storage scheme that supports both real-time computational operations (with an RDBMS), and ad-hoc querying on historical data and efficient data warehousing (with a Distributed Storage). The main data entry point of the system is an outlet-based streaming pipeline wrapped around the Datastreamer API (http://www.datastreamer.io). This subsystem acts as a messaging queue and fetches, in real-time, postings from a specific set

[1]https://en.wikipedia.org/wiki/Likert_scale

of social media accounts along with their reactions. These incoming data streams are processed, and the corresponding news articles are extracted. For these transformations, the system leverages the distributed file system of Hadoop (http://hadoop.apache.org), and the distributed computational framework Spark (http://spark.apache.org), for parallel data processing and storing. The data synchronization between the RDBMS and the Distributed Storage is made through a daily data migration process.

Data Management and Model Training. As we show in our use-case (§4), an essential aspect of our system is the computation of analytics on top of particular segments of our data. These segments are combinations of content-based supervised topics of news and quality-based categories of news outlets.

More specifically, regarding the content-based segmentation, the system performs a probabilistic hierarchical clustering on the articles and assigns one or more topics to each one of them. These topics can be very generic (e.g., Health) or very specific (e.g., COVID-19). On the other hand, regarding the outlet quality-based segmentation, the system groups the articles based on the news outlet that they are published and then groups the outlets that have similar quality. The quality of an outlet is either computed using the expert reviews or imported from external sources (e.g., in §4 we use a ranking published by the American Council on Science and Health).

Finally, our system periodically trains Machine Learning models on top of the Distributed Storage, accessing the full history of our data. These models are used for the extraction of the quality indicators that we described in §3.1.

Indicators API. The last core component of our system is the Indicators API, which is responsible for the real-time article evaluation. Its architecture is based on micro-services, which are lightweight, loosely coupled services that support parallel execution. The main functionality of this component is to compute and serve quality indicators of articles to the web application.

4. DEMONSTRATION PLAN

Our demonstration will showcase the SciLens News Platform on the topic of the pandemic of *Coronavirus Disease 2019* (*COVID-19*). *COVID-19* is a topic with highly trending nature which triggers an abundance of news articles and social media discussions. Given such a prominent topic, the task of discerning between low and high-quality articles becomes very challenging for non-experts in the fields of medicine and epidemiology. On that end, we present how fused information retrieved from our system allows end-users to i) assess the quality of individual news articles, and ii) obtain aggregated insights for the topic of *COVID-19*.

To prepare the *COVID-19* data segment, the system uses a shortlist, published by the American Council on Science and Health [1], that contains 45 mainstream news outlets accompanied by their quality ranking. The time frame of the data collection, in the context of this paper, covers the 60-day period from 2020-01-15 to 2020-03-15. A live demonstration of the platform, with up-to-date news articles and quality indicators, can be found here: http://scilens.epfl.ch.

4.1 Single Article Assessment

As we explain in §3, an end-user of the platform can explore in real-time, a wide range of automatically extracted quality indicators combined with manually-operated expert reviews. A snapshot of this enhanced view of news articles is depicted in Figure 3. This functionality is available for all the articles in our news collection as well as for any arbitrary news article that a user wants to evaluate.

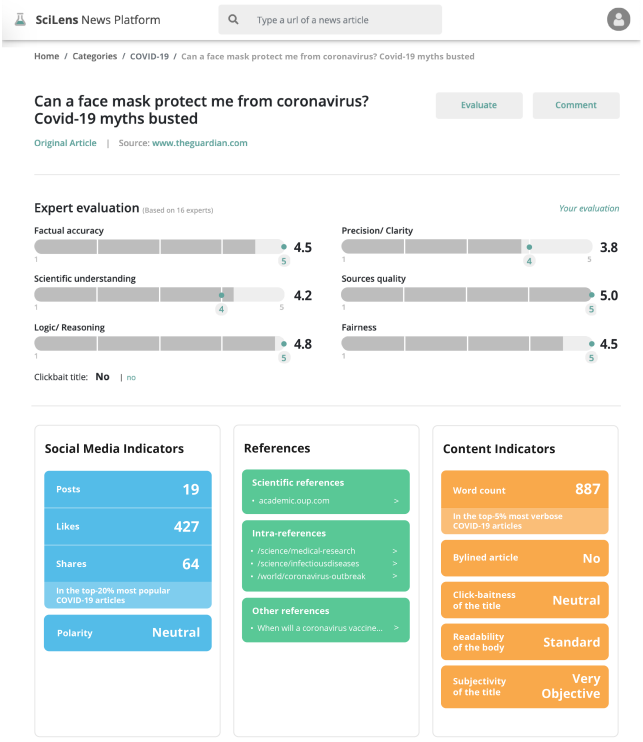

Figure 3: User interface of the SciLens News Platform. A wide range of automatically extracted quality indicators combined with manually-operated expert reviews.

4.2 News Topic Insights

Apart from the single article assessment, a user can interact with aggregated insights regarding a news topic (in our case *COVID-19*). The news outlets that publish articles regarding *COVID-19* are evaluated based on three axes, namely their newsroom activity, evidence seeking and social engagement.

To study the newsroom activity, the system computes the distribution of daily posts for each of the outlets. Then, it groups all the media outlets by their quality ranking and creates a time-series of the mean percentage of daily posts per rating class. The results an end-user will see are presented in Figure 4.

We observe that in the early stages of the discussion on the pandemic, both low and high-quality outlets posted with the same frequency. However, by the end of the first month, low-quality outlets started dedicating a larger percentage of their published articles on this topic. The latter implies a trade-off between the quantity and the quality of the articles. Low-quality outlets seem to be driven by the breaking news, whereas high-quality outlets are more conservative on their publication rate; however, as we see next, they have better scientific foundations.

Moreover, the system provides the end-user with insights regarding i) the social engagement (i.e., the number of social media reactions), and ii) the evidence seeking (i.e., in our use-case, the ratio of scientific references used) of the news outlets. As shown in Figure 5, one can verify the assumption that low-quality outlets tend to publish more and thus to acquire a higher amount of social media reach than high-quality outlets. Conversely, high-quality outlets base their findings more on well-established scientific references.

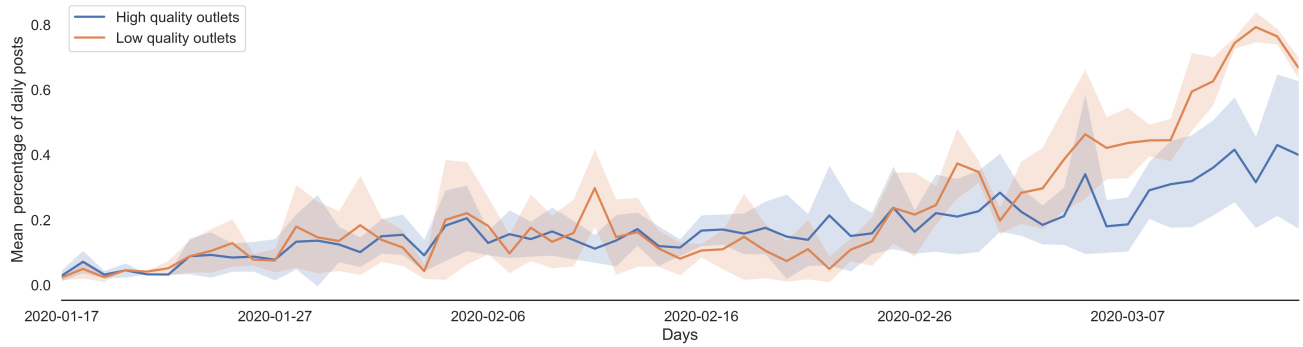

Figure 4: Mean percentage of daily posts referred to COVID-19 per rating category. Low-quality outlets seem to be driven by the breaking news, whereas high-quality outlets are more conservative on their publication rate.

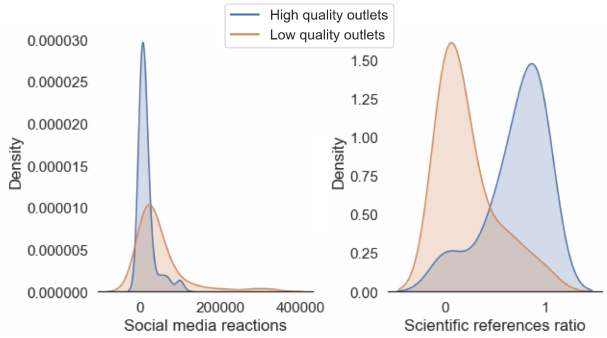

Figure 5: Kernel Density Estimation (KDE) of the number of Social Media Reactions (left) and Scientific References Ratio (right). The low-quality outlets tend to have a wider distribution of reactions but lower number of scientific references, whereas the high-quality outlets show the inverse behaviour.

References

[1] A. Berezow. *Infographic: The Best and Worst Science News Sites*. American Council on Science and Health, 2017.

[2] S. Cazalens, P. Lamarre, J. Leblay, I. Manolescu, and X. Tannier. A content management perspective on fact-checking. In P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 565–574. ACM, 2018. doi: 10.1145/3184558.3188727.

[3] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational Fact Checking from Knowledge Networks. *PLOS ONE*, 10(6):e0128193, jun 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0128193.

[4] Elisa Shearer. Social media outpaces print newspapers in the U.S. as a news source. *Pew Research Center*, 10/12/2018.

[5] M. Fernández and H. Alani. Online misinformation: Challenges and future directions. In P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 595–602. ACM, 2018. doi: 10.1145/3184558.3188730.

[6] N. Hassan, F. Arslan, C. Li, and M. Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 1803–1812. ACM, 2017. ISBN 978-1-4503-4887-4. doi: 10.1145/3097983.3098131.

[7] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017*, pages 1003–1012. ACM, 2017. ISBN 978-1-4503-4914-7. doi: 10.1145/3041021.3055133.

[8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, 19(1):22–36, 2017. doi: 10.1145/3137597.3137600.

[9] P. Smeros, C. Castillo, and K. Aberer. Scilens: Evaluating the quality of scientific news articles using social media and scientific literature indicators. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 1747–1758. ACM, 2019. doi: 10.1145/3308558.3313657.

[10] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, E. Bice, S. Hawke, D. R. Karger, and A. X. Mina. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 603–612. ACM, 2018. doi: 10.1145/3184558.3188731.