Revisionista.PT: Uncovering the News Cycle Using Web Archives

Flávio Martins(✉) and André Mourão

NOVA LINCS, School of Science and Technology, Universidade NOVA de Lisboa,
Lisbon, Portugal
flavio.martins@fct.unl.pt, a.mourao@campus.fct.unl.pt

Abstract. In this demo, we present a meta-journalistic tool that reveals post-publication changes in articles of Portuguese online news media. *Revisionista.PT* can uncover the news cycle of online media, offering a glimpse into an otherwise unknown dynamic edit history. We leverage on article snapshots periodically collected by Web archives to reconstruct an approximate timeline of the changes: additions, edits, and corrections. *Revisionista.PT* is currently tracking changes in about 140,000 articles published by 12 selected news sources and has a user-friendly interface that will be familiar to users of version control systems. In addition, an *open source* browser extension can be installed by users so that they can be alerted of changes to articles they may be reading. Initial work on this demo was started as an entry submitted into *Arquivo.PT*'s 2019 Prize, where it received an award for second place.

1 Introduction

Nowadays, online media plays a critical role in the dissemination of the news. In the age of *online first* and the *24 h news cycle*, news publishing has been deeply transformed to allow more agility and flexibility, and (online) news articles are now updated seamlessly as new information becomes available. However, the desire to be first to publish and achieve high click-through rates has led to an increase in *clickbait*, both pre- and post-publication editorialization of titles and headlines that may misrepresent the actual content of the articles [1]. These trends may have contributed to the increased media mistrust that is discussed today under the controversial term *fake news*, described as “a poorly-defined and misleading term that conflates a variety of false information, from genuine error through to foreign interference in democratic processes” [2].

In some instances, undisclosed post-publication changes have swung the overall message and tone of news articles and, when revealed, have angered readers [4]. How can news readers be sure that the articles they are reading were not altered secretly after their original publication? Can we trust publishers to guarantee that a link you share with a friend will contain the same content you have read?

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 465–469, 2020.
https://doi.org/10.1007/978-3-030-45442-5_59

Revisionista.PT[1] allows readers to see all kinds of post-publication changes, and allows scholars to research this phenomenon and reason about how these changes affect the readers perception of news in the age of social media.

2 Proposed Solution

In this demo paper, we propose a meta-journalistic tool that reveals post-publication changes in articles published by Portuguese online news media. The goal of *Revisionista.PT* is to discover significant post-publication changes in news articles and to present them in a transparent auditable timeline. The most important raw materials to achieve this objective are historical collections of web page snapshots from news outlets that are preserved by the Portuguese Web Archive, *Arquivo.PT*[2].

To create this project, we had to build solutions to solve a number of technical challenges: *Given a web page, how to extract the textual content of the article?*, *What are the significant differences between article revisions?*, and, most importantly, *What dimensions can be used to classify the changes found?*.

In Fig. 1 we presents a simplified overview of the pipeline that we use to process a collection of news article URLs and generate the summaries of the changes that will be shown in the user interface. The pipeline can be described in three main steps: (1) fetching of HTML snapshots, (2) text extraction from HTML, and (3) changes extraction from text.

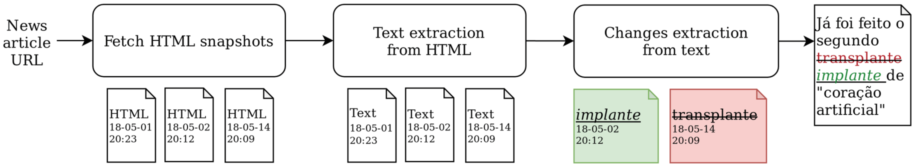

News article URL
Fetch HTML snapshots
Text extraction from HTML
Changes extraction from text
Já foi feito o segundo ~~transplante~~ *implante* de "coração artificial"
HTML 18-05-01 20:23
HTML 18-05-02 20:12
HTML 18-05-14 20:09
Text 18-05-01 20:23
Text 18-05-02 20:12
Text 18-05-14 20:09
implante 18-05-02 20:12
~~transplante~~ 18-05-14 20:09

Fig. 1. From a news article URL to a set of edits

News Article URLs. To build a list of news article URLs our strategy was to develop URL *slug* templates that are matched when fetching news articles of a given news source (e. g., `https://observador.pt/<YYYY>/<MM>/<DD>/<title>`. We used a semi-automatic interactive process to create the *slug* templates that we then used to whitelist URLs that will be fetched from each news source. The Portuguese Web Archive (*Arquivo.PT*) is currently the main data source. Using an established Web archive allowed us to focus on tasks that are orthogonal to the preservation of the web and leverage on their preserved web page snapshots.

[1] https://revisionista.pt.
[2] https://arquivo.pt.

Fetch HTML Snapshots. Given a list of news article URLs, our crawler downloads multiple snapshots for each of the URLs in the list. We adopted a more conservative crawling strategy that can be described as a bisection method over the largest time span (furthest pair of snapshots), with downloads proceeding only on the branches where changes are found.

Text Extraction from HTML. At this stage, the goal is to extract the textual content and other interesting metadata such as title, author names and publication date, that we want to consider for version control. News web pages contain a myriad of extra content, such as sidebars with breaking news, and comment sections that are undesirable for our purposes and not to be included.

To guarantee a high-fidelity at this stage, we opted to build site-specific specialised content extractors for each target news source. The extractors essentially select the main HTML element containing the article (e. g., <div class="article">) and extracts all the text inside all the inner HTML elements except from known site-specific undesired elements identified semi-automatically.

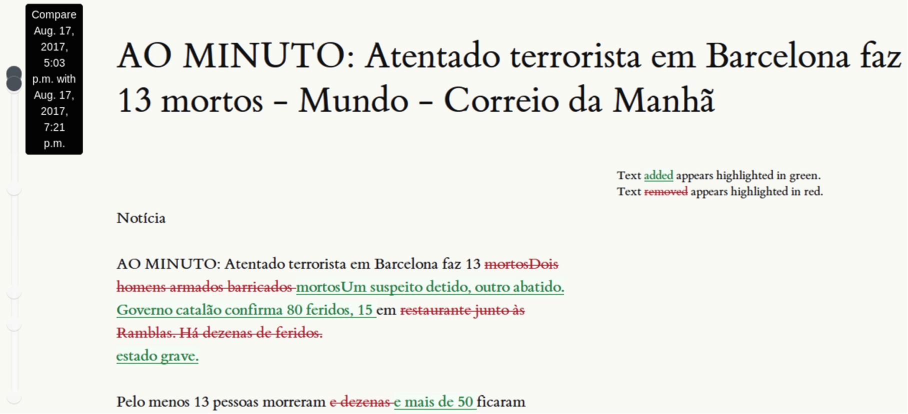
In addition, we normalised relative time expressions such as, *Hoje, 20:21* (Today, 20:21) into a standard format (27-01-2018, 20:21:00).

Changes Extraction from Text. Diffs for prose are different from code diffs. Character-based diffs are not as useful on prose. Diffs on word boundaries or even concept boundaries are more useful. After extracting diff at multiple levels (char, word and line), we decided that word based level offered the best balance between conciseness and usability. For *Revisionista.PT*, we opted to customise an algorithm that is used in Google Docs to Track Changes Google [3] but in word-based mode. Word-based level diffs are important to allow interpreting the differences between versions (e. g., entities changes, typos corrected, etc.), improving readability and enabling simple browsing and search on diff text.

Changes Categorisation. Finally, once all the changes are computed we can do a broad categorisation of revisions found using an automatic rule-based approach:

- **Republished:** The publishing date of the article changed.
- **Undisclosed correction:** Text changes not explicitly disclosed.
- **Disclosed correction:** Text changes are accompanied by a note.
- **Disclosed update:** Updates the article with a note and new text only.
- **Live:** Coverage of live events or ongoing developments.

In Fig. 2, we can see a representation of the changes categorised as *Live.* "AO MINUTO" can be roughly translated to "minute by minute".


Compare Aug. 17, 2017, 5:03 p.m. with Aug. 17, 2017, 7:21 p.m.

AO MINUTO: Atentado terrorista em Barcelona faz 13 mortos - Mundo - Correio da Manhã

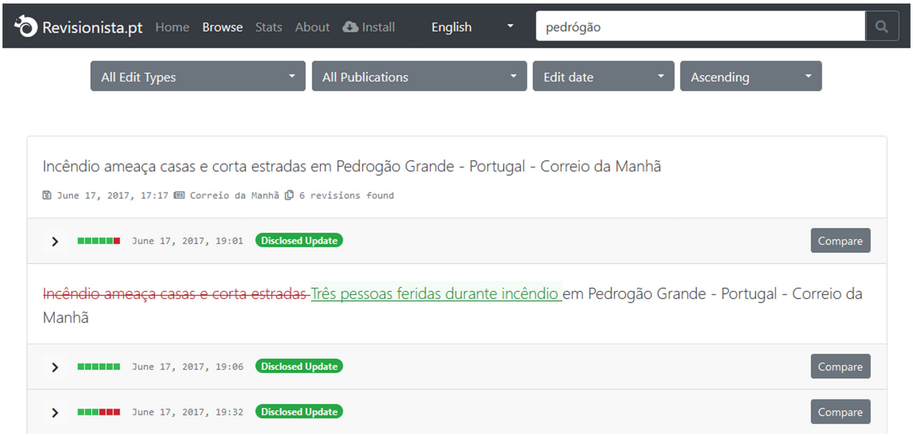
Text added appears highlighted in green.
Text ~~removed~~ appears highlighted in red.

Notícia

AO MINUTO: Atentado terrorista em Barcelona faz 13 ~~mortosDois homens armados barricados~~ mortosUm suspeito detido, outro abatido. Governo catalão confirma 80 feridos, 15 em ~~restaurante junto às Ramblas. Há dezenas de feridos.~~ estado grave.

Pelo menos 13 pessoas morreram ~~e dezenas~~ e mais de 50 ficaram

Fig. 2. Inline differences in *Revisionista.PT*



Incêndio ameaça casas e corta estradas em Pedrogão Grande - Portugal - Correio da Manhã

June 17, 2017, 17:17 Correio da Manhã 6 revisions found

June 17, 2017, 19:01 Disclosed Update Compare

~~Incêndio ameaça casas e corta estradas~~ Três pessoas feridas durante incêndio em Pedrogão Grande - Portugal - Correio da Manhã

June 17, 2017, 19:06 Disclosed Update Compare

June 17, 2017, 19:32 Disclosed Update Compare

Fig. 3. Searching and browsing in *Revisionista.PT*

3 *Revisionista.PT*

Revisionista.PT allows search and browsing of online news articles that were found to have post-publication changes. We examined 139,561 articles, of which 6,793 were edited after publication (4.87%), for a total number of edits of 7,405 (some articles were edited more than once). These articles are a part of 12 online Portuguese news platforms, from 2015 to 2018. In Fig. 3, we show the search interface, where relevance-based ranking can be overridden (edit date, number of edits, edit size) and results can be filtered by publication and edit type. In addition to showing added and removed text inline with the revisions, a special article view presents the changes contextualised within the article (see Fig. 2).

Revisionista.PT *Companion Extension*[3] is *open source* and can check on *Arquivo.PT* for different snapshots of the current page and show the changes. Our goal is to enable users to find undisclosed post-publication edits in the wild.

We intend to continue building new features into *Revisionista.PT* to provide more assurances and help journalists do their job with confidence. Our goal is to build automatic classifiers to help identify articles that need correction and significant changes that are too important to not include an explicit disclosure. Following this direction, we will create an interface to allow expert journalists to annotate changes found by *Revisionista.PT* according to different parameters to build a useful dataset. In addition, we will create a new tool that will allow journalists to subscribe to alerts helping them to keep their articles up-to-date with fresh information arriving from other news sources and published articles.

References

1. Andrew, B.C.: Media-generated shortcuts: do newspaper headlines present another roadblock for low-information rationality? Harv. Int. J. Press/Polit. **12**(2), 24–43 (2007). https://doi.org/10.1177/1081180X07299795
2. Digital, Culture, Media and Sport Committee: Disinformation and "fake news": Final report, vol. HC 1791. British House of Commons (2018)
3. Google: google/diff-match-patch: diff match patch is a high-performance library in multiple languages that manipulates plain text (2019). https://github.com/google/diff-match-patch
4. VillageVoice: Why did the New York times change their Brooklyn bridge arrests story? (2011). https://web.archive.org/web/20111003160551/blogs.villagevoice.com/runninscared/2011/10/why_did_the_new_1.php

[3] https://github.com/revisionista.