Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results

Rafael A. Monteiro[1], Roney L. S. Santos[1], Thiago A. S. Pardo[1], Tiago A. de Almeida[2], Evandro E. S. Ruiz[3], and Oto A. Vale[4]

[1] Interinstitutional Center for Computational Linguistics (NILC)
University of São Paulo - São Carlos, Brazil
{rafael.augusto.monteiro, roneysantos}@usp.br, taspardo@icmc.usp.br
[2] Federal University of São Carlos - Sorocaba, Brazil
talmeida@ufscar.br
[3] University of São Paulo - Ribeirão Preto, Brazil
evandro@usp.br
[4] Interinstitutional Center for Computational Linguistics (NILC)
Federal University of São Carlos - São Carlos, Brazil
otovale@ufscar.br

Abstract. Fake news are a problem of our time. They may influence a large number of people on a wide range of subjects, from politics to health. Although they have always existed, the volume of fake news has recently increased due to the soaring number of users of social networks and instant messengers. These news may cause direct losses to people and corporations, as fake news may include defamation of people, products and companies. Moreover, the scarcity of labeled datasets, mainly in Portuguese, prevents training classifiers to automatically filter such documents. In this paper, we investigate the issue for the Portuguese language. Inspired by previous initiatives for other languages, we introduce the first reference corpus in this area for Portuguese, composed of aligned true and fake news, which we analyze to uncover some of their linguistic characteristics. Then, using machine learning techniques, we run some automatic detection methods in this corpus, showing that good results may be achieved.

Keywords: Fake news · Reference corpus · Linguistic features · Machine learning

1 Introduction

Since the earliest times, long before the advent of computers and the web, fake news (also known as deceptive news) were transmitted through the oral tradition, in the form of rumors (face to face) or in the yellow/sensational press, either to "innocently" talk about other people lives, or to intentionally harm the reputation of other people or rival companies. Nowadays, social networks and instant messenger apps have allowed such news to reach an audience that was never imagined before the web era. Due to their appealing nature, they spread rapidly [20], influencing people behavior on several subjects, from healthy issues (e.g., by revealing miraculous medicines) to politics and economy (as in the recent Cambridge Analytica/Facebook scandal[5] and in the Brexit situation[6]).

As the spread of fake news has reached a critical point, initiatives to fight back fake news have emerged. On the one hand, journalistic agencies have supported fact checking sites (e.g., Agência Lupa[7] and Boatos.org[8]) and big digital companies (as Facebook[9]) have attempted to block fake news and to educate users. On the other hand, academic efforts have been made by studying how such news spread, the behavior of the users that produce and read them, and language usage characteristics of fake news, in order to identify such news. This last research line - on language characteristics - has been mainly explored in the Natural Language Processing (NLP) area.

In NLP, the attempts to deal with fake news are relatively recent, both on the theoretical (e.g., [7, 12, 24]) and practical points of view ([1, 11, 16, 18]). Some previous work has showed that humans perform poorly on separating true from fake news [3, 10] and that the domain may affect this [14], but others have produced promising automatic results. Despite the advances already made, the lack of available corpora may compromise the evaluation of different approaches.

To fill this important gap, in this paper we investigate the issue of fake news detection for the Portuguese language. Inspired by previous initiatives for other languages, to the best of our knowledge, we introduce the first reference corpus in this area for Portuguese. This corpus is composed of aligned true and fake news, which we analyze to uncover some of their linguistic characteristics. Then, using traditional machine learning techniques and following some of the ideas of [16] and [22], we perform tests on the automatic detection of fake news, achieving good results. One of our main goals is that our corpus and methods may support future researches in the area.

The remainder of this paper is organized as follows. In Section 2, we briefly review the essential related work. Section 3 offers details about the newly-created corpus. In Section 4, we report our machine learning approaches for fake news detection. Finally, Section 5 concludes this paper.

2 Related Work

According to [17], there are three main types of deception in texts: (i) the ones with humor, clearly for fun, using sarcasm to produce satires and parodies; (ii)

[5] http://fortune.com/2018/04/10/facebook-cambridge-analytica-what-happened/

[6] https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy/

[7] http://piaui.folha.uol.com.br/lupa/

[8] http://www.boatos.org/

[9] https://newsroom.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/

fake content, which intends to deceive people and to cause confusion; and (iii) rumors, which are non-confirmed and usually publicly accepted information. Fake content, in particular, may appear in different contexts. Fake news are a type of it, as well as fake reviews, for instance, that are tailored to harm or to promote something.

Although the recent interest growing in the area, there are several available corpora of different types of deception. In [15], the authors present three datasets related to social topics, such as opinions on abortion, death penalty, and feelings about a best friend, containing 100 deceptive and 100 truthful sentences. In [18], the authors build two datasets containing satirical and true news in four different domains (civics, science, business, and "soft" news), totalizing 240 samples. In [14], two datasets are collected on the celebrity news domain. The first one consists in emulating journalistic writing style, using Amazon Mechanical Turk, resulting in 240 fake news. The second one is collected from the web, following similar guidelines to the previous dataset (aiming to identify fake content that naturally occurs on the web), resulting in 100 fake and 100 legitimate news. Other corpora are available in English, such as the Emergent [8] and LIAR [21] corpora. For Portuguese, it is possible to find some websites that compile true and fake news for fact checking (as the ones cited in the previous section), but they often present comments about the news (and not the original texts themselves) and are not ready-to-use corpora for NLP purposes.

Some methods for detecting deceptive content have been investigated, using varied textual features, as commented by [5] and systematized in [24]. [1] studies false declarations in social networks, looking for clues of falsification (lies, contradictions and distortions), exaggeration (modifiers and superlatives), omission (lack of information, half truths) and deception (subject change, irrelevant information and misconception). [22] proposes to look at the amount of verbs and modifiers (adjectives and adverbs), complexity, pausality, uncertainty, non-immediacy, expressivity, diversity and informality features. In [15], [16] and [14], the authors compare the performance of classifiers using n-grams/bag of words, part of speech tags and syntactic information, readability metrics and word semantic classes.

Despite the efforts already made, as far as we know, there is no public and labeled dataset of fake news written in Portuguese. The absence of representative data may seriously impact the processes of development, evaluation and comparison of automatic detection methods. In what follows, we report our efforts to build the first reliable corpus in this area for Portuguese.

3 The Fake.Br Corpus

In order to create a reliable corpus, we have collected and labeled real samples written in Portuguese. The corpus – simply called "Fake.Br Corpus" – is composed of true and fake news that were manually aligned, focusing only on Brazilian Portuguese. To the best of our knowledge, there is no other similar available corpus for this language.

Collecting texts to the corpus was not a simple task. It took some months to manually find and check available fake news in the web and, then, to semi-automatically look for corresponding true news for each fake one. The manual step was necessary to check the details of the fake news and if they were in fact fake, as we wanted to guarantee the quality and reliability of the corpus.

The alignment of true and fake news is relevant for both linguistic studies and machine learning purposes, as positive and negative instances are important for validating linguistic patterns and automatic learning, depending on the adopted approach. Besides this, the alignment is a desired characteristic of the corpus, as pointed by [17], which also suggests the following for assembling the corpus: news should be in plain text format, as this is usually more appropriate for NLP; the news must have similar sizes (usually in number of words) in order to avoid bias in learning, but, if this is not the case, size normalization (e.g., text truncation) may be carried out when necessary; specification of a time period for collecting the texts, as writing style may change in time and this may harm the corpus purposes; maintenance of pragmatic factors, e.g., the original link to the news, as such information may be useful in the future for fact checking tasks [13].

Overall, we collected 7,200 news, with exact 3,600 true and 3,600 fake news. All of them are in plain text format, with each one in a different file. We kept size homogeneity as much as we could, but some true news are longer than the fake ones. We established a 2 years time interval for the news, from January of 2016 to January of 2018, but there were cases of fake news in this time period that referred to true news of a time before this. We did not consider this as a problem and kept these news in the corpus. Finally, we saved all the links and other metadata information (such as the author, date of publication, and quantity of comments and visualizations, when possible) that was available.

We manually analyzed and collected all the available fake news (including their titles) in the corresponding time period from 4 websites: *Diário do Brasil* (3,338 news[10]), *A Folha do Brasil* (190 news), *The Jornal Brasil* (65 news) e *Top Five TV* (7 news). Finally, we filtered out those news that presented half truths[11], keeping only the ones that were entirely fake.

The true news in the corpus were collected in a semiautomatic way. In a first step, using a web crawler, we collected news from major news agencies in Brazil, namely, *G1*, *Folha de São Paulo* and *Estadão*. The crawler searched in the corresponding webpages of these agencies for keywords of the fake news, which were nouns and verbs that occurred in the fake news titles and the most frequent words in the texts (ignoring stopwords). About 40,000 true news were collected this way. Then, for each fake news, we applied a lexical similarity measure (the cosine measure presented in [19]), choosing the most similar ones to the fake news, and performed a final manual verification to guarantee that the fake and true news were in fact subject-related. It is interesting to add that there were cases in that the true news explicitly denied the corresponding fake one (see,

[10] We could realize that most of the checked sites shared many fake news.

[11] Half truth may be defined as the case in which some actual facts are told in order to give support to false facts [4].

e.g., the first example in Table 1), but others were merely on the same topic (second example in Table 1).

Table 1. Examples of aligned true and fake news.

Fake	True
Michel Temer propõe fim do carnaval por 20 anos, "PEC dos gastos". Michel Temer afirmou que não deve haver gastos com aparatos supérfluos sem pensar primeiramente na educação do Brasil. A medida pretende calcelar o carnaval de 2018.	*Michel Temer não quer o fim do Carnaval por 20 anos. Notícias falsas misturam proximidade dos festejos, crise econômica e medidas impopulares do governo do peemedebista.*
Acabou a mordomia ! Ingresso mais barato pra mulher é ilegal. Baladas que davam meia entrada para mulher, ou até mesmo gratuidade, esto na ilegalidade agora. Acabou o preconceito com os homens nas casas de show de todo o Brasil.	*Ingresso feminino barato como marketing 'não inferioriza mulher', diz juíza do DF. Afirmação consta em decisão sobre preços diferentes para homens e mulheres em festa no Lago Paranoá. 'Prática permite que mulher possa optar por participar de tais eventos sociais', diz texto.*

Overall, the collected news may be divided into 6 big categories regarding their main subjects: politics, TV & celebrities, society & daily news, science & technology, economy, and religion. In order to guarantee consistency and annotation quality, the texts were manually labeled with the categories. Table 2 shows the distribution of texts by category. As expected, politics is the most frequent one.

Table 2. Amount of documents per category in the Fake.Br corpus.

Category	Number of samples	%
Politics	4,180	58.0
TV & celebrities	1,544	21.4
Society & daily news	1,276	17.7
Science & technology	112	1.5
Economy	44	0.7
Religion	44	0.7

Table 3 shows a overall comparison of the news, including the average number of tokens and sentences, as well as several other features. It is interesting to notice some differences, e.g., spelling errors were more frequent in the fake news.

Finally, we adopted the proposal of [22] to compute other linguistic features that may serve as indications of fake content, to know: (i) pausality, which checks the frequency of pauses in a text, computed as the number of punctuation signals over the number of sentences; (ii) emotiveness, which is an indication of language

Table 3. Basic statistics about the Fake.Br corpus.

Features	Fake news	True news
Avg number of tokens	216.1	1,268.5
Avg number of types (without punctuation and numbers)	119.2	494.1
Avg size of words (in characters)	4.8	4.8
Type-token ratio	0.68	0.47
Avg number of sentences	12.7	54.8
Avg size of sentences (in words)	15.3	21.1
Avg number of verbs (norm. by the avg number of tokens)	14.3	13.4
Avg number of nouns (norm. by the avg number of tokens)	24.5	24.6
Avg number of adjectives (norm. by the avg number of tokens)	4.1	4.4
Avg number of adverbs (norm. by the avg number of tokens)	3.7	4.0
Avg number of pronouns (norm. by the avg number of tokens)	5.0	5.2
Avg number of stopwords (norm. by the avg number of tokens)	31.0	32.8
Percentage of news with spelling errors	36.0	3.0

expressiveness in a message [23], computed as the sum of the number of adjectives and adverbs over the sum of nouns and verbs; (iii) uncertainty, measured by the number of modal verbs and occurrences of passive voices; and (iv) non-immediacy, measured by the number of 1st and 2nd pronouns. Table 4 shows the values of these features in the corpus. The higher differences in uncertainty and non-immediacy values are due to the size difference of the texts, as these two metrics are not normalized.

Table 4. Linguistic features of [22] in the Fake.Br corpus.

Features	Fake news	True news
Avg pausality per text	2.46	3.04
Avg emotiveness per text	0.20	0.21
Avg uncertainty per text	4.48	23.24
Avg non-immediacy per text	0.62	4.05

In what follows, we present our experiments on fake news detection using the above corpus.

4 Experiments and Results

Motivated to create an automatic classifier of fake news, we run some tests using machine learning over the Fake.Br corpus. To guarantee a fair classification, we have normalized the size of the texts (in number of words) by truncating the longer texts to the size of their aligned counterparts.

Following [16], we run the widely used SVM technique [6] (the LinearSVC implementation in Scikit-learn, with default parameters). We tried different features of [16] and [22]:

- bag of words/unigrams (simply indicating whether each word occurred or not in the text, using boolean values), after case folding, stopword[12] and punctuation removal, and stemming;
- the (normalized) number of occurrences of each part of speech tag, as indicated by the NLPNet tagger [9];
- the (normalized) number of occurrences of semantic classes, as indicated by LIWC for Brazilian Portuguese [2], which is an enriched lexicon that associates to each word one or more possible semantic classes (from a set of 64 available classes);
- and the pausality, emotiveness, (normalized) uncertainty and (normalized) non-immediacy features.

Still following the work of [16], we used an evaluation strategy of 5-fold cross-validation. We computed the traditional precision, recall and F-measure metrics for each class, as well as general accuracy. Table 5 shows the average results that we achieved for different feature sets. The first three rows refer to features of [16], while the fourth is a combination of them; the next four rows are the features of [22], also followed by their combination; we then combine the best features of both initiatives; and, finally, we combine all the features (in the last row).

Table 5. Classification results.

Features	Precision		Recall		F-Measure		Accuracy
	Fake	True	Fake	True	Fake	True	
Part of speech (POS) tags	0.76	0.74	0.73	0.77	0.74	0.76	0.75
Semantic classes	0.73	0.73	0.73	0.73	0.73	0.73	0.73
Bag of words	0.88	0.89	0.89	0.88	0.88	0.88	0.88
POS tags + semantic classes + bag of words	0.88	0.89	0.89	0.88	0.89	0.89	0.88
Pausality	0.52	0.52	0.58	0.46	0.55	0.49	0.52
Emotiveness	0.57	0.56	0.53	0.61	0.55	0.58	0.56
Uncertainty	0.51	0.51	0.46	0.57	0.48	0.54	0.51
Non-immediacy	0.53	0.51	0.16	0.86	0.24	0.64	0.51
Pausality + emotiveness + uncertainty + non-immediacy	0.57	0.56	0.53	0.60	0.55	0.58	0.57
Bag of words + emotiveness	0.88	0.89	0.89	0.88	0.89	0.89	0.89
All the features	0.89	0.89	0.89	0.89	0.89	0.89	0.89

Bag of words alone could (surprisingly) achieve good results (88% of F-measure, for both true and fake news), and other features (including the ones of [22]) did not help to significantly improve this. It is interesting that most of the methods performed similarly for the two classes.

We show in Table 6 the confusion matrix for the bag of words classification. One may see that there is still room for improvements. In our opinion, misclassifying (and, consequently, filtering out) true news is more harmful than not

[12] We also remove numeric values in order to help avoiding sparsity.

detecting some fake news (the same logic of spam detection), and this must have more attention in the future.

Table 6. Confusion matrix for bag of words classification.

		Actual classes	
		True	Fake
Classified as	True	3,192	432
	Fake	408	3,168

We checked that the classification errors are correlated with the news categories in the following way: 11.6% of the political texts were misclassified; for TV & celebrities, 10.4%; for society & daily news, 12.3%; for science & technology, 16.1%; for economy, 18.1%; and, for religion, 20.4%. Economy and religion categories appear to be the most difficult ones, but this may have happened due to fewer learning instances that we have for such categories.

We have also run some other machine learning techniques, from different paradigms, as Naïve-Bayes, Random Forest, and Multilayer Perceptron (with the default parameters of Scikit-learn). Additionally, we tried bag of words with different minimum numbers of occurrence in the corpus, as well as other values for the occurrence of words, as their (normalized) frequency (instead of boolean 0 or 1 values). Multilayer Perceptron could achieve 90% of accuracy. Considering words with at least 3 occurrences produced the same results; from 5 to more occurrences, the results start to slightly fall. Using word frequency (instead of boolean values) did not improve the results.

One final test was to run the experiments without truncating the size of the texts. The use of full texts achieved impressive 96% of accuracy with bag of words, but this classification is probably biased, as true texts are significantly longer than the fake ones.

It is interesting that, in our case, differently from [16], part of speech tags did not produce the best results. Such difference is probably explained by the dataset. While our dataset is "spontaneous" (to the extent that such nomenclature makes sense for fake news), collected from the web, [16] used a dataset of a different nature (in fact, the authors used sentences), produced by hired people to the task.

Overall, the achieved results were above our expectations. One factor that may help explaining such good results is that we have filtered out news with half truth, which might make things more complex (and equally interesting). This remains for future work, as we comment below.

5 Conclusions

To the best of our knowledge, we have presented the first reference corpus for fake news detection for Portuguese - the Fake.Br corpus. More than this, we have run some experiments, following some well known attempts in the area, and produced good results, considering the apparent difficulty of the task. We hope that our corpus may foster research in the area and that the methods we tested instigate new ones in the future.

For future work, we hope to identify other features that may help distinguishing the remaining misclassified examples, as well as to test other classification techniques, using, e.g., distributional representations and methods. We also aim at dealing with other deception types, such as satiric texts and fake opinion reviews, and with more complex cases, as the news including half truth.

More information about this work and the related tools and resources may be found at the OPINANDO project website[13].

Acknowledgments

The authors are grateful to FAPESP, CAPES and CNPq for supporting this work.

References

1. Appling, D.S., Briscoe, E.J., Hutto, C.J.: Discriminative models for predicting deception strategies. In: Proceedings of the 24th International Conference on World Wide Web. pp. 947–952 (2015)
2. Balage Filho, P.P., Pardo, T.A., Alu›sio, S.M.: An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. pp. 215–219 (2013)
3. Charles F. Bond, J., DePaulo, B.M.: Accuracy of deception judgments. Personality and Social Psychology Review **10**(3), 214–234 (2006)
4. Clem, S.: Post-truth and vices opposed to truth. Journal of the Society of Christian Ethics **37**(2), 97–116 (2017)
5. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: Methods for finding fake news. In: Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community. pp. 82:1–82:4 (2015)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
7. Duran, N.D., Hall, C., McCarthy, P.M., McNamara, D.S.: The linguistic correlates of conversational deceprion: comparing natural language processing technologies. Applied Psycholinguistics **31**(3), 439–462 (2010)
8. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1163–1168. Association for Computational Linguistics (2016)
9. Fonseca, E.R., Aluísio, S.M.: A deep architecture for non-projective dependency parsing. In: Proceedings of the NAACL-HLT Workshop on Vector Space Modeling for NLP (2015)

[13] https://sites.google.com/icmc.usp.br/opinando/

10. George, J.F., Keane, B.T.: Deception detection by third party observers. In: Paper presented at the deception detection symposium, 39th Annual Hawaii International Conference on System Sciences (2006)
11. Gimenes, G., Cordeiro, R.L., Rodrigues-Jr, J.F.: Orfel: Efficient detection of defamation or illegitimate promotion in online recommendation. Information Sciences **379**, 274 – 287 (2017)
12. Hauch, V., Blandn-Gitlin, I., Masip, J., Sporer, S.L.: Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. Personality and Social Psychology Review **19**(4), 307–342 (2015)
13. Musskopf, I.: A ciência da detecção de fake news. https://medium.com/data-science-brigade/a-ci%C3%AAncia-da-detec%C3%A7%C3%A3o-de-fake-news-d4faef2281aa (September 2017)
14. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. CoRR **abs/1708.07104** (2017)
15. Pérez-Rosas, V., Mihalcea, R.: Cross-cultural deception detection. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 440–445 (2014)
16. Pérez-Rosas, V., Mihalcea, R.: Experiments in open domain deception detection. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1120–1125 (2015)
17. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: Three types of fakes. Proceedings of the Association for Information Science and Technology **52**(1), 1–4 (2015)
18. Rubin, V.L., Conroy, N.J., Chen, Y., Cornwell, S.: Fake news or truth? using satirical cues to detect potentially misleading news. In: Proceedings of 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 7–17 (2016)
19. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)
20. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018)
21. Wang, W.Y.: “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC, Canada (2017)
22. Zhou, L., Burgoon, J., Twitchell, D., Qin, T., Nunamaker Jr., J.: A comparison of classification methods for predicting deception in computer-mediated communication. Journal of Management Information Systems **20**(4), 139–165 (2004)
23. Zhou, L., Twitchell, D.P., Qin, T., Burgoon, J.K., Nunamaker, J.F.: An exploratory study into deception detection in text-based computer-mediated communication. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003 (2003)
24. Zhou, L., Zhang, D.: Following linguistic footprints: Automatic deception detection in online communication. Communications of the ACM - Enterprise Information Integration: and other tools for merging data **51**(9), 119–122 (2008)