FakeRecogna: a new Brazilian corpus for fake news detection

Gabriel L. Garcia*[0000−0003−1236−7929], Luis C. S. Afonso[0000−0002−5543−3896], and João P. Papa[0000−0002−6494−7514]

School of Sciences, São Paulo State University, Bauru, Brazil
gabriel_linao@hotmail.com,{luis.afonso, joao.papa}@unesp.br

Abstract. Fake news has become a research topic of great importance in Natural Language Processing due to its negative impact on our society. Although its pertinence, there are few datasets available in Brazilian Portuguese and mostly comprise few samples. Therefore, this paper proposes creating a new fake news dataset named FakeRecogna that contains a greater number of samples, more up-to-date news, and covering a few of the most important categories. We perform a toy evaluation over the created dataset using traditional classifiers such as Naive Bayes, Optimum-Path Forest, and Support Vector Machines. A Convolutional Neural Network is also evaluated in the context of fake news detection in the proposed dataset.

Keywords: Fake News · Corpus · Portuguese.

1 Introduction

Fake news has a significant impact on society, as it affects people's education, decision-making, and attitudes [5]. According to Rubin [27], fake news can be categorized into three main groups: (i) hoaxes, (ii) serious fabrications, and (iii) humorous fakes. Hoaxes are intended to mislead the audience by posing themselves as genuine news. They may cause material damage or even harm to the victim. Serious fabrications stand for articles written by the so-called "yellow press". They can use "clickbait", i.e., a lying headline that does not match the content or hype to get traffic and financial gain. Last but not least, humorous fakes are distinguished from fabricated news, for a reader aware of the satirical intent of the content will not be willing to believe the information.

To reduce the spreading of fake news, news agencies have created and supported many fact-checking pages to verify the veracity of news and explain why the news is fake. Also, many works using Natural Language Processing (NLP) have addressed such a problem. Aphiwongsophon [6] proposed the use of machine learning techniques to detect fake news using three popular methods in the experiments. Ahmed [4] introduced a fake news detection model that uses n-gram

* Author to whom correspondence should be addressed.

analysis and machine learning techniques. Gilda [14] explored term frequency-inverse document frequency (TF-IDF) of bi-grams and probabilistic context-free grammar detection in a corpus of about 11,000 articles. Jain [17] proposed a combination of two datasets that contain an equal number of both true and fake news articles on politics. The work extracted linguistic/stylometric features, a bag of words, and TF-IDF features to feed different machine learning models.

Brazil is no different. We can mention the works of Silva et al. [28], which proposed a dataset of labeled real and fake news in Portuguese and performed a comprehensive analysis of machine learning methods for fake news detection. Queiroz et al. [3] compared machine learning algorithms in three languages (English, Portuguese, and Spanish) describing how the results are successful in describing false, satirical, and legitimate news in three different languages. Souza et al. [29] proposed an extended method that, in addition to the grammatical classification and polarity-based sentiment analysis, also applied the analysis of emotions to detect fake news.

However, to tackle such a problem in Brazil, we must have a dataset of local news. Currently, there are only a few datasets in Brazilian Portuguese, which are mostly outdated and not large enough. As an example, one of the most used ones is the *Fake.Br Corpus*, with news dating back from 2016 to 2018. This work proposes building up a new fake news dataset focusing on Brazilian news. The main idea is to collect the most updated news (real and fake ones) from well-known agency news web pages, such as G1, UOL, and Extra, and to increase the number of samples for research. Hence, the main contributions of this work are three-fold:

- A new fake news corpus in Brazilian Portuguese called *FakeRecogna*;
- A larger and updated corpus;
- To foster the research on fake news in Brazilian Portuguese.

The remainder of this paper is organized as follows: Section 2 provides a review of related works, while Section 3 describes the proposed dataset. Sections 4 and 5 present the methodology and results of a toy-evaluation performed over the proposed dataset, respectively. Finally, Section 6 states conclusions.

2 Related Works

This section presents and describes the primary datasets concerning fake news in Brazilian Portuguese.

- Fake.Br: This corpus [20] figures as one of the first and most used datasets concerning fake news in Brazil comprising 7,200 news where 3,600 are fake, and the remaining 3,600 are real ones. The collection was manually analyzed, and only those that were entirely fake were kept in the dataset. An interesting characteristic of this dataset is that for each fake news, a real one was searched by performing a lexical similarity measure using keywords from the fake ones.

- FACTCK.BR: The FACTCK.BR [21] corpus is a dataset in Portuguese structured according to the ClaimReview framework, which was created to ease the sharing of verified news among technological companies. The dataset comprises 1,309 claims with non-binary labels, which are: false, true, impossible to prove, distorted, exaggerated, controversial, without context, and inaccurate, among others. However, the claims in FACTCK.BR are divided between False, Half True, and True.
- Boatos.org: The Kaggle platform[1] provides a dataset comprised of 1,900 fake news verified by Boatos.org, which are either in Portuguese or Spanish. For each fake news, there is a link to a page rebutting it.
- Covid-19 Rumor: The dataset comprises rumors and non-rumors related to COVID-19 collected from three sources: (i) the Brazilian Ministry of Health official website, (ii) a journalistic initiative named Boatos.org focused on debunking online rumors, including COVID-19, and (iii) the O Globo news agency. The COVID-19 RUMOR dataset has 1,291 rumors and 8 non-rumors [12] labeled by the teams of journalists from the sources.

3 FakeRecogna Corpus

This section presents the details behind the design of the proposed corpus. FakeRecogna is a dataset comprised of real and fake news. The real news is not directly linked to fake news and vice-versa, which could lead to a biased classification. The news collection was performed by crawlers developed for mining pages of well-known and of great national importance agency news. The web crawlers were developed based on each analyzed webpage, where the extracted information is first separated into categories and then grouped by dates. The plurality of news on several pages and the different writing styles provide the dataset with great diversity for natural language processing analysis and machine learning algorithms.

The fake news mining was mainly focused on pages created between 2019 and 2021 and mentioned by the Duke Reporters Lab[2], which provides a list of pages that verify the veracity of news worldwide. There were 160 active fact-checking agencies in the world in 2019, and Brazil figures as a growing ecosystem with currently 9 initiatives. Table 1 presents the current initiatives as well as the number of fake news collected from each source. Due to content restrictions, there were considered 6 out of the 9 pages during search with a great variation in the number of fake news extracted from each one, ending in 5,951 samples.

Concerning the real news, the crawlers searched portals such as G1[3], UOL[4] and Extra[5], which are publicly recognized as reliable news outlets, besides the Ministry of Health of Brazil[6] home page, resulting in a collection of over 100,000

[1] https://www.kaggle.com/rogeriochaves/boatos-de-whatsapp-boatosorg
[2] https://reporterslab.org/fact-checking/
[3] https://g1.globo.com/
[4] https://www.uol.com.br/
[5] https://extra.globo.com/
[6] https://www.gov.br/saude/pt-br

Table 1. Fact-checking agencies in Brazil.

agency	web address	# news
AFP Checamos	https://checamos.afp.com/afp-brasil	509
Agência Lupa	https://piaui.folha.uol.com.br/lupa/	–
Aos Fatos	https://aosfatos.org	–
Boatos.org	https://boatos.org	2,605
Estadão Verifica	https://politica.estadao.com.br/blogs/estadao-verifica	–
E-farsas	https://www.e-farsas.com	812
Fato ou Fake ("Fact or Fake")	https://oglobo.globo.com/fato-ou-fake	1,055
Projeto Comprova	https://projetocomprova.com.br	388
UOL Confere	https://noticias.uol.com.br/confere	582
total		5,951

samples. From this set, there were filtered out 5,951 samples to keep the balance between classes and, thus, resulting in a dataset comprised of 11,902 samples. Each sample has 8 metadata fields, as described in Table 2.

Table 2. Metadata used to describe each sample.

columns	description
Title	Title of article
Sub-title (if available)	Brief description of news
News	Information about the article
Category	News grouped according to your information
Author	Publication author
Date	Publication date
URL	Article web address
Class	0 for fake news and 1 for real news

The collected texts are distributed into six categories in relation to their main subjects: Brazil, Entertainment, Health, Politics, Science, and World. These categories are defined based on the journal sections where the news were extracted. The distribution of news by category and its percentages are described in Table 3.

Table 4 provides a comparison among FakeRecogna dataset and the ones mentioned in Section 2. Notice that only FACTCK.BR makes uses of a third class, i.e., half true, that considers that a piece of news may contain facts to support a fake idea. However, the class "half true" is not used in the experiments. The FakeRecogna[7] dataset is available as a single XLSX file that contains 8 columns for the metadata, and each row stands for a sample (real or fake news).

4 Methodology

This section presents a toy evaluation employed over the proposed dataset.

[7] https://github.com/recogna-lab/datasets/tree/master/FakeRecogna

Table 3. Amount of news per category in the FakeRecogna.

category	# news	%
Brazil	904	7.6
Entertainment	1,409	12.0
Health	4,456	37.4
Politics	3,951	33.1
Science	602	5.1
World	580	4.9
total	11,902	100.0

Table 4. Comparison between datasets.

datasets	# news	# real	# fake	# half true
FakeRecogna	11,902	5,951	5,951	−−
Fake.Br	7,200	3,600	3,600	−−
FACTCK.BR	1,309	411	528	370
Boatos.org	1,900	−−	1,900	−−
Covid-19 Rumor	1,299	8	1,291	−−

4.1 Pre-processing

The preprocessing step comprises four steps:

1. Truncation: this step is essential to avoid any classification bias due to the significant variation in size among sentences, especially between fake and real ones. The latter tend to be much longer than the former.
2. Standardization of terms: removal of words that may bias the news, such as, “enganoso”, “boato”, “#fake” and, so on. Punctuation, special characters, and URLs were also removed and the standardization to lowercase letters.
3. Lemmatization: The lemmatization comprises the morphological analysis of the words resulting in their canonical form [15]. Lemmatization also considers the context of the word to solve other problems, such as disambiguation, that is, differentiating the meaning of identical words depending on the context.
4. Removal of stop words: removal of words considered irrelevant for the understanding of the news, such as articles and prepositions.

All pre-processing steps were performed using the SpaCy library [16], as it uses state-of-the-art approaches for such a purpose. Its performance is usually superior when compared to the Natural Language Toolkit, a.k.a. NLTK [7]. SpaCy also builds up a syntax tree for each sentence, a more robust method that produces much more information about the text.

4.2 Text Representation

The experiments considered two techniques to compute the text representation: (i) Bag-of-Words (BoW) [18], and (ii) FastText [8]. Bag-of-Words computes the

text representation as a bag (multiset) of words, disregarding the grammar and even the order of the words, but maintaining the multiplicity, that is, the frequency of each word in the text. The FastText extends the Word2Vec model by representing each word as an n-gram of characters, thus helping to capture the meaning of shorter words and allows embeddings to understand suffixes and prefixes. The parameter values for FastText were: embedding size equals to 200 dimensions, maximum number of unique words as $4,000$, and the maximum amount of tokens for each sentence equals to $1,000$. Since BoW is a simple text representation, we did not perform experiments using CNN.

4.3 Classifiers

The experiments were performed using the following classifiers:

- Multi-Layer Perceptron (MLP) [2]
- Naive Bayes (NB) [19]
- Optimum-Path Forest (OPF) [24, 23]
- Random Forest (RF) [10]
- Support Vector Machines (SVM) [9]
- Convolutional Neural Networks (CNN): [11, 13]

Concerning the implementation and parameters, all classifiers but OPF[8] and CNN[9] used the Scikit-learn library [25] with their default parameter values. Regarding the CNN training procedure, the following setup was used: Adam as the optimizer [10] and the Binary Cross Entropy as loss function[11].

4.4 Evaluating Measures

The classifiers were evaluated over four measures: (i) precision, (ii) recall, (iii) f1-score, and (iv) accuracy. The average results over a 5-fold cross-validation are presented in the next section.

4.5 Additional experiments

No removal of words One of the preprocessing steps perform the removal of words that might bias the news (e.g., "enganoso", "boato", and "#fake"). This experiment evaluates whether removing the standardization of terms step from the workflow has any influence in the classification process and the lemmatization step. The experiment is performed using all aforementioned text representation techniques and classifiers.

[8] We used an implementation provided by OPFython library [26].
[9] We used the well-known TensorFlow library [1]
[10] https://keras.io/api/optimizers/adam/
[11] https://keras.io/api/losses/probabilistic_losses/

Data Augmentation The second additional experiment studies the impact of adding the data from FakeRecogna to other datasets. This experiment uses the Fake.Br and FACTCK.BR datasets being classified by the CNN and FastText for text representation in two rounds. The first round classifies the original datasets, whereas the second round merges the datasets mentioned above with FakeRecogna. Notice that the experimental protocol is the same as discussed in the following section.

5 Experimental Results

Table 6 presents the average results for each text representation and classification techniques, with the best results in bold. A more in-depth analysis shows us that the best result with the BoW was achieved by the MLP classifier, followed by RandomForest and SVM, with results that surpassed more than 90% of correct answers. The outcomes show that even using a standard natural language processing technique, i.e., BoW, the achieved results were very interesting.

Table 5. Experimental results over FakeRecogna corpus.

classifier	precision		recall		f1-score		accuracy	
	BoW	FastText	BoW	FastText	BoW	FastText	BoW	FastText
MLP	**0.931**	**0.850**	**0.931**	**0.848**	**0.930**	**0.848**	**93.1%**	**84.8%**
NB	0.896	0.712	0.898	0.680	0.897	0.666	89.7%	68.1%
OPF	0.834	0.784	0.834	0.784	0.834	0.782	83.4%	78.4%
RF	0.924	0.840	0.922	0.840	0.922	0.840	92.3%	84.0%
SVM	0.926	0.832	0.925	0.810	0.926	0.804	92.5%	80.8%
CNN	**-**	**0.942**	**-**	**0.942**	**-**	**0.942**	**-**	**94.28%**

On the other hand, using a more robust text pre-processing architecture like FastText and the Convolutional Neural Network, the best results surpassed 94%. FastText was considered here for its enriched representations where each word is represented by an embedding of the entire word itself plus its n-grams. SkipGram, for example, provides simpler presentations that take into account only the word itself.

5.1 No removal of words

This experiment evaluates the influence of both keeping words and terms that may bias the classification and removing the lemmatization step. By the results, FastText provided a slight increase in precision and accuracy. However, BoW as text representation faced a degradation in its precision. We believe the experiments using BoW were deeply affected by the removal of the lemmatization step, since that words and its variations are kept. For instance, the words *playing*, *plays*, and *played* all become *play* after lemmatization.

Table 6. Experimental results over FakeRecogna corpus without pre-processing.

classifier	precision		recall		f1-score		accuracy	
	BoW	FastText	BoW	FastText	BoW	FastText	BoW	FastText
MLP	**0.926**	**0.848**	**0.926**	**0.844**	**0.926**	**0.844**	**92.6%**	**84.3%**
NB	0.896	0.730	0.896	0.688	0.896	0.674	89.6%	69.1%
OPF	0.680	0.778	0.660	0.776	0.652	0.776	65.9%	77.7%
RF	0.918	0.840	0.918	0.840	0.918	0.840	91.8%	84.0%
SVM	0.918	0.820	0.918	0.796	0.918	0.786	91.8%	79.5%
CNN	–	**0.942**	–	**0.942**	–	**0.942**	–	**94.26%**

5.2 Augmentation Study

As mentioned earlier, the experiments concerning data augmentation were performed using the same CNN from the previous section since it achieved the best results. The main goal of this experiment is to evaluate how much the accuracy can be increased when the FakeRecogna dataset is used to augment Fake.Br and FACTCK.BR datasets. Table 7 presents the experimental results for the original and augmented datasets.

Table 7. Experimental results concerning data augmentation.

Dataset	**Precision**	**Recall**	**F1-score**	**Accuracy**
Fake.Br	**0,954**	**0,954**	**0,954**	**95.4%**
Fake.Br + FakeRecogna	0.946	0,946	0,946	94.6%
FACTCK.BR	0,871	0,871	0,871	87.7%
FACTCK.BR + FakeRecogna	**0,933**	**0,933**	**0,933**	**93.3%**

According to the results, the best outcome (95.4%) was achieved over the original Fake.Br dataset. By joining Fake.Br to FakeRecogna, the accuracy has a minor decrease. One reason for such a decrease is that each fake news has some level of similarity to some real news, which is a characteristic that is not present in the FakeRecogna dataset. Notice that a few other works, such as one of Okano et al. [22], achieved a higher accuracy (96%) but using the raw text, which may bias the results since real news are usually much longer than fake ones. Regarding the FACTCK.BR dataset, we obtained 87.7% of accuracy over the original data and 93.2% by merging it with FakeRecogna, an improvement of more than 5%. This scenario would be the closest to reality since both datasets' real and fake news do not share any similar relationship between them.

6 Conclusions and Future Works

In this paper, we proposed a new corpus for fake news detection in Brazilian Portuguese called FakeRecogna. The proposed corpus presents a series of advantages over the existing ones, which include:

- Up-to-date news (2019-2021);
- Greater number of categories;
- Greater number of news;
- It covers current issues, such as the Covid-19 pandemic;
- Multiple sources;
- Internationally recognized sources.

By building up the dataset, we hope to foster the research on fake news detection over texts in Brazilian Portuguese since the low number of samples available drives most works to English-based datasets.

Additionally, we performed experiments in which we managed to produce interesting results considering the problem's difficulty. There were employed two text representation techniques and six classifiers, with accuracies higher than 90% in some cases.

As future works, we intend to extract more fake news from other sources to increase the dataset size and use other word embeddings. Regarding deep learning, we intend to consider new architectures such as attention models and BERT.

Acknowledgments

The authors are grateful to FAPESP grants #2013/07375-0, #2014/12236-1, and #2019/07665-4, and CNPq grants #307066/2017-7 and #427968/2018-6.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation. p. 265–283. OSDI'16, USENIX Association, USA (2016)
2. Abirami, S., Chitra, P.: Chapter fourteen - energy-efficient edge based real-time healthcare support system. In: Raj, P., Evangeline, P. (eds.) The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases, Advances in Computers, vol. 117, pp. 339–368. Elsevier (2020)
3. Abonizio, H.Q., de Morais, J.I., Tavares, G.M., Barbon Junior, S.: Language-independent fake news detection: English, portuguese, and spanish mutual features. Future internet **12**(5) (2020)
4. Ahmed, H., Traore, I., Saad, S.: Detection of online fake news using n-gram analysis and machine learning techniques. pp. 127–138 (2017)
5. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives **31**, 211–236 (2017)
6. Aphiwongsophon, S., Chongstitvatana, P.: Detecting fake news with machine learning method. In: 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 528–531 (2018)

7. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. (2009)
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics **5**, 135–146 (06 2017)
9. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. p. 144–152. COLT '92, Association for Computing Machinery, New York, NY, USA (1992)
10. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001)
11. Dalto, M., Matuško, J., Vašak, M.: Deep neural networks for ultra-short-term wind forecasting p. 1657–1663 (2015)
12. Endo, P.T., Nascimento Campos, G.R., de Lima Xavier, M.E., Carvalho Monteiro, K.H., Ferreira da Silva Barros, M.H.L., Silva, I., Santos, B.: Covid-19 rumor: a classified dataset of covid-19 related online rumors in brazilian portuguese. Mendeley Data **V3** (2021)
13. Ferreira, A., Giraldi, G.: Convolutional neural network approaches to granite tiles classification **84**, 19–29 (2017)
14. Gilda, S.: Evaluating machine learning algorithms for fake news detection. 2017 IEEE 15th Student Conference on Research and Development (SCOReD) pp. 110–115 (2017)
15. Hippisley, A.: Lexical analysis. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, Second Edition, pp. 31–58. Chapman and Hall/CRC (2010)
16. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
17. Jain, M.K., Gopalani, D., Meena, Y.K., Kumar, R.: Machine learning based fake news detection using linguistic features and word vector features. 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) pp. 1–6 (2020)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Bengio, Y., LeCun, Y. (eds.) 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings (2013)
19. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
20. Monteiro, R.A., Santos, R.L.S., Pardo, T.A.S., de Almeida, T.A., Ruiz, E.E.S., Vale, O.A.: Contributions to the study of fake news in portuguese: New corpus and automatic detection results pp. 324–334 (2018)
21. Moreno, J.a., Bressan, G.: Factck.br: A new dataset to study fake news. In: Proceedings of the 25th Brazillian Symposium on Multimedia and the Web. pp. 525–527. WebMedia '19, Association for Computing Machinery, New York, NY, USA (2019)
22. Okano, E.Y., Liu, Z., Ji, D., Ruiz, E.E.S.: Fake news detection on fake.br using hierarchical attention networks. In: Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., Gonçalves, T. (eds.) Computational Processing of the Portuguese Language. pp. 143–152. Springer International Publishing, Cham (2020)
23. Papa, J.P., Falcão, A.X., Albuquerque, V.H.C., Tavares, J.M.R.S.: Efficient supervised optimum-path forest classification for large datasets. Pattern Recognition **45**(1), 512–520 (2012)

24. Papa, J.P., Falcão, A.X., Suzuki, C.T.N.: Supervised pattern classification based on optimum-path forest. International Journal of Imaging Systems and Technology **19**(2), 120–131 (2009)
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
26. de Rosa, G.H., Papa, J.P., Falcão, A.X.: Opfython: A python-inspired optimum-path forest classifier (2020), https://arxiv.org/abs/2001.10420
27. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: Three types of fakes. In: Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community. ASIST '15, American Society for Information Science, USA (2015)
28. Silva, R.M., Santos, R.L., Almeida, T.A., Pardo, T.A.: Towards automatically filtering fake news in portuguese. Expert Systems with Applications **146**, 113199 (2020)
29. de Souza, M.P., da Silva, F.R.M., Freire, P.M.S., Goldschmidt, R.R.: A linguistic-based method that combines polarity, emotion and grammatical characteristics to detect fake news in portuguese. In: Proceedings of the Brazilian Symposium on Multimedia and the Web. pp. 217–224. WebMedia '20, Association for Computing Machinery, New York, NY, USA (2020)