The Fake.Br corpus - a corpus of fake news for Brazilian Portuguese

Roney L. S. Santos, Rafael A. Monteiro, and Thiago A. S. Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
{roneysantos, rafael.augusto.monteiro}@usp.br, taspardo@icmc.usp.br

Abstract. Fake news are a problem of our time. Although they have always existed, the volume of fake news has recently increased, affecting several aspects of our lives, from healthy issues to politics. To make things worse, the scarcity of labeled datasets prevents us from training classifiers to automatically filter out such documents. In this extended abstract, inspired by previous initiatives for other languages, we briefly report our effort on building the Fake.Br corpus, composed of aligned true and fake news in Brazilian Portuguese.

Keywords: Fake news · Reference corpus · Brazilian Portuguese

1 The Corpus

In this abstract, we introduce the "Fake.Br" corpus, which is composed of aligned true and fake news written in 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. Besides this, the alignment is a desired characteristic of the corpus, as pointed by [1].

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. For this reason, we also provide a version of the corpus with size normalization, in which, for each true-fake news pair, the longer text is truncated (in number of words) to the size of the shorter aligned text. 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, title, date of publication, among others) that was available.

We manually analyzed and collected all the available fake news in the corresponding time period from 4 websites: *Diário do Brasil*, *A Folha do Brasil*, *The Jornal Brasil*, and *Top Five TV*. Finally, we filtered out those news that presented half truths[1], 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 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. About 40,000 true news were collected this way. Then, for each fake news, we applied a lexical similarity measure, 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, but others were merely on the same topic.

Finally, 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. The amount of documents per category in Fake.Br corpus were: 4,180 samples of politics, 1,544 samples of TV & celebrities, 1,276 samples of society & daily news, 112 samples of science & technology, 44 samples of economy, and 44 samples of religion.

2 Corpus Availability and Applications

The corpus is freely available in the OPINANDO project website[2]. It is presented in two versions: with the full texts and the size-normalized texts.

The OPINANDO project is a large project on opinion mining for the Portuguese language. Finding relevant documents to process and filtering out fake news (or, in more general terms, deceptive content) consist in one of the first steps in this project. Actually, initial experiments on fake news detection have already been performed over the Fake.Br corpus and good results have been achieved.

We invite the interested reader to visit the project website, where all the related information may be found.

References

1. 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)

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

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