Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

The Suisse Romande Local News Dataset

Victor Bros[1, 2], **Daniel Gatica-Perez**[1, 2]

[1]Idiap Research Institute
[2]EPFL
victor.bros@idiap.ch, daniel.gatica-perez@idiap.ch

Abstract

This paper introduces a comprehensive dataset of news articles sourced from ESH Médias, a prominent local press agency in Romandy, the French-speaking region of Switzerland. The dataset encompasses all articles published on their digital platforms from January 2015 through June 2022. With over 130,000 articles written in French, this dataset offers a rich insight into local news from the French-speaking cantons of Switzerland. The articles cover a diverse range of topics and provide valuable material for Natural Language Processing and media studies. To respect privacy and legal considerations, journalists' names have been anonymized, and the dataset is made available for research purposes under a specific agreement with ESH Médias. The dataset adheres to the FAIR principles, and a detailed datasheet is provided to facilitate its use. The dataset is accessible via a DOI link.

1 Introduction

Local press occupies a unique niche within the media ecosystem, reflecting the distinct regional interests and concerns of its readership. Historically, it has differentiated itself from national titles, which often represent the interests of dominant regions within countries (Hobbs 2018). This distinction has allowed local press to assume a significant societal role, acting as a key tie between communities.

However, the media landscape in Switzerland has also undergone significant transformation. The rise of free newspapers targeting supra-regional audiences and a shift toward advertising-centric business models have reconfigured the press ecosystem (BFS 2007). As a result, local news coverage has diminished in favor of broadly appealing content (Vogler, Weston, and Udris 2023; Salerno 2016). The resulting "news deserts" and substantial job cuts raise concerns about democratic engagement and the weakening of local journalism's role (Abernathy 2018; swissinfo.ch 2024, 2023; Pauline and Romy 2022). Recognizing the need to preserve local journalism, several initiatives aim to revitalize local news through accessible information and community-focused reporting (Gulyas and Hess 2024; Vos and Thomas 2024). However, these efforts take place within a complex political landscape. Politically, the tensions in Switzerland related to the support of national and local news are visible through its direct democracy system. On one hand, Swiss citizens voted in 2018 against an initiative that proposed to eliminate subsides to the national public broadcaster (France-Presse 2018). On the other hand, Swiss citizens voted in 2022 against a federal government proposal to financially support regional media, including newspapers, radio, and television (DETEC 2015). These political decisions reflect the ongoing debate about the role and funding of media in Switzerland, further complicating the efforts to sustain local journalism.

Recent upheavals, such as changes in readership habits and economic pressures on editorial newsrooms, have altered the media ecosystem's dynamics (Stites 2018). To adapt, local newspapers have transitioned to digital formats or modified their offerings to attract new readers (Nielsen 2015). These transformations potentially threaten the local press's capacity to address social cohesion and democratic practice challenges (Ballarini 2015).

Given its unique characteristics, local press serves as a key subject for understanding media ecosystem dynamics. To further such studies, we present a dataset of online local newspaper articles. The data comprises articles from the web platforms of three newspapers owned by the ESH Médias press group: *Le Nouvelliste* (Canton of Valais), *La Côte* (Canton of Vaud), and *Arc Info* (Canton of Neuchâtel). The articles, totaling 130,155 items, are stored in their original French language.

This substantial dataset is ideal for large-scale analyses of local press, particularly those conducting Natural Language Processing experiments.

1.1 Addressing a Gap in Existing Resources

While there are existing resources that cover Swiss news and legal texts, they often lack comprehensive coverage of local press articles from the French-speaking regions of Switzerland. For example, SwissBERT (Vamvas, Graen, and Sennrich 2023) is a multilingual language model trained on news articles from Swiss media, but it does not include sources from the local newspapers we consider in our dataset. Similarly, datasets used in studies such as (Vogler, Weston, and Udris 2023) focus on specific aspects like geographical diversity in Swiss news media but are limited in size (n=5,173 articles in that specific study) and may not

comprehensively represent the local press landscape. Other works, like (T.Y.S.S et al. 2024), compile datasets on legal judgments, which differ significantly from news articles.

Our dataset addresses this gap by providing a substantial collection of over 130,000 articles from local newspapers (*Le Nouvelliste*, *La Côte*, and *Arc Info*) in the French-speaking cantons of Switzerland. This resource offers researchers access to content that has not been extensively documented or analyzed in existing datasets. By focusing on local press, our dataset enables studies that can explore regional language use, local discourse, and media practices in a way that geographically broader datasets cannot.

This substantial dataset is ideal for large-scale analyses of local press, particularly those conducting Natural Language Processing experiments.

2 Dataset Description

2.1 Newspaper Context in Romandy

ESH Médias and Its Local Daily Newspapers. The dataset was obtained from the ESH Médias platform, which operates three French-language daily local newspapers in Romandy, Switzerland: *Le Nouvelliste* (Canton of Valais), *ArcInfo* (Canton of Neuchâtel), and *La Côte* (Canton of Vaud). Each of these newspapers focuses on covering news within its respective canton, covering national or international stories only when particularly relevant.

Despite the decline in certain segments of print media, these three outlets have maintained relatively stable readership until recent years. The daily circulation stands at approximately 45,000 printed copies for *Le Nouvelliste* (serving around 100,000 readers), about 32,000 copies for *ArcInfo* (reaching an estimated 55,000 readers), and fewer than 10,000 copies for *La Côte* (precise readership figures are unavailable) (OFS 2023). Each newspaper is published every day and is owned by ESH Médias, a private regional press group.

Information gleaned from conversations with editorial staff highlights subtle differences in scale and editorial emphasis among these newspapers. For instance, *Le Nouvelliste* employs approximately 50 permanent journalists, compared to 35 at *ArcInfo* and 15 at *La Côte* (at the time of the discussion), suggesting variations in the volume of local reportage. Across all three, journalists seldom contribute to more than one newspaper. Following the pandemic, editors emphasized the need to fortify their local focus further, offering news content specifically tailored to their readerships' unique social, cultural, and political contexts. Roughly 95% of online articles also appear in print, and minimal automation is used (primarily for article recommendation on each publication's website).

Representativeness in Romandy. Romandy's media ecosystem comprises both local daily newspapers such as those owned by ESH Médias and other publications with broader coverage (e.g. *24 heures*, *La Tribune de Genève*, and *Le Temps*). In the specific cantons of Valais, Neuchâtel, and Vaud, however, *Le Nouvelliste*, *ArcInfo*, and *La Côte* are considered staple daily outlets with substantial local reach. As such, while our dataset does not represent every possible news source in French-speaking Switzerland, it does capture a significant segment of daily, local-level reporting in three major cantons. This affords researchers a meaningful window into regional journalistic practices, albeit one that should be supplemented with larger-scale Romandy or national datasets for studies requiring broader geographic coverage.

Articles published			
Le Nouvelliste	La Côte	Arc Info	Total
43,393	38,283	48,479	130,155
(21,581)	(12,618)	(22,830)	(83,243)

Table 1: Number of articles collected for each newspaper. The number of unique articles is in parentheses. The final total of unique articles does not equal the sum of unique articles from each source because duplicates are not counted and only the first occurrence is retained.

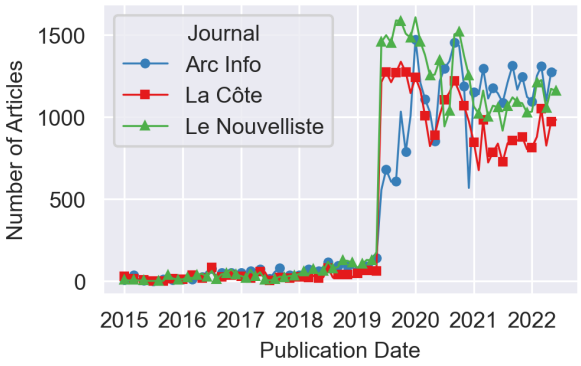

Figure 1: Monthly distribution of articles from January 2015 to June 2022 for each newspaper.

2.2 Statistical Description

In the process of data collection, articles were retrieved from the websites of three newspapers, namely *lenouvelliste.ch*, *lacote.ch*, and *arcinfo.ch*, spanning from January 2015 to June 2022. The total number of articles collected from each source is detailed in Table 1.

The numbers in Table 1 reflect the count of unique articles to give an overview of potential information overlap. However, **all articles, including duplicates, remain in the final dataset**.

In addition to the overall counts, Figure 1 shows the monthly distribution of articles for each newspaper in our dataset from January 2015 through June 2022. This visualization highlights how publication rates evolved over time and reveals a noticeable change in data volume and consistency around mid-2019, probably coinciding with the newsroom's more systematic approach to digital publication.

The articles are stored in JSON files, organized hierarchically by newspaper, year, and month of publication. Within

each monthly directory, individual JSON files contain all articles published on a given day.

Each article is assigned a unique ID, structured as follows: *source_year_month_day_index*.

Each article is stored as a JSON object, including fields for the title, headline, and content of the article. To respect privacy, the names of the journalists have been anonymized and replaced with unique identifiers. Additional fields provide information on the publication date, journalist-annotated tags, and the presence of any accompanying illustrations. All articles are stored in French, their original language.

The collected data encompasses all articles available on the respective web platforms, thus reflecting the diverse range of content published in the newspapers, from sports results to opinion pieces and human-interest stories.

3 Data Collection Process

In this section, we briefly explain the process of data collection and preparation. The articles were collected by scraping the web platforms of three newspapers, utilizing the Python library BeautifulSoup (Richardson 2007).

For each article, we extracted the title, headline, and content. To ensure the quality of the data, we implemented a script to remove HTML tags from the text of the articles. This step was important to eliminate any potential noise that could interfere with further analyses.

As part of the data preparation, the names of the journalists were anonymized to protect their privacy. Each journalist's name was replaced with a unique identifier, and the mapping between names and identifiers is not publicly available. This step adheres to privacy and ethical standards.

It should be noted that additional data cleaning procedures may be necessary depending on the specific requirements of the Natural Language Processing analysis to be conducted. This could include, but is not limited to, the correction of typos, the removal of stop words or boilerplates, stemming, and lemmatization.

4 FAIR Compliance

The dataset has been developed in accordance with the FAIR principles (Wilkinson et al. 2016) to ensure that it is Findable, Accessible, Interoperable, and Reusable:

- **Findable** The dataset is assigned a persistent DOI (https://doi.org/10.34777/6w61-tn46), making it easily findable. Metadata describing the dataset are provided to enhance discoverability.
- **Accessible** The dataset is hosted on a data-sharing platform that allows for authorized access to researchers, ensuring that it can be retrieved using standard protocols : (https://www.idiap.ch/en/scientific-research/data/suisse-romande-news-datasetthat).
- **Interoperable** The data is stored in standard JSON format, which is widely used and compatible with common data analysis tools. This promotes interoperability with various software applications.
- **Reusable** Comprehensive metadata and documentation are provided to facilitate understanding and reuse of the dataset in different research contexts. The dataset is shared under specific terms that allow for reuse in academic research.

5 Datasheet for the Dataset

Motivation The dataset was created to provide a comprehensive collection of local news articles from French-speaking Switzerland, facilitating research in NLP, media studies, and social sciences.

Composition The dataset contains over 130,000 articles from three newspapers (*Le Nouvelliste*, *La Côte*, and *Arc Info*) between January 2015 and June 2022. The articles are in French and cover a wide range of topics including local news, sports, politics, culture, and opinion pieces.

Collection Process Articles were collected by scraping the newspapers' websites using the Python library BeautifulSoup. Automated scripts navigated the websites' archives to retrieve article content, titles, headlines, publication dates, author information, and tags.

Preprocessing Journalists' names were anonymized by replacing them with unique identifiers to protect privacy. Duplicate articles were identified based on content similarity, and only the first occurrence was retained. HTML tags and other non-textual elements were removed from the article content.

Uses The dataset can be used for NLP tasks such as language modeling, topic modeling, sentiment analysis, and studies on media content and discourse analysis. It is also suitable for investigating trends in local news reporting and the evolution of regional media landscapes.

Distribution The dataset is available under specific terms for research purposes. Access is provided via the DOI: (https://doi.org/10.34777/6w61-tn46).

Maintenance Any updates or corrections to the dataset will be documented and made available through the same DOI. Users are encouraged to report any issues or errors they encounter.

6 Legal Considerations

The dataset has been shared with the permission of ESH Médias, who is the copyright owner of the news articles. ESH Médias has granted rights to use the articles for research purposes. To adhere to privacy and ethical standards, the journalists' names have been anonymized by replacing them with unique identifiers. The mapping between names and identifiers is not public.

Users of the dataset are required to agree to the terms of use, which specify that the dataset is to be used solely for academic research and not for any commercial applications. Redistribution of the dataset or any portion of it is not permitted without prior consent from ESH Médias.

7 Potential Applications and Usage

The dataset presented in this report is designed for hierarchical and efficient data storage. However, it can be manipulated for quick access to facilitate more comprehensive analyses of all the articles.

These articles can be utilized in statistical analyses or experiments employing Natural Language Processing techniques to study various phenomena concerning the local French-speaking Swiss press. Potential applications include:

- **Language Modeling**: Training language models on the French language, with a focus on local dialects and terminology.
- **Sentiment Analysis**: Analyzing public sentiment on local issues as reflected in news articles.
- **Topic Modeling**: Identifying prevalent themes and topics in regional news coverage over time.
- **Media Studies**: Examining news reporting patterns, bias, and the evolution of media narratives in local press.

A significant observation to highlight is the distinct difference in both the number of articles and the accuracy of their collected content available before and after June 2019. This difference is likely due to a more systematic approach to the digitization of articles implemented from this date onwards by the newsroom. For experiments necessitating high-quality data, we recommend limiting the timeframe to the period between June 2019 and June 2022.

Generalizability. Although the three newspapers in our dataset encompass a major fraction of daily local journalism in the cantons of Valais, Neuchâtel, and Vaud, they are not exhaustive. Researchers seeking to study other regions or broader Swiss coverage may need to supplement these data with outlets from other cantons or national-level newspapers. Nevertheless, because each newspaper is a principal source of local reporting for its readership, our dataset offers a robust representation of local press dynamics within these three cantons, and thus constitutes a unique resource for long-term studies of Swiss local media.

8 Conclusion

In this paper, we introduced a dataset of news articles, encompassing popular local titles (*Le Nouvelliste*, *La Côte*, and *Arc Info*) in three French-speaking cantons of Switzerland. The data comprises over 130,000 articles, all written in French, providing comprehensive coverage of the three newspapers. This dataset has been assembled with the intention of facilitating computational analyses, specifically in the context of local press. By making this dataset available, we aim to support a wide range of research activities in natural language processing and media studies.

Acknowledgments

We thank ESH Médias (especially Isabelle Segarini) for providing access to their news data. This work was supported by the AI4Media project (Grant 951911, H2020 Programme ICT-48-2020) and the ELIAS project (Grant 101120237, Horizon Europe Programme), funded by the European Commission.

References

Abernathy, P. M. 2018. *The Expanding News Desert*. The Center for Innovation and Sustainability in Local Media, School of Media and Journalism, University of North Carolina at Chapel Hill.

Ballarini, L. 2015. Julia Cagé, Sauver les médias. Capitalisme, financement pariticipatif et démocratie. *Questions de communication*, 28(2): 350–352. 010.

BFS. 2007. *La diversité de la presse en Suisse*. Bundesamt für Statistik. ISBN 978-3-303-16084-8.

DETEC. 2015. Paquet médias - DETEC. https://www.uvek.admin.ch/uvek/fr/home/detec/votations/paquet-medias.html. Accessed: 2025-04-23.

France-Presse, A. 2018. Switzerland votes overwhelmingly to keep its public broadcaster.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gulyas, A.; and Hess, K. 2024. The Three "Cs" of Digital Local Journalism: Community, Commitment and Continuity. *Digital Journalism*.

Hobbs, A. 2018. *A Fleet Street In Every Town: The Provincial Press in England, 1855-1900*. Open Book Publishers. 04.

Nielsen, R. K., ed. 2015. *Local journalism: the decline of newspapers and the rise of digital media*. I.B. Tauris & Co. Ltd in association with the Reuters Institute for the Study of Journalism, University of Oxford. 02.

OFS. 2023. La diversité de la presse en Suisse. https://www.bfs.admin.ch/asset/fr/343686. Accessed: 2025-04-23.

Pauline, T.; and Romy, K. 2022. Swiss media workforce keeps shrinking.

Richardson, L. 2007. Beautiful soup documentation. *April*.

Salerno, S. 2016. Media challenges and strategies in Romandie and Ticino. *Studies in Communication Sciences*, 16(1): 2–7.

Stites, T. 2018. About 1,300 U.S. communities have totally lost news coverage, UNC news desert study finds. *Poynter*. 03.

swissinfo.ch. 2023. Swiss newspaper group Tamedia announces job cuts. https://www.swissinfo.ch/eng/business/swiss-newspaper-group-tamedia-announces-job-cuts/48830520. Accessed: 2025-04-23.

swissinfo.ch. 2024. Swiss publisher Tamedia makes massive job cuts and closes printing plants. https://www.swissinfo.ch/eng/workplace-switzerland/tamedia-makes-massive-job-cuts-and-closes-printing-plants/87443927. Accessed: 2025-04-23.

T.Y.S.S, S.; Baumgartner, N.; Sturmer, M.; Grabmair, M.; and Niklaus, J. 2024. Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset. In *International Conference on Language Resources and Evaluation*.

Vamvas, J.; Graen, J.; and Sennrich, R. 2023. SwissBERT: The Multilingual Language Model for Switzerland. *ArXiv*, abs/2303.13310.

Vogler, D.; Weston, M.; and Udris, L. 2023. Investigating News Deserts on the Content Level: Geographical Diversity in Swiss News Media. *Media and Communication*, 11.

Vos, T.; and Thomas, R. 2024. "They're Making It More Democratic": The Normative Construction of Participatory Journalism. *Digital Journalism*, 12(6): 869–893.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures?
   Yes, the dataset has been prepared with careful attention to privacy norms and ethical considerations, including anonymization of personal data as described in Section 6.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope?
   Yes, the abstract and introduction accurately describe the scope and contributions of the dataset as detailed in the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made?
   Yes, the methodology is detailed in Section 3, demonstrating its appropriateness for data collection and preparation.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions?
   Yes, potential biases and limitations, such as differences in article quantity before and after June 2019, are discussed in Section 7.
   (e) Did you describe the limitations of your work?
   Yes, limitations are discussed in Sections 2 and 7, including data quality variations over time.
   (f) Did you discuss any potential negative societal impacts of your work?
   Yes, potential impacts are considered, and measures to mitigate negative effects are outlined in the Legal Considerations sections.
   (g) Did you discuss any potential misuse of your work?
   Yes, potential misuse is addressed in the Legal Considerations section, and the terms of use restrict the dataset to research purposes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings?
   Yes, steps such as anonymization of personal data and controlled access to the dataset are described in Sections 3 and 6.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   Yes, we have reviewed the ethics guidelines and ensured compliance throughout the paper.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results?
   Not applicable, as the paper does not involve hypothesis testing.
   (b) Have you provided justifications for all theoretical results?
   Not applicable.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results?
   Not applicable.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study?
   Not applicable.
   (e) Did you address potential biases or limitations in your theoretical framework?
   Not applicable.
   (f) Have you related your theoretical results to the existing literature in social science?
   Not applicable.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain?
   Not applicable.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results?
   Not applicable, as the paper does not include theoretical proofs.
   (b) Did you include complete proofs of all theoretical results?
   Not applicable.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   Not applicable, as no machine learning experiments were conducted in this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   Not applicable.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   Not applicable.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
Not applicable.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made?
Not applicable.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance?
Not applicable.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators?
Yes, we have cited all sources, including the use of BeautifulSoup (Richardson 2007).

(b) Did you mention the license of the assets?
Yes, the dataset is made available under specific terms for research purposes, as described in the Legal Considerations section.

(c) Did you include any new assets in the supplemental material or as a URL?
Yes, the dataset is accessible via the DOI provided in the abstract and throughout the paper.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
Yes, consent was obtained from ESH Médias, the copyright owner, as described in Section 6.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
Yes, personally identifiable information such as journalists' names has been anonymized, and the dataset was explored semi-automatically by the researcher without encountering offensive content.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see Wilkinson et al. (2016))?
Yes, we discuss the FAIR compliance in Section 4.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))?
Yes, a datasheet is provided in Section 5 of the paper.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots?
Not applicable, as no crowdsourcing or human subjects research was conducted.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals?
Not applicable.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
Not applicable.

(d) Did you discuss how data is stored, shared, and deidentified?
Not applicable.