Perfil Público: Automatic Generation and Visualization of Author Profiles for Digital News Media

Nuno Guimarães
INESC TEC
University of Porto

Ricardo Campos
INESC TEC
University of Beira Interior

Alípio Jorge
INESC TEC
University of Porto

{nuno.r.guimaraes,ricardo.campos,alipio.jorge}@inesctec.pt

Abstract

Interest in the news has been declining and digital news subscriptions are still a hard sell for the average Internet user who is often used to consuming news through social media without any fees. To attract readers and engage with them, digital news outlets are forced to look for and integrate innovative solutions. In this work, we propose Perfil Público, a web platform that allows users to find news media authors based on their writing style and the topics they write about. Our solution combines a framework to generate authors' profiles automatically with a web platform that aims to facilitate the search, filter, and recommendation of digital news media authors.

1 Introduction

The massification of the Internet had an impact on the way consumers read news (Martinez-Alvarez et al., 2016). The COVID-19 pandemic only helped to accelerate the transition towards a digital-dominated media ecosystem. The declining interest in news, low newspaper sales, and only a small percentage of readers (17%) willing to pay for digital news subscriptions (Newman et al., 2023) had several impacts on the business model and format of journalism, which can only be overcome with innovative solutions and features. Several digital news media (such as the New York Times, but also Observador, Expresso, Público, and Correio da Manhã in Portugal) have adhered to subscription-only articles. Some of them, such as Jornal Público are working towards data journalism and interactive infographies to increase the number of paid subscribers. Nevertheless, a stronger commitment to the development of new solutions to captivate readers is necessary to guarantee the sustainability of the different digital news media. In that sense, we argue that one way to engage users is by allowing them to connect to authors based on their topics of interest and the author's writing style. The idea draws parallels with book authors, where readers have their preferences based on genre and writing style.

To cope with this, we developed Perfil Público, a platform that allows readers to find digital news media authors based on their writing style and topics of interest. Towards this end, we present a framework that aims to generate each author's profile based on a time span of news articles collected from the Arquivo.pt (Gomes et al., 2013) [1]. These profiles are then presented in a web platform, with search, recommendation, and filtering functionalities to promote easy navigation and captivate users' interest in the solution provided. To showcase this, we have developed a demo (`http://perfilpublico.dcc.fc.up.pt/`) on top of Público news outlet articles. We rely on Arquivo.pt for data retrieval and author profile generation, making this a scalable solution that can be easily adapted to other Portuguese news media, giving smaller and region-based news media a plug-and-play solution without the need for additional data to be stored locally.

Perfil Público methodology can be divided into two components: 1) the framework, which is responsible for the automatic generation of the authors' profiles and 2) the web platform for readers to find the authors that most suit their preferences. The framework and web platform are available in the GitHub repository [2].

2 Profile Generation Framework

The framework can be divided into three steps: 1) extraction of the required data 2) feature extraction at the article level and 3) authors' profile generation.

Data Extraction: The first step in creating author profiles is to collect the articles. Although,

[1] `https://arquivo.pt/`

[2] `https://github.com/nrguimaraes/PerfilPublico`

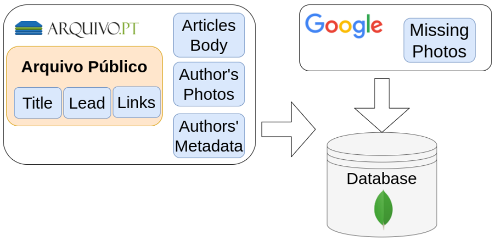


Figure 1: Data Extraction Diagram

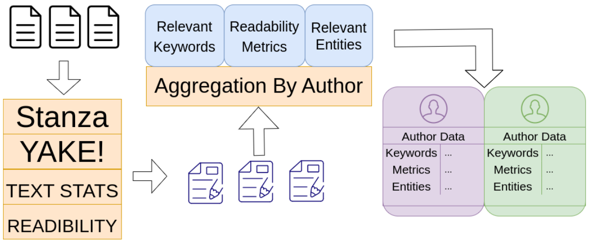


Figure 2: Feature Extraction and Profile Generation

nowadays, each digital news media has its webpage with the articles, it is not 100% guaranteed that all the articles' history is preserved and made available to the general public. In addition, smaller or local news digital media usually do not have the resources (financial or otherwise) to maintain a large archive of past articles. Therefore, to provide a scalable solution independent of the current state of the different digital media websites, Perfil Público builds its data extraction process on top of Arquivo.pt, which provides easy access to webpages and news articles from different Portuguese digital news sources through the years.

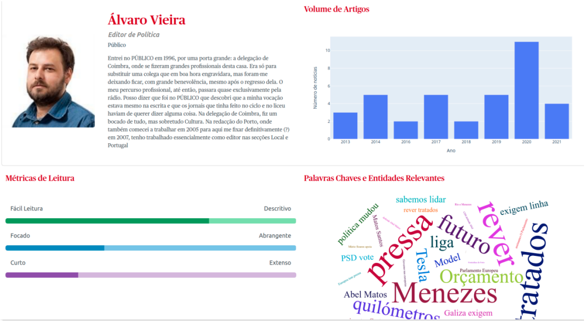
To collect articles from Arquivo.pt we rely on Public Archive (Campos et al., 2023), a tool that facilitates the extraction of news media articles from the Portuguese web archiving infrastructure. This tool allows the extraction of the title, author, lead, and link from the archives of 5 Portuguese digital news media websites. We complement Arquivo Público with modules to extract the body of the text, as well as the author's metadata (role and description) and photo. As a large number of authors did not have their profile photos available, we also used an unsupervised method supported by Google Images API to extract the remaining photos, using as a query the author's name and the name of the news media. Finally, all the information retrieved was stored in a database. A diagram of the data extraction workflow is presented in Figure 1.

Feature Extraction: In this step, we first clean each article by removing possible HTML tags and non-ASCII characters. Next, we run Stanza (Qi et al., 2020) Named-Entity Recognition (NER) model for Portuguese to extract Persons, Organizations, and Locations in each article. Similarly, we applied the state-of-the-art keyword extractor Yake (Campos et al., 2020) to extract the most relevant unigram, bigram, and trigram keywords. This data is essential to get a grasp on the topics that each article addresses.

Another set of features we focus on are text statistics such as the number of words and sentences, the mean of words per sentence, and the mean of syllables per word. In addition, to understand each author's writing style, we computed four readability metrics: Flesch-Kincaid, Gunning Fog, ARI, and Coleman-Liau. These metrics were already adapted to Brazilian Portuguese (Moreno et al., 2022). However, we modified the complex word list used to better suit the European Portuguese language. Similar to (Moreno et al., 2022), we considered a complex word if is not present in the first 5000 words available in Linguateca frequent tokens resources [3] after applying some filters to remove non-word entries.

Profile Generation: To automatically build each author's profile, we first calculate the aggregated features from all the articles of an author and average the numeric ones (readability metrics and text statistics). Concerning the entities, we select the top-10 most frequent entities for each category to characterize the author. In the keywords, we used Yake score and for each different keyword, we sum all the scores. Then, we selected the top-10 most relevant keywords based on those scores for each n-gram selected. We also established 3 metrics to characterize the writing style of each author. The first, concerns the average article length of the author (measured using the number of words). The second, evaluates the readability of the author (using the average of the four readability metrics extracted). Finally, the third tries to grasp the descriptiveness of the articles by leveraging the number of entities mentioned and their diversity. The intuition is that the diversity of the entities allows a richer contextualization. For example, an article that mentions at least a location, organization, and person is closer to following the 5w1h framework used in journalism (Bleyer, 1932) to answer when, where, who, why, and how. Additional entities will further

[3] https://www.linguateca.pt/acesso/tokens/formas.totalpt.txt



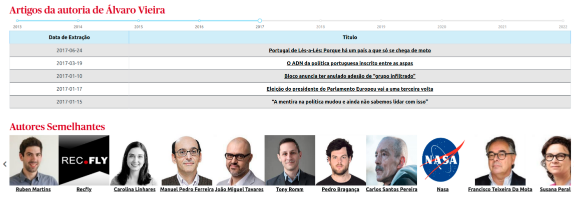
Álvaro Vieira

Editor de Política

Público

Entrei no PÚBLICO em 1996, por uma porta grande: a delegação de Coimbra, onde se fizeram grandes profissionais desta casa. Era só para substituir uma colega que em boa hora engravidara, mas foram-me deixando ficar, com grande benevolência, mesmo após o regresso dela. O meu percurso profissional, até então, passara quase exclusivamente pela rádio. Posso dizer que foi no PÚBLICO que descobri que a minha vocação estava mesmo na escrita e que os jornais que tinha feito no ciclo e no liceu haviam de querer dizer alguma coisa. Na delegação de Coimbra, fiz um bocado de tudo, mas sobretudo Cultura. Na redacção do Porto, onde também comecei a trabalhar em 2005 para aqui me fixar definitivamente (?) em 2007, tenho trabalhado essencialmente como editor nas secções Local e Portugal

Figure 3: Author profile with the features extracted



Artigos da autoria de Álvaro Vieira

Data de Extração	Título
2017-06-24	Portugal de Lés-a-Lés: Porque há um país a que só se chega de moto
2017-03-19	O ADN da política portuguesa inscrito entre as aspas
2017-01-10	Bloco anuncia ter anulado adesão de "grupo infiltrado"
2017-01-17	Eleição do presidente do Parlamento Europeu vai a uma terceira volta
2017-01-15	"A mentira na política mudou e ainda não sabemos lidar com isso"

Autores Semelhantes

Figure 4: Articles timeline and recommendations

enrich the context of the article. These 3 metrics were computed using the features extracted from each author's article and averaged by the collection of articles. Figure 2 presents the workflow for the feature extraction and profile generation processes.

3 Web Platform

Perfil Público web platform can be divided into three different sections.

Main Page: The main page features a search bar to search for a specific author's name. It also presents a visual hierarchy focused on horizontal scrolling with a set of topics and the most relevant authors associated with each one of them.

Advance Filters Page: Allows users to filter authors based on the readability metrics or topics they write. It provides three range sliders to adjust the interval for each metric. In addition, it also provides the user with a search bar to find authors based on specific topics.

Author Page: Each author's page combines the author's metadata and profile generated. The web profile includes a profile picture, name, description, role, and the number of articles published by year. The features mentioned in Section 2 are presented in different visualizations (e.g. the author's entities and keywords are converted in word clouds and the metrics are presented in progress bars). The web profile also integrates 1) a timeline with the titles, publication dates, and links to each article's preserved page in Arquivo.pt and 2) a recommendation section of authors with similar writing styles (based on the Euclidean distance of the 3 metrics proposed). Figure 3 and 4 show the features presented in each digital news media author's profile.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 | https://doi.org/10.54499/LA/P/0063/2020. The authors Alípio Jorge and Ricardo Campos would like to acknowledge the National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

References

W.G. Bleyer. 1932. *Newspaper Writing and Editing*. Houghton Mifflin.

Ricardo Campos, Diogo Correia, and Adam Jatowt. 2023. Public News Archive: A Searchable Sub-archive to Portuguese Past News Articles. In *Advances in Information Retrieval*, volume 13982, pages 211–216. Springer Nature Switzerland, Cham.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Daniel Gomes, David Cruz, João Miranda, Miguel Costa, and Simão Fontes. 2013. Search the Past with the Portuguese Web Archive. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 321–324. Association for Computing Machinery.

Miguel Martinez-Alvarez, Udo Kruschwitz, Gabriella Kazai, Frank Hopfgartner, David Corney, Ricardo Campos, and Dyaa Albakour. 2016. First International Workshop on Recent Trends in News Information Retrieval (NewsIR'16). In *Advances in Information Retrieval*, volume 9626, pages 878–882. Springer International Publishing.

Gleice Carvalho de Lima Moreno, Marco P. M. de Souza, Nelson Hein, and Adriana Kroenke Hein. 2022. ALT: um software para análise de legibilidade de textos em língua Portuguesa. ArXiv:2203.12135.

Nic Newman, Richard Fletcher, Kirsten Eddy, Craig T Robinson, and Rasmus Kleis Nielsen. 2023. Reuters Institute digital news report 2023. Technical report, Reuters Institute for the Study of Journalism.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.