FACTCK.BR: A New Dataset to Study Fake News

João Moreno
joao.guilherme.moreno@usp.br
University of São Paulo
São Paulo, Brazil

Graça Bressan
graca.bressan@usp.br
University of São Paulo
São Paulo, Brazil

ABSTRACT

Machine learning algorithms can be used to combat fake news propagation. For the news classification, labeled datasets are required, however, among the existing datasets, few separate verified false from skewed ones with a good variety of sources. This work presents FACTCK.BR, a new dataset to study Fake News in Portuguese, presenting a supposedly false News along with their respective fact check and classification. The data is collected from the ClaimReview, a structured data schema used by fact check agencies to share their results in search engines, enabling data collect in real time.

CCS CONCEPTS

• **Applied computing → Document searching**.

KEYWORDS

fake news, fact check, information extraction, dataset

ACM Reference Format:
João Moreno and Graça Bressan. 2019. FACTCK.BR: A New Dataset to Study Fake News. In *Brazilian Symposium on Multimedia and the Web (WebMedia '19), October 29-November 1, 2019, Rio de Janeiro , Brazil.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3323503.3361698

1 INTRODUCTION

Although not a new phenomenon, fake news has drawn public attention for its scope and influence in the debates leading up to the 2016 US presidential elections, and more recently the 2018 Brazilian elections.

The phenomenon can be explained by the massification of internet access and the polarization of public debate created an favorable environment for the dissemination of informational content that corroborates beliefs previously accepted by users who identify with one of the poles of the debate [5].

To combat the spread of fake news, has been proposed the use of machine learning algorithms for automated fact-checking (AFC) [1]. The work [2] defines an AFC as the "Holy Grail" of fact-checking. It also say it would not come as a surprise if we can get too close to that goal but never reach the utopian model.

WebMedia '19, October 29-November 1, 2019, Rio de Janeiro , Brazil

ACM ISBN 978-1-4503-6763-9/19/10…$15.00
https://doi.org/10.1145/3323503.3361698

Duke Reporters Lab started an initiative designed to achieve an operational AFC. The Claimbuster tool, still under development, aims to act as an end-to-end AFC [3].

In order to train this type of classifiers, we need to get access to datasets with a large number of Fake News. But given the difficulty, many datasets use lists of websites that propagated fake news at any moment. The problem with this approach is the difficulty in establishing criteria about which sites are on the list.

The Poynter Institute recently go back after published a list of such sites when received complaints about the criteria for a site to be on the list[1].

The FACTCK.BR corpus on the other hand, is based on claims in Portuguese individually checked by professional fact-checking initiatives.

In this section, we present the 3 types of fake news, the fact-checking agencies in Brazil and the ClaimReview, a framework to share structured data about fact-checking.

In the next section we detail the FACTCK.BR dataset. In the following section we compare the dataset with related works.

1.1 Fake News Types

Genuine news in the digital media is the articles from well-established journalistic sources, such as O Estado de São Paulo[2] and O Globo[3] in Brazil, or a well-established journalistic blog. On the other hand, fake news can be divided into three main groups [6]:

(1) **Hoaxes:** They are intended to mislead the audience by posing itself as genuine news. May cause material damage or harm to the victim.
(2) **Serious Fabrications:** These are articles written by the so-called yellow press. They can use *"clickbaits"*, lying headline that does not match the content, or hype to get traffic and financial gain.
(3) **Humorous Fakes:** They are distinguished from fabricated news, for a reader aware of the satirical intent of the content will not be willing to believe the information. An example of this content in Portuguese is the Sensacionalista website[4].

Hoax is the most dangerous from the 3 types of fakes and the focus of this work. The term fake news will be used as a synonym for hoaxes. One example[5] in Portuguese is displayed in Table 1.

ClaimReview[6] is an open source framework from Schema.org, an organization with a mission to promote schemas for structured

[1] https://www.poynter.org/letter-from-the-editor/2019/letter-from-the-editor/
[2] https://www.estadao.com.br
[3] https://oglobo.globo.com
[4] https://www.sensacionalista.com.br
[5] https://aosfatos.org/noticias/e-falso-que-lula-aparece-na-lista-de-pessoas-mais-ricas-do-mundo-da-revista-forbes/
[6] www.schema.org/ClaimReview

Table 1: Fake News Example

(Fake News Claim) Profissão: Metalúrgico. Fortuna: U$ 2 bilhões (segundo a revista Forbes, 2006). Agora eu pergunto a você que vota no PT: você conhece algum trabalhador honesto que tenha tamanha fortuna?
(Fact Check) É falsa a informação de que o ex-presidente Luiz Inácio Lula da Silva teve sua fortuna orçada em US$ 2 bilhões pela revista Forbes. Além de não ter sido citado em nenhuma das listas de pessoas mais ricas do mundo produzida anualmente pela publicação, o patrimônio de Lula estava orçado em R$ 12,3 milhões.

data on the Internet, used by fact-checking organizations to make it easier to share a verification story with technological companies.

Inserted into the webpage source code, the schema can store structured data about the claim, its authors (if available), the fact-checking organization, the review, and the rating.

If the article checks a set of claims, for example in a political interview or TV debate, a ClaimReview structure is used for each published claim.

With ClaimReview companies like Google and Facebook are able to feature stories from fact-checking agencies on their products.

Figure 1 shows an verification story displayed on Google Search to a claim in Portuguese language.

Figure 1: Fact Check Displayed on Google.

1.2 Fact-checking agencies

There are 160 active fact-checking agencies in the world in 2019, according to *Duke Reporters LAB* website[7]. In Brazil it is a growing ecosystem with currently 9 initiatives, displayed in Table 2.

Table 2: Fact-Checking Agencies in Brazil

Agency	Website
Aos Fatos	www.aosfatos.org
Lupa	www.piaui.folha.uol.com.br/lupa/
Truco	www.apublica.org/checagem/
Checamos	www.checamos.afp.com
É Isso Mesmo?	www.blogs.oglobo.globo.com/eissomesmo/
UOL Confere	www.noticias.uol.com.br/confere/
Hoax Reports	www.ebc.com.br/hoax
E-farsas	www.e-farsas.com
Boatos.org	www.boatos.org

[7]https://reporterslab.org/fact-checking/

In order to create the FACTCK.BR corpus, the information present in the ClaimReview was captured using a web crawler, from articles of the 3 most active Brazilian fact-checking agencies, from its first available article until the date of Jun/19.

The three selected agencies are part of a Google initiative that shows fact checks on Google Search and Google News products using ClaimReview[8], they are listed below:

(1) **Aos Fatos:** One of the largest Brazilian fact-checking agencies. More than 350 fact-check articles, active since 2015.
(2) **Lupa:** It is considered the first fact-checking agency in Brazil. More than 450 fact-check articles published, active since 2015.
(3) **Truco:** Fact-checking initiative from Publica Investigative Journalism Agency. More than 260 articles published since 2015.

Unfortunately, not yet all fact-checking initiatives use ClaimReview to share their articles. But this scenario may change as they find some reason to use it.

2 FACTCK.BR CORPUS

The FACTCK.BR corpus is a dataset in the form of a 9 column table with 1309 lines, each one corresponding to a claim. The corpus is storage in a text file with *tabs* separating the columns. The Table 3 describes the attributes present in each column of the corpus.

Table 3: FACTCK.BR Attributes

Columns	Description
URL	Check article web address
Author	Initiative id.
datePublished	Check publication date
claimReviewed	Claim analyzed
reviewBody	Check text
Title	Title of the article
ratingValue	Numerical classification
bestRating	Lenght of the scale
alternativeName	Text label

Each fact-check agency uses its own labels to classify claims. The Labels can be for example: false, true, impossible to prove, distorted, exaggerated, controversial, without context, inaccurate, among others. In the claimReview the classification of the claim should be performed on a linear scale, where 1 is false and the higher number at the scale, set by the value in BestRating is true, with the interval representing half trues. The distributions of that classifications from each agency is displayed in Table 4.

The corpus stores the claim and the review provided in ClaimReview. We also add the title of the article, that can be many representative about the check. In the case of a article with multiple checks, the title will be the same. The agency Lupa don't include the Review in ClaimReview schema. Table 5 displayed average number of characters from each attribute to each agencies.

The FACTCK.BR corpus can be easily updated by web crawling new articles. The ClaimReview is located in the source code by

[8]https://aosfatos.org/noticias/google-une-se-aos-fatos-e-lanca-selo-de-verificacao-de-fatos/

Table 4: Claim Review Classification Distribution.

Agencies	False	Half True	True	Total
Ag. Publica	146	180	85	411
Lupa	469	25	34	528
Aos Fatos	328	41	15	370
Total	943	246	120	1309

Table 5: Average Char. Length

Agencies	Title	Claim	Review
Ag. Publica	74	188	184
Aos Fatos	87	130	407
Lupa	83	97	None

searching for an script of the type *LD+JSON*. Using Python language we use the library *Feedparser* to get a RSS feed from the agencies and *BeautifulSoup* to navegate in the source code.

3 RELATED WORK

Other datasets have been proposed for the study of fake news, each having its strengths and weaknesses.

3.1 LIAR

LIAR is a dataset of manually labeled short statements from the US fact-checking website Politifact[9]. LIAR has collected more than 10000 short statements over a decade [7].

3.2 Fake.Br

It is one of the first database of fake news in Portuguese [4]. This database has 4000 articles, 2000 of which are considered true articles, obtained in digital versions of traditional vehicles and another 2000 articles classified as false by the authors, obtained from sites considered to be false news spreaders.

A characteristic of the Fake.Br database is that the fake news articles were collected from a set of websites considered to be fake news propagators, but each article was not individually reviewed. The news labeled false may consist of some hoaxes, but most of it is serious fabrications written from the so-called yellow press.

3.3 Boatos.org

The Kaggle platform offers a database of rumors checked by Boatos.org[10] fact-checking website. This database contains 1900 rumors written in Portuguese or Spanish languages, with the link to the respective fact check article on the website.

Unlike the Fake.Br database on the Boatos.org database, the fake news was individually reviewed by the site staff and published as a fact check article. The Table 6 shows a comparison of the 3 fake news datasets in portuguese: Fake.Br, Boatos.org and FACTCK.BR.

[9] politifact.com
[10] https://www.kaggle.com/rogeriochaves/boatos-de-whatsapp-boatosorg

Table 6: Fake News Datasets in Portuguese

Datasets	Fake.Br	Boatos.org	Factck.Br
Claim	X	X	X
Fact Checking		X	X
Yellow Press	X		
Full Text	X	X	
Multiple Sources	X		X
Upgradeable			X

4 CONCLUSIONS

The proposed corpus has a number of advantages over existing datasets. First, for each claim is presented the text of its respective check and a linear numerical classification for its veracity.

Each claim was analyzed manually and individually and was not based on lists of unreliable news sources.Also it is based in multiple sources, not being skewed at a single source.

Another advantage is that the database can be easily updated as new articles are published, keeping up to date for use in applications.

The **FACTCK.BR** dataset and the script to update the corpus are available at Github on https://github.com/jghm-f/FACTCK.BR.

5 FUTURE WORK

The next step of the research is to use machine learning to extract information from the claims to validate in the corresponding review.

Another interesting activity would be to expand the database with checks from another countries and languages.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

REFERENCES

[1] Lucas Graves. 2018. Understanding the promise and limits of automated fact-checking. *Factsheet* 2 (2018), 2018–02.
[2] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium*.
[3] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (2017), 1945–1948.
[4] Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. 2018. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In *Computational Processing of the Portuguese Language*. Springer International Publishing, 324–334.
[5] Marcio Moretto Ribeiro Pablo Ortellado. 2018. *Polarização e desinformação online no Brasil* (44 ed.).
[6] Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. American Society for Information Science, 83.
[7] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).