Media Observatory Initiative

REPORT – NOVEMBER 2020

JÉRÉMIE RAPPAZ

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Office of Communications OFCOM

EPFL

INITIATIVE FOR MEDIA INNOVATION

LE TEMPS

Abstract

Moving toward digitalization, the news ecosystem goes through a transition phase that eases access to topical issues but tends to amplify disinformation, consequently undermining public trust in media institutions. The Media Observatory Initiative aims to collect news article at a global scale, to propose algorithmic solutions to monitor the news landscape and to research processing method that benefit the public. The project is a collaboration between École Polytechnique Fédérale de Lausanne (EPFL) and the newspaper *Le Temps*. The platform is accessible at the following address: https://newsteller.io.

Context

The news landscape, in Switzerland and in the world, has been largely affected by a rapid evolution of the information sector and by the digital transformation. The advent of the Web has changed news consumption, with more articles read on screen, but also news production, with newsrooms making increasingly use of information technologies.

Media in the Digital Age

Traditionally, readers had to choose their primary source of information from well-known news entities with comprehensive editorial lines, in full knowledge of ideological or political leanings associated to them. The Web broke this equilibrium: news sources of heterogeneous type and quality have emerged online and are now sharing same formats and distribution schemes as traditional news actors. Distinguishing between various types of digital offers is, consequently, increasingly difficult for the public. Moreover, the access to massive potential audiences at extremely low costs has enabled a rapid proliferation of misleading or deceptive content, such as news stories with underlying political or financial goals. This phenomenon creates both incomprehension from the public, for whom it is hard to assess the credibility of a piece of content, and from the media, who need to reinvent their authoritative role in news landscape.

Social Media

Beyond the evolution of online news formats, the digital shift fundamentally changed the way news are carried to the reader. Indeed, Social Media now play a critical role in the discovery and spread of news stories. This observation implies that readers' exposition to specific types of content and media outlets, is dictated by the structure of social networks and by undocumented and unregulated algorithms. Those algorithms are typically built for retaining user attention and optimizing engagement metrics which are not necessarily aligned with informational objectives. Research[1] suggests that the perceived credibility of news is reduced on Social Media, where articles are mixed with various types of entertainment and non-informational content. Finally, Social Media is known to make brand recognition difficult for news outlets, since reader reportedly pay less attention to the source of an information online, consequently deteriorating even more public trust in news institutions.

[1] https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

Misinformation

The pervasive use of information technologies has significantly increased the rate at which individuals and organizations publish news-worthy content. The problem is broader than deceptive news articles since every individual is now susceptible to publish and share misleading or rumorous content, with the potential to target Web-scale audiences. In this distributed context, factual information is often difficult to isolate from large spectra of divergent opinions, rumors or agenda-setting discourses. Numerous examples show the potential impact of misinformation on financial markets, on health issues and on the democratic process. As the problem is not directly tied to the media ecosystem, it creates a climate of online distrust that contributes to decrease trust towards online information. A recent study across countries showed that less than four in ten respondents (38%) said they trust most news most of the time [5].

The decreasing perception of credibility online, caused by an environment in which truthful information is often indistinguishable from misleading content, leave citizens unarmed against malicious actors at a time at which accessing verified information is highly needed. Fact-checking organizations propose solutions to fight the rise of fabricated stories, but their responses require manual labor and, therefore, only allow to be applied at small scale.

Quality assessment

Classic economics theories [1] describe how uncertainty about the quality of products and the inability of consumers to properly assess this quality can depress the entire market. This theory can be transposed to current news market, in which the reader cannot distinguish legitimate news from deceptive content anymore. In this hypothetical scenario, sellers are incentivized to sell defective products since their real value is inferior to market price, thus making the willingness-to-pay from buyers decrease. Research concludes that, without means of assessing quality, the presence of individuals who are willing to offer inferior goods can, theoretically, drive the market out of existence. Potential causes of distrust are hypothesized to revolve around the digital gap, but they remain open research questions and seem to create a paradox in which news media and their audiences being steadily more connected while having increasing trouble understanding each other.

Problems

The main objective of the project is to fight misinformation and restore public trust. We identify the following key obstacles (O) that hamper the development of large-scale analyses of news ecosystem.

- **O1**: **The decentralized nature of news media** represents a barrier to access real-time information about world events.
- **O2**: **Diffusion on Social Media**: exposition to news article is often governed by Social Media platforms which suffer from algorithmic biases and expose reader to the so-called *echo-chamber* effect.
- **O3**: **The lack of ground truth** to assess the quality and credibility of news articles: fact-checking can only validate factual statements and is not scalable to large corpora of articles.
- **O4**: **Research is still in its infancy** in fields such as media bias, audience engagement or discourse framing. Studies on these subjects are hard to conduct, due to the limited information made available to researchers.

The Media Observatory Initiative is designed to mitigate these obstacles by providing real-time analyses of news articles published in Switzerland and in the world.

Scope and Objectives

As pointed out by the European Commission [4], regulatory actions are generally not conceivable without harming free speech and liberty of the press. Instead, we believe that key responses to the decreasing trust from the general public lie in the promotion of transparency, information literacy and critical thinking of both the public and the media.

Given the obstacles listed above, we establish the objectives for the Media Observatory Initiative:

- **Fostering information literacy** by providing indicators, both at sources and article level.
- **Monitoring public reactions** in order to quickly identify trends, for information professional to identify news worthy

content, or controversial topics, for helping fact-checkers to propose responses.

- **Communication and collaborations** around the topic of information diffusion with media and research partners.

Media Observatory is a breath-focused initiative, as reflected in the structure of this report. In the following sections, we describe various examples of computational methods, collaborative projects and reports that have been implemented on top of the core infrastructure. Beyond insights provided in this document, we believe that the significance of the platform is to empower future projects of this kind and to rapidly deploy research-driven methods to the public domain.

Newsteller.io

Intended audience

Results of research and development are accessible at a platform named NewsTeller[2]. The platform is intended for three types of audiences: (i) the general public, for which we provide access to news with improved context, (ii) media practitioners, with tools to monitor reactions and trends in real-time, and (iii) researchers, with historical access to news data.

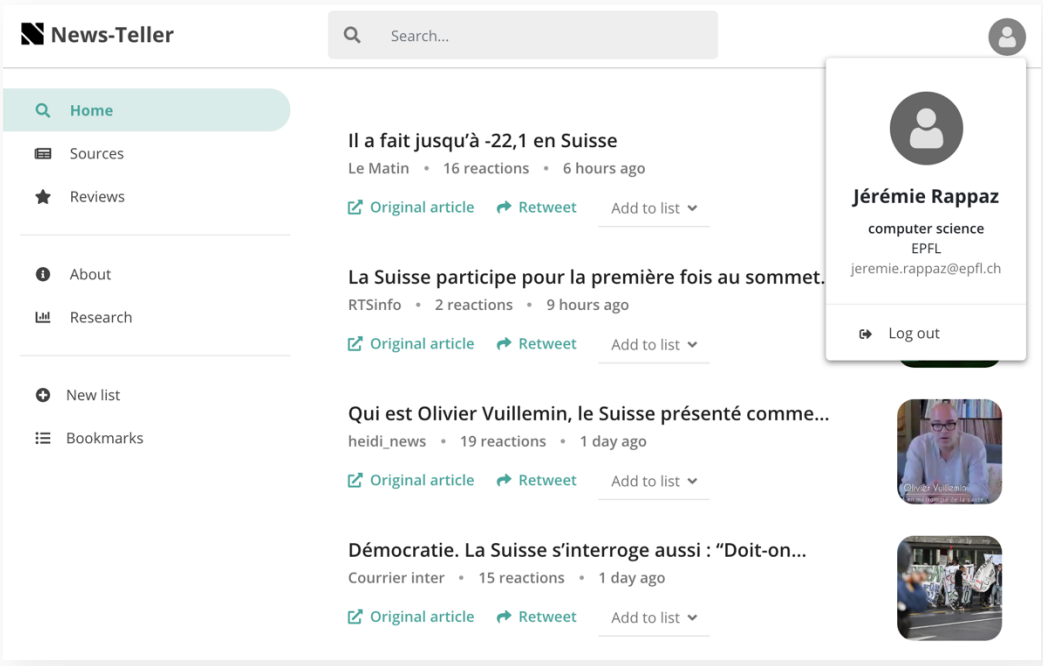

Figure 1: Front page of newsteller.io showing the result of a search with the keyword "Suisse".

[2] https://newsteller.io

Platform features

By taking obstacles mentioned above (O1, O2) into consideration, search mechanisms have been implemented on the platform, in order to facilitate access to articles outside Social Media. Some of the features deployed include[3]:

- **Article search engine** using keywords. Searches are performed on real-time data.
- **Filtering** results by language, news source. Filtering methods using categories will soon be available (sport, politics, etc).
- **Sorting results of a search** by recency, or by a combination of recency and volume of reactions (e.g. "hot" sorting).
- **Comparative news source analysis** by showing the 5 news sources with the most similar editorial lines using the method described in section *Research*, or by showing named entities frequently covered by the source.
- **Account management** allows users to create an account and to store lists of articles. These lists will be helpful in the future for experts to review and share emerging news subjects.

The user interface has been designed with modern Web technologies, with the objective to reduce information overload for the user and to ease the access to topical trends.

Data storage

News articles are currently stored in two different ways. The short-term storage gives public access to articles for a period of three months after their publications. During this period, articles are indexed into a publicly available high-performance search engine. Passed this period, articles are transferred to long-term storage, in order to be made available to the research community. Historical data are currently only available upon request, but future development aims to provide an API access.

Ethical Engagement and Security

Our initiative takes strong engagement on privacy, security and tracking. As such, the Web platform is tracker-free and do not make use of third-party analytics services. Data extracted from Social Media are anonymized before storage by using strong hashing methods. All data are stored and processed in Switzerland and the infrastructure is secured behind strong authentication mechanisms.

[3] Features are described as of the writing of this document (Nov. 2020) and are likely to evolve in future releases.

Information Extraction

Storing, indexing and enriching information about the online news landscape represents one of the core objectives of the Media Observatory. The link between news articles and audience reactions is generally difficult to establish. Therefore, indexing news and reactions data in a single database represents an important achievement of the initiative. Section *Reporting* proposes examples of reporting that can be extracted from the platform and suggest the importance of such a database for future media research. We briefly summarize our methods of data-processing in this section.

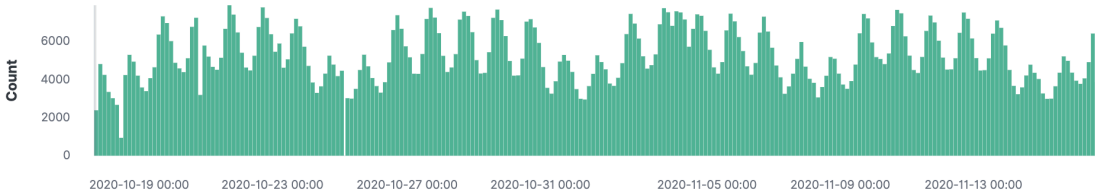

Figure 2: News articles collected and processed by the data pipeline over one month. In this signal, one could observe the circadian rhythms, as well as the low publishing rate over weekends.

Data Collection

The discovery of news article is a critical step for any news-related analysis. It consists in detecting newly published articles and extracting their respective URLs. In order to select a suitable discovery method, we need to consider several dimensions. The method should allow to discover articles quickly after publication, to contain reaction signals from the audience, and possibly to contains meta-information on the topic of the article. Existing lines of research generally make use of third-party data services in order to compare news sources, i.e. GDELT[4], that conveniently provide metadata for tracking topical issues across different media [2,3]. One issue with this approach is the poor coverage of non-American sources. For example, we report Swiss sources to be almost absent from GDELT, which limits its applicability to our scenario. As an alternative, our discovery mechanism currently uses Social Media connectors (Twitter), as it enables the tracking of large amounts of news channels and to gather links at publication time[5]. To circumvent the absence of additional information, *ad hoc* methods have been

[4] https://www.gdeltproject.org/

[5] We are aware of the non-exhaustivity of the data collection process: The Observatory currently focuses on article published through Social Media. Future releases will increase the coverage of our analysis with other discovery mechanisms.

developed to generate meta-data in quasi-real-time. The processing steps at article-level are briefly described below. The processing of source-level indicators is described in section *research*.

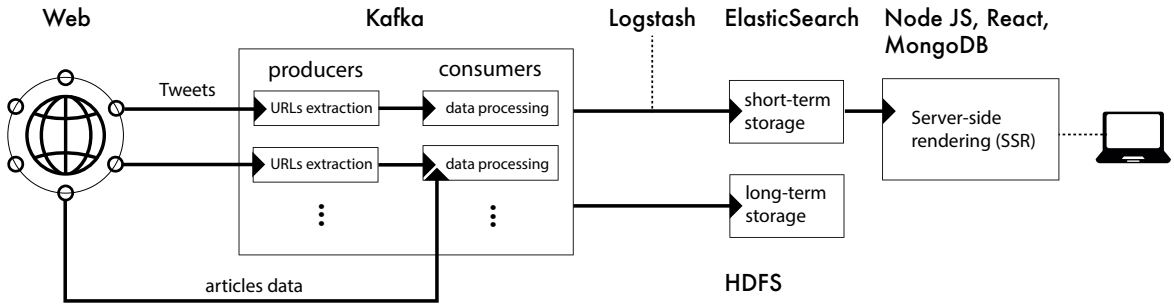

Figure 3: High level representation of Newsteller data pipeline that extracts, processes, indexes and serves news articles.

- **Text extraction** is a costly operation that downloads the title and the text of an article given its URL. The pipeline implements a multi-process approach in order to guarantee to process the required throughput of around1.2 articles per second.
- **Text processing** is mostly used to clean HTML tags and extract references, e.g. external links present in the text of the article.
- **Entity extraction** is the process of extracting *named entities* from text, for example names, organizations or locations.
- **News classification** allow to automatically infer the category of an article, for example sport, technology or politics.
- **Reactions tracking** and is given by the Twitter streaming API. The pipeline processes in average around 25 reactions per second.

The complete data pipeline is show in *Figure 3*. Statistics on the volume of articles collected every month is shown in *Table 1*.

	Value	Description
Media	913	Media tracked by the platform
Articles	1.3M / month	Number of collected articles
Retweets	28M / month	Number of retweets (Twitter)
Replies	16M / month	Number of replies (Twitter)
Quotes	9M / month	Number of quotes (Twitter)
Favorites	121M / month	Number of favorites (Twitter)
References	6M / month	Number of references in the article

Table 1: Data collection statistics for one month

Reporting

In this section, we provide analyses of the news landscape in order to demonstrate the scope of possibilities offered by the Media Observatory. We first give an example of a macro-level analysis, by comparing news sources on their tendency to provoke reactions, and of a micro-level analysis, by analyzing the coverage of chosen sources on a given topic.

Macro-level analysis

As described in section *Context*, new forms of online media can be seen emerging and are most frequently referred to as alternative media. Investigative work, supported by the expertise of our media partner *Le Temps*, led to identify and analyze characteristics that distinguish them from traditional news sources. Alternative media outlets are known for their use of « sensationalism », which provokes public interest or excitement, at the expense of accuracy. This last statement is observable through the lens of our database. *Figure 4* shows the difference in the volume of reactions for both traditional and alternative media.

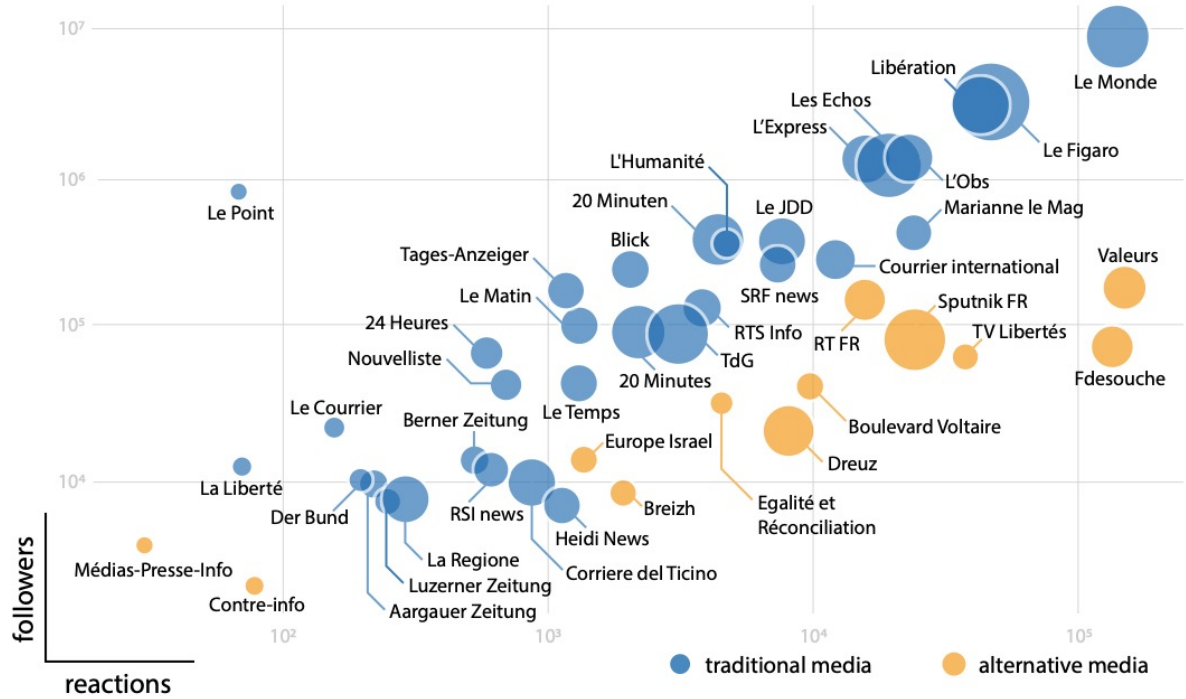

Figure 4: Relationship between audience size (followers) and volume of reactions (RT, replies, likes) for various francophone media. Alternative media are show in orange and exhibit significantly stronger reaction patterns. Media are classified as "alternative" if they appear in the list [Wikipedia: réinformation[6]], except for RT and Sputnik that have been described as propaganda outlets by multiple sources[7][8].

[6] https://fr.wikipedia.org/wiki/R%C3%A9information
[7] https://en.wikipedia.org/wiki/RT_(TV_network)
[8] https://en.wikipedia.org/wiki/Sputnik_(news_agency)

In order to analyze media on comparable scales, we investigate the volume of reactions by taking into account the number of their followers (Twitter). One could observe the log-linear relationship among traditional channels. However, alternative media differ from the observed trend and provoke, on average, larger volumes of reactions than their traditional counterparts. Content analysis would be needed to confirm the presence of sensationalism, but our findings clearly show the higher-than-usual engagement from alternative media communities.

Micro-level analysis

The landscaping possibilities offered by the Media Observatory is also useful for investigating information diffusion at micro-level, e.g. articles addressing a specific topic, published by specific channels. We exemplify such process through the mediatic coverage analysis of a selected topic: the case of Darius Rochebin. As shown in *Figure 5*, this topic has extensively been covered by the French-speaking press while only few articles have been published in German-speaking cantons. Similarly, those articles have provoked significant volumes of reactions from French-speaking cantons while being almost completely absent from German-speaking Social Networks. This example shows the possibilities for investigating mediatic attention from various outlets or regions. This is especially important for policy makers and planners who need to understand public concerns before taking decisions on specific topics.

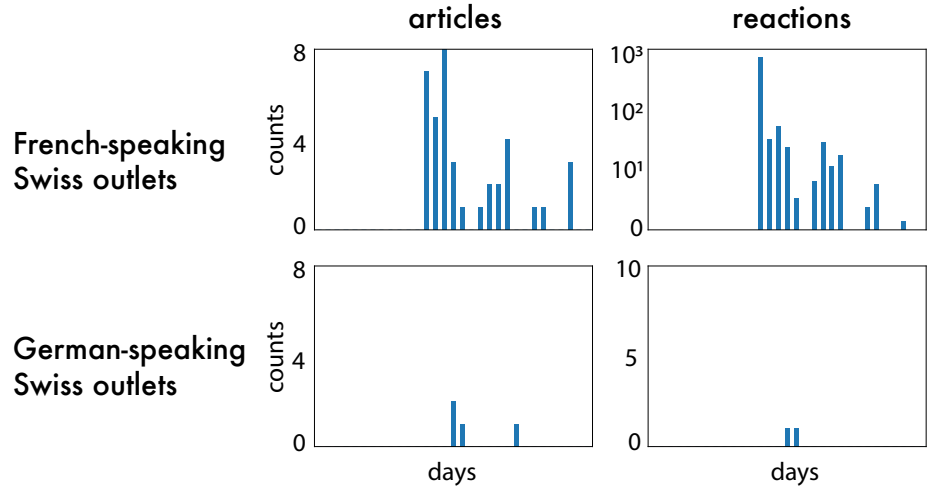

Figure 5 : publication and reaction patterns for French-speaking and German-speaking Swiss outlets for articles about Darius Rochebin[9]. French-speaking outlets are Le Temps, Le Matin, 20 Minutes, Heidi.news, 24 Heures, Tribune de Genève. German-speaking outlets are 20 Minuten, Aargauer Zeitung, NZZ, Der Bund, Tages-Anzeiger, Berner Zeitung.

[9] https://www.letemps.ch/suisse/rts-darius-rochebin-loi-silence

The infrastructure currently enables *ah hoc* visualization of specific themes by processing historical data. Future developments could make this option available for Web users.

Research

Beyond its data collection capabilities, the Media Observatory aims to develop news computational methods to analyze the news landscape. In this section, we provide a high-level description of the method developed to captures differences in the publishing patterns at semantic and thematic level.

As stated in section *Information Extraction*, the absence of metadata is a barrier to compare the respective selection of subjects of news channels. For example, GDELT provides a unique identifier to *news events*[10] that enables the tracking of information across channels. To circumvent the absence of information about events, our research efforts have focused on alternative methods to detect variance across different media.

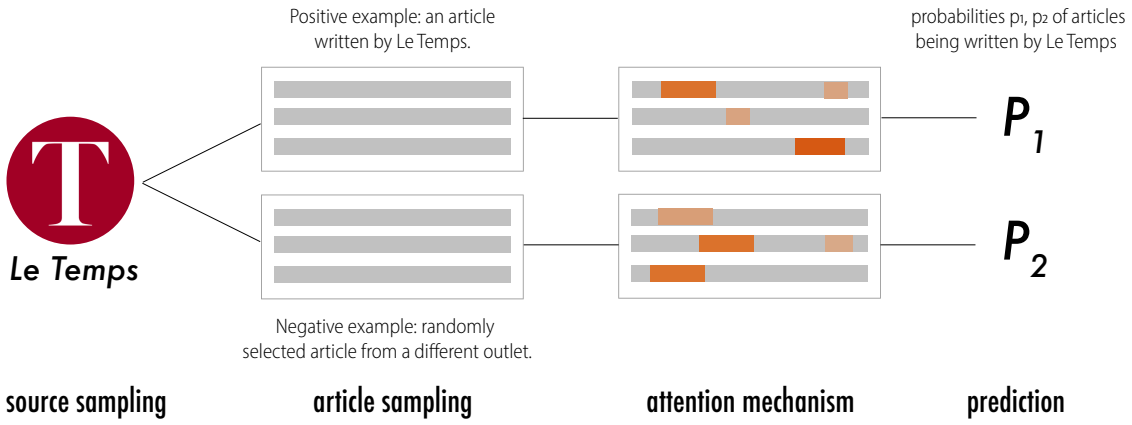

Figure 6: Illustration of the learning process of our method. The model learns to predict the probability of an article being written by a specific source.

The choice of a method is guided by consideration of performance, interpretability and type of data available. First, the model should be able to identify writing patterns that make a media distinct from its peers. Second, the model should be interpretable to avoid the presence of spurious relationships (e.g. the presence of the media name in the text) and the results to be validated by human experts.

[10] GDELT is "an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source".
Source: https://www.gdeltproject.org/about.html

Last, the model should be able to infer patterns from textual data only.

Machine learning methods.

Recent machine learning approaches fulfill the aforementioned requirements: the selected model is a self-attentive recurrent neural network that learns to identify the source of an article from its text. The model is trained using a discriminative learning procedure: a positive example, consisting of a media and an article it published is compared with a randomly sampled article, published from a different source. The model learns, in a pairwise fashion, to maximize the probability of the positive media-article pair, while minimizing the probability of the negative example (see *figure 6*). The additional attention mechanism allows the model to focus on specific portions of the text. This attention mechanism can be further leveraged in order to detect highly discriminative parts of the text, e.g. portions that are predictive of its source.

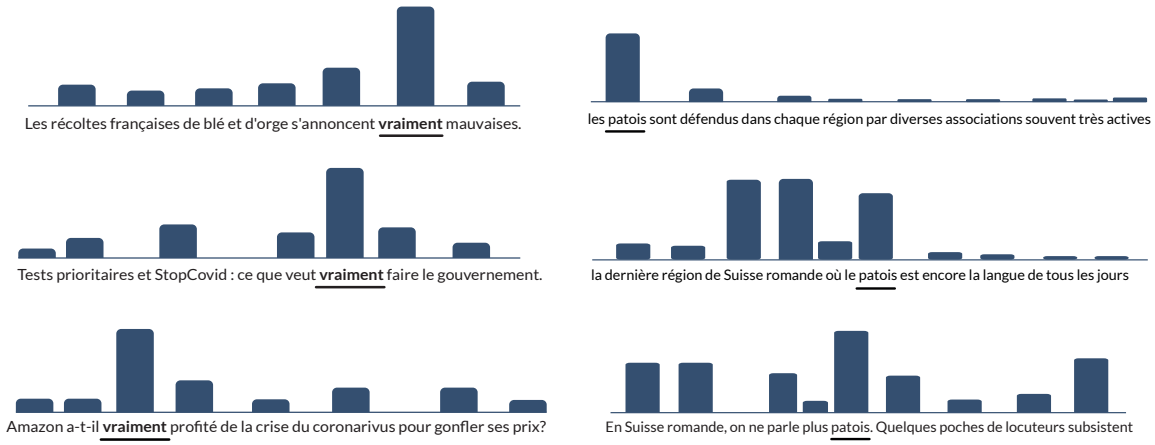

Figure 7 : Examples of the model's attention patterns on selected examples. On the left, the model focuses its attention on a syntactic pattern of BMFTV, identifying a frequent use of the term "vraiment". On the right, the model focuses its attention on a topic frequently addressed by Le Temps: Swiss dialects.

Attention weights, as it can be seen in *figure 7*, are informative of syntactic and thematic patterns that define the writing style and the editorial line of a media. If this is, by no means, a way to fully characterize a news source, this model represents a promising direction for the quick identification of patterns helping experts identifying potential biases[11] in the media landscape.

Visualizing sources

In this model of the media landscape, news channels are symbolized by a numerical representation. This representation is could be visualized using projection methods[12] that show the publishing

[11] Here we adopt the term bias in its technical sense. As such, the term *bias* qualifies any disproportionate use of specific themes and terms in comparison to other news sources.
[12] Projected using the dimensionality reduction method t-SNE.

characteristics of channels learned by the method. The proximity of a source with its peers is generally informative of the editorial line of the media. For example, the projection shown in *Figure 8* captures factors such as geographic proximity: all channels from a specific world region are grouped together, since the model identifies their propensity to discuss local news events. We report this representation to also be predictive of media ownership (for example in the case of a broadcast network) and even of political alignment[13].

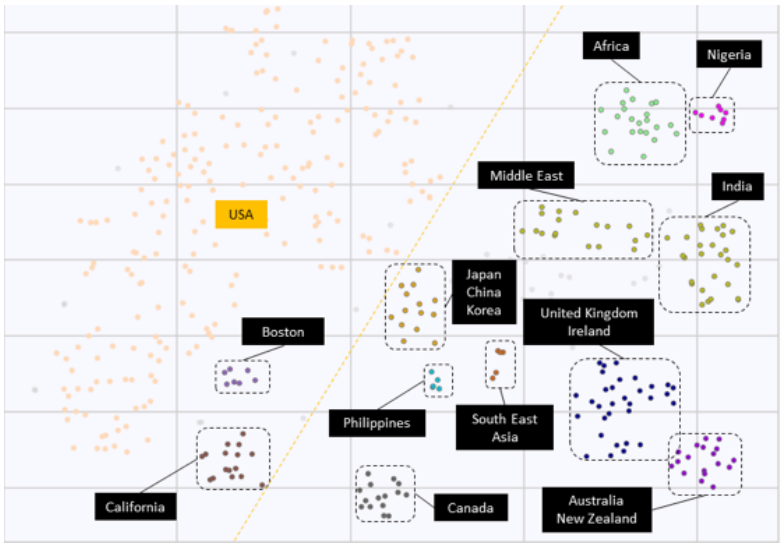

Figure 8 : 2D projection of the numerical representation of news channels by the model.

The final conclusions on this model, including classification performances, as well as the code for reproducing experiments, will be made publicly available after the publication of our results in a scientific venue. All code will be made available in our repository[14].

Communication

Our initiative benefited from a partnership with a Swiss newspaper, *Le Temps*, that provided expertise all along the project. One important objective of this partnership is the research of novel ways to communicate around the topic of information diffusion.

Media collaboration

During early stages of the COVID-19 pandemic, we investigated the publication rate at which articles about the virus were published. We correlated this signal with the number of Tweets and Google

[13] On US sources, predicting the bias as extracted from https://www.allsides.com/media-bias/media-bias-ratings
[14] https://github.com/News-Teller

searches mentioning the pandemic. The collaboration between engineers and journalists enabled the publication of a data-supported article containing interactive visualizations. Compared to standard data analyses, this type of collaborations enables the contextualization of data by journalists in the text of the article.

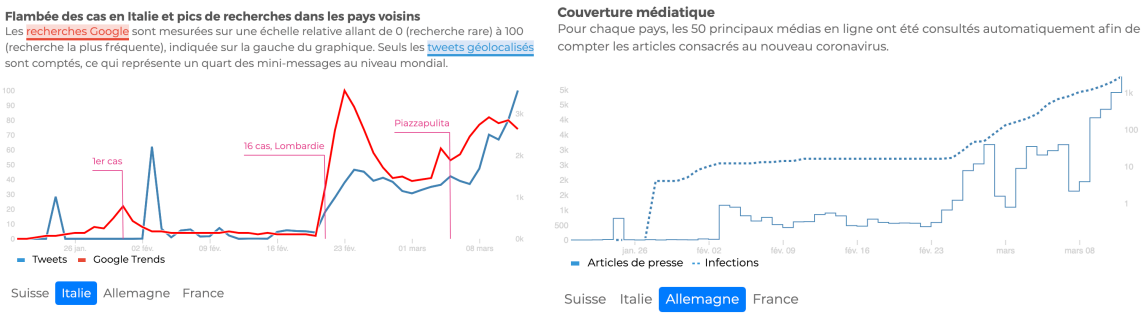

Figure 9: Number of press articles, Tweets and Google searches during the early stages of the COVID-19 pandemic. Extracted from "Covid-19: histoire d'une médiatisation", Le Temps[15].

The resulting article exemplifies short-term collaborative projects empowered by the infrastructure of the Media Observatory. This project has been well received by the public and has been cited in a broader study on the mediatization of COVID-19, conducted by *L'institut National de l'Audiovisuel* (INA)[16].

Additionally, we release an open-source analysis toolkit for evaluating polarization on Twitter, that reflects the investigative process developed during the project[17].

Collaborations

One of the goals of the initiative is to develop a worldwide network of researchers working on information diffusion and media-related problems. To this end, different events have been organized in order to build a research network in the international dimension.

Conferences and workshops

- *Mediate: Social and News Media Misinformation Workshop*[18] held at the AAAI International Conference on Web and Social Media (ICWSM)[19].

[15] https://labs.letemps.ch/interactive/2020/covid-trends/
[16] http://www.herve.name/pmwiki.php/Main/Etude-Coronavirus
[17] https://github.com/Fanfou02/twitter-retweets-analysis
[18] https://mediateworkshop.github.io/
[19] https://www.icwsm.org/2020/index.html

- *Narrative Framing and its Linguistic Forms in Online Media* held at the Communication in Multicultural Society Conference (CMSC)[20].

The platform was also introduced to the research community in a talk given at the *Conference for Truth and Trust Online* 2020[21] (UK). Finally, we co-organize an interest group hosted by the *Alan Turing Institute* (UK) [22] called *Media in the Digital Age* whose goal is to bring together researchers and media practitioners.

Societal Impact

Democracy requires vigilance and vigilance requires access to information. However, news information is increasingly brought to readers by large technology companies that (i) have substantial portions of their revenues coming from user tracking and (ii) can alter the access to information at their discretion. This environment is not a fair and transparent way to access information. Our initiative represents a unique experimenting field in which to develop methods to index, sort and diffuse news content exclusively with informational objectives: the platform developed by the Media Observatory Initiative allows searching in the news stream in real-time with clear filters and without the use of personalization methods. Moreover, the close collaboration with news institutions makes the platform a valuable tool to promote news literacy and information accountability.

Organization

Media Observatory is a collaboration between EPFL and the newspaper *Le Temps*.

Prof. Karl Aberer	Full professor - EPFL
Jérémie Rappaz	Doctoral researcher - EPFL
Panayiotis Smeros	Doctoral researcher - EPFL
Marco Romanelli	Engineer – EPFL
François Quellec	M.S. Student - EPFL
Gaël Hürlimann	Chief editor - *Le Temps*
Catherine Frammery	Journalist - *Le Temps*
Paul Ronga	Journalist - *Le Temps*

[20] https://cmsc2020.com/
[21] https://truthandtrustonline.com/2020-proceedings/
[22] https://www.turing.ac.uk/

The project has received funding from the Initiative for Media Innovation (IMI) based at Media Center, EPFL, Lausanne, Switzerland and from the Federal Office of Communications (OFCOM). The platform has also received support for the project SciLens [6, 7], from the Open Science fund at EPFL and from the Swiss Academy of Engineering Sciences (SATW) with the participation of OFCOM.

Future Opportunities

Computational journalism is an emerging research field of high potential impact since it offers a framework in which to reason about news data. The field is still in its infancy but an increasing community of researchers tackles problems such as media bias, fact checking and political discourse framing. Our project represents a unique opportunity to build tools and resources for supporting future research initiatives in this field. Furthermore, it enables a rapid deployment of future research in the public domain through the Newsteller platform.

The project has received further support from the Initiative for Media Innovation (IMI) for the project "Media Laboratory" with two new objectives: (i) developing novel data-supported article formats in partnership with the Academy of Journalism and Media (University of Neuchâtel). (ii) Automating experts' interventions on trending news topics.

Conclusion

The Media Observatory has successfully deployed a large-scale data collection infrastructure, developed new methods to analyze online news media and has been made available to the public. In parallel, the initiative has developed novel approaches to communicate around information diffusion with a media partner and has developed an international research network around media related topics. The Initiative will continue to give access to news data for non-profit institutions and to bring its reporting and landscaping capabilities to the public.

References

[1] Akerlof, George A. "The market for “lemons”: Quality uncertainty and the market mechanism." *Uncertainty in economics*. Academic Press, 1978. 235-251.

[2] Rappaz, Jérémie, Dylan Bourgeois, and Karl Aberer. "A Dynamic Embedding Model of the Media Landscape." *The World Wide Web Conference*. 2019.

[3] Bourgeois, Dylan, Jérémie Rappaz, and Karl Aberer. "Selection bias in news coverage: learning it, fighting it." *Companion Proceedings of the The Web Conference 2018*. 2018.

[4] The digital transformation of news media and the rise of disinformation and fake news. European Commission, 2018.

[5] Newman, Nic, et al. "Digital news report 2020." *Reuter Institute for the Study of Journalism. Recuperado de: https://bit. ly/2BhczTN* (2020).

[6] Romanou, Angelika, et al. "Scilens news platform: a system for real-time evaluation of news articles." *arXiv preprint arXiv:2008.12039* (2020).

[7] Smeros, Panayiotis, Carlos Castillo, and Karl Aberer. "SciLens: evaluating the quality of scientific news articles using social media and scientific literature indicators." *The World Wide Web Conference*. 2019.