THE SUISSE ROMANDE LOCAL NEWS DATASET

Victor Bros Daniel Gatica-Perez

Idiap-Com-03-2023

NOVEMBER 2023

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

The Suisse Romande Local News Dataset

VICTOR BROS, Idiap Research Institute & EPFL, Switzerland
DANIEL GATICA-PEREZ, Idiap Research Institute & EPFL, Switzerland

This report introduces a comprehensive database of news articles sourced from ESH Médias, a prominent local press agency in Romandy, the French-speaking region of Switzerland. The database encompasses all articles published on their digital platforms from January 2015 through June 2022. Given the popularity of ESH Médias' titles, this database offers a rich and unique insight into local news from the French-speaking cantons of Switzerland. With a total of over 130 000 articles, this database presents a significant opportunity for extensive Natural Language Processing (NLP) analysis.

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 represent the interests of dominant regions within countries [3]. This distinction has allowed local press to uptake a significant societal role.

However, recent upheavals, such as changes in readership habits and economic pressures on editorial newsrooms, have altered the media ecosystem's dynamics [5]. To adapt, local newspapers have transitioned to digital formats or modified their offerings to attract new readers [1]. These transformations potentially threaten the local press's ability to address social cohesion and democracy practice challenges [2].

Given its unique characteristics, local press serves as an intriguing subject for understanding media ecosystem dynamics. To further such studies, we propose a database of digitized local newspaper articles. The database comprises articles from the web platforms of three newspapers owned by the ESH média press group: Le Nouvelliste (Canton of Valais), La Cote (Canton of Vaud), and Arc Info (Canton of Neuchâtel). The articles, totaling 130 155, 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.

2 DATABASE 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.

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

Table 1. The numbers of articles collected for each newspapers. The number of unique articles are in parenthesis. The final total of unique articles does not equal the sum of unique, because we kept the first occurrence of duplicates.

Authors' addresses: Victor Bros, victor.bros@idiap.ch, Idiap Research Institute & EPFL, Switzerland; Daniel Gatica-Perez, gatica@idiap.ch, Idiap Research Institute & EPFL, Switzerland.

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*.

The Json dictionary for each article includes fields for the title, headlines, and content of the article. Additional fields provide information on the authors, publication date, journalist-annotated tags, and any accompanying illustrations. All articles are stored in their original language, most if not all in French.

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.

3 DATA COLLECTION PROCESS

In this section, we briefly explain the process of data collection. The articles were collected by scraping the web platforms of three distinct newspapers, utilizing the Python library, BeautifulSoup [4].

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

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

4 POTENTIAL APPLICATIONS AND USAGE

The database 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.

A significant observation to highlight is the distinct difference in both the quantity and quality of articles available before and after June 2019. This difference is likely due to a more systematic approach to the digitalization 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.

5 CONCLUSION

In this report, we have introduced a newly compiled database of news articles, encompassing popular local titles (Le Nouvelliste, La Cote, and Arc Info) in three French-speaking cantons of Switzerland. The database comprises over 130 000 articles, all written in French, providing a comprehensive coverage of the three newspapers. This database has been assembled with the intention of conducting computational analyses, specifically for the context of local press.

6 ACKNOWLEDGMENT

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

REFERENCES

[1] Rasmus Kleis Nielsen (Ed.). [n. d.]. *Local journalism: the decline of newspapers and the rise of digital media* (London ; New York, 2015). I.B. Tauris & Co. Ltd in association with the Reuters Institute for the Study of Journalism, University of Oxford. 02.

[2] Loïc Ballarini. [n. d.]. Julia Cagé, Sauver les médias. Capitalisme, financement participatif et démocratie. 28, 2 ([n. d.]), 350–352. https://doi.org/10.4000/questionsdecommunication.10247 08.

[3] Andrew Hobbs. [n. d.]. *A Fleet Street In Every Town: The Provincial Press in England, 1855-1900.* Open Book Publishers. https://www.openbookpublishers.com/books/10.11647/obp.0152 04.

[4] Leonard Richardson. 2007. Beautiful soup documentation. *April* (2007).

[5] Tom Stites. [n. d.]. *About 1,300 U.S. communities have totally lost news coverage, UNC news desert study finds.* https://www.poynter.org/business-work/2018/about-1300-u-s-communities-have-totally-lost-news-coverage-unc-news-desert-study-finds/ 03.