Fake News Detection on Fake.Br Using Hierarchical Attention Networks

Emerson Yoshiaki Okano[1(✉)], Zebin Liu[2], Donghong Ji[2], and Evandro Eduardo Seron Ruiz[1]

[1] Departamento de Computação e Matemática – FFCLRP, Universidade de São Paulo, Avenida dos Bandeirantes, 3900 – Monte Alegre, 14040-901 Ribeirão Preto, SP, Brazil
{okano700,evandro}@usp.br

[2] School of Cyber Science and Engineering, Wuhan University, #299, Bayi Road, Wuchang District, Wuhan 430072, China
dhji@whu.edu.cn, 1579670185@qq.com

Abstract. Automatic fake news detection is a challenging problem in natural language processing, and contributions in this field may induce immense social impacts. This article examines the use of Hierarchical Attention Network (HAN) as a method for automatic fake news detection. We evaluate the proposed models in the Brazilian Portuguese fake news parallel corpus Fake.Br using its original full text, and also in the truncated version. We run the HAN varying the size of word embedding from 100 to 600, and by maintaining and removing the stop words. This method achieved an accuracy of 97% for full texts using the word embedding size of 600 from GloVe. However, when comparing running this method for truncated texts, this method presents similar results (90% accuracy) to the baseline established by the simple machine learning methods presented in the original presentation work of the Fake.Br (89% accuracy). Overall, keeping or removing stop words and varying the size of the word embeddings also shows a negligible advantage.

Keywords: Hierarchical attention networks · Deep learning · Text classification · Fake news

1 Introduction

Lately, the term fake news is of frequent use, a neologism, referring either as misinformation, fabricated news, or false information presented in a way that it may seem factually accurate. Lazer and colleagues define "fake news" as fabricated information that mimics news media content in form but not in organizational process or intent [8]. One of the most notorious cases of fake news in science is related to the 1998 paper by Wakefield [17], published in The Lancet, reporting a link between autism and the measles, mumps, and rubella (MMR) vaccine. As numerous other studies fail to replicate these associations, this is a real sign of falsified data in the original study. Nowadays, Wakefield's dangerous claims are

P. Quaresma et al. (Eds.): PROPOR 2020, LNAI 12037, pp. 143–152, 2020.
https://doi.org/10.1007/978-3-030-41505-1_14

still causing direct harm to children whose parents insist not to vaccinate them due to the alleged lifelong conditions imposed by autism.

Another landmark was in 2016 when the word fake news became very famous due to two decisive political events (the Brexit referendum in the UK and the presidential election in the US). These events coincide with the rise of searches for the term "fake news" in Google[1] in this year. On another hand, this buzzword becomes famous in Brazil in 2018 due to the presidential election.

Presently this fake news nightmare turned to be the high-speed phenomena carrying this fabricated news from cellphones to cellphones, from families to families. The consequences of fake news are everywhere, from the anti-vaccine movement to politics [1].

Recently there has been much effort to fight against fake news. Companies such as Facebook, Google, and Bing have joined forces to create "The Trust Project"[2], that created the Trust indicator which helps people to differentiate real news from fake news.

The content of fake news is rather diverse in terms of topics, styles, and media platforms. To help mitigate the adverse effects caused by fake news, it is of utmost importance to develop methods that automatically detect fake news on social media [14]. Although the language has no direct connection to distinguish real and fake news, legitimate and faked news articles use language differently in ways that can be detected by algorithms. The distinctive ways language is used to detect fake news have been explored by a number of research groups, such as [4, 8,9,11,14] and [16]. In this article, we employ a deep learning based automated detector approach using a three-level Hierarchical Attention Network (HAN) accurate detection of fake news. Regarding the work of Monteiro *et al.* [9] that establishes a baseline classification metric using machine learning techniques, the present work tests the effectiveness of a HAN to detect deception on written texts. We also expect that by applying a HAN model, its attention layers could highlight the most relevant sentences for the classifier.

The remainder of this article is organized as follows. In Sect. 2, we start by overviewing related works on fake news detection. In Sect. 3, we briefly discuss the datasets and explain the network model adopted. The results are presented in Sect. 4, and we conclude in Sect. 5.

2 Related Work

As James Kershner pointed out in his book [7] "So what makes fake news fake? If news refers to an accurate account of a real event, what does fake news mean?" Actually, 'fake news' has become a buzzword. Nowadays, this term does not differentiate misinformation in the media from actually fabricated news, nor a news satire from large-scale hoaxes, rumors. Only recently, on an article by Tandoc Jr. and colleagues [16], the authors clearly identified the dimensions that guided previous definitions of fake news, offering a typology based on such

[1] http://bit.ly/2og4zvV.
[2] https://thetrustproject.org.

dimensions that include news satire, news parody, news fabrication, advertising material, propaganda (referring to news stories created by a political entities), and even photo manipulation.

Earlier, Conroy, Rubin, and Chen [4] also propose a typology to asses veracity composed of methods emerging from two major categories, which are: linguistic cue approaches and network analysis approaches. They concluded proposing operational guidelines based on a hybrid approach that combine both previously cited categories. In another article [12], we understand that these same authors inaugurated a more computer methodological focused research on deception detection. They analyzed rhetorical structures, discourse constituent parts, and their coherence relations in deceptive and in real news samples. They further applied a vector space model to cluster the news by discourse feature similarity, achieving 63% accuracy.

Recently Pérez-Rosas and co-authors [11] focused on the automatic detection of fake online news. Their dualistic work consists of two novel datasets, covering seven different news topics, for fake news disclosure; and a set of learning experiments able to detect fake news achieving accuracies of up to 76%. Some of the linguistic features are also presented in the work of Okano and Ruiz [13], where they tested 76 linguistic cues on the Fake.Br [9] parallel corpus and achieved accuracy results of 75% using support vector machines (SVM), random forest, and logistic regression. We address the work of Monteiro and colleagues [9] about the construction of the Fake.Br, the first Brazilian Portuguese parallel fake news corpus, later, on Sect. 3.1.

One of the first attempts to use HAN in social media was aimed to detect rumor in microblogs was reported by [5]. They build a hierarchical bidirectional long short-term memory (BLSTM) model for representation learning. The social contexts were incorporated into the network via attention mechanism in a way that that important semantic information was introduced to the framework for robust rumor detection. They achieved 94% using F1-measure on rumorous Weibo microblogs and 82.5% on tweets.

Hierarchical Attention Networks were first introduced in 2016 by Yang and collaborators [19], a partnership between Carnegie Mellon and Microsoft. As they emphasize, they propose this novel structure that mirrors the hierarchical structure of documents. This model incorporates the knowledge of document structure in the model architecture. It also has two levels of attention mechanisms that, when applied to the word and sentence-level, enable more or less importance to the content when constructing the document representation. They applied this new neural network on six large scale document classification data sets for two classification tasks: sentiment estimation and topic classification, and concluded that the proposed architecture outperforms previous methods by a significant margin.

Following Yang's article, we call attention to the work of Yang, Mukherjee, and Dragut [18] who used an attention mechanism and some linguistic features to recognize satirical cues at the paragraph level. They investigated the difference between paragraph-level features and document-level features and revealed

what features are essential at which level. Singhania and collaborators [15] also have tested this three-level hierarchical attention network, which they called 3HAN, but aiming for accurate detection of fake news. By their experiments, they argue that despite other deep learning models, the 3HAN model provides an 'understandable output through the attention weights given to different parts of an article'. In this paper, we innovate by proposing a HAN for document classification written in Brazilian Portuguese as disjoint labels of fake or real news, and also by adopting a categorical cross-entropy as loss function that measures the performance of the classification model in the $[0, 1]$ interval.

3 Dataset and Methods

In this section, we discuss the available dataset and present the algorithms used for fake news detection. We also focus on evaluation metrics for this task.

3.1 Fake.Br Corpus

The Fake.Br Corpus [9] is the first fake news reference corpus for Portuguese. It is composed of aligned pairs of authentic and fake news. This corpus is composed of 7,200 news, where 3,600 are fake news, and 3,600 are their real correspondent news. Table 1 shows the distributions of the news in each category.

Table 1. Sampling distribution per category in Fake.BR. Adapted from Monteiro et al. [9]

Category	Samples (# of)	%
Politics	4180	58.0
TV & Celebrities	1544	21.4
Society & Daily News	1276	17.7
Science & Technology	112	1.5
Economy	44	0.7
Religion	44	0.7

The authors manually collected and analyzed all the fake news from an interval of two years (from 01/2016 to 01/2018) from 4 websites, which are: *Diário do Brasil*, *A Folha do Brasil*, *The Journal Brasil* and *Top Five TV*, filtering out the news that presented half-truth. On the other hand, the authors collected the real news using a web crawler from three major Brazilian news agencies: *G1*, *Folha de São Paulo*, and *Estadão*. After collecting the news, the authors used a lexical similarity measure to choose the most similar real news to the fake news collected. They also performed a manual verification to guarantee that the fake news and real news were subject related.

As the average number of tokens $\bar{t}$ in each corresponding class in Fake.Br is so big, fake news $\bar{t} = 216.1$ tokens and true news $\bar{t} = 1268.5$ tokens, the authors of the Fake.Br also offered a truncated version of the corpus in which both real and fake news, have the same number of characters based on the length (number of tokens) of the shortest of the pair. The HAN was applied to both corpora, the full, and the truncated one.

3.2 Hierarchical Attention Network

We adopt a general HAN architecture for document representation, displayed in Fig. 1, initially proposed by Yang [19]. The overall architecture is represented in two major parts: a word-level attention layer, and a sentence-level attention layer, therefore representing two levels of abstraction. They are both preceded by the corresponding encoders, a word encoder, and a sentence encoder.

Word encoder. Given a sentence with words $w_{it}, t \in [1, T]$ for T maximum number of words per sentence, they are embedded to vectors x_{iT}, see Eq. 1. This model uses a bidirectional GRU network where the words are summarized in two directions, the forward hidden state $\overrightarrow{h_{it}}$ and and backward hidden state $\overleftarrow{h_{it}}$, (Eqs. 2 and 3) therefore summarizing the information of the whole sentence centered around w_{it}, as in Eq. 4.

Word attention layer. The attention mechanism devised in the next phase (seen as α_w in Fig. 1), extracts the most meaningful words to contribute to the sentence meaning. In other words, it gives weights to words considering the importance α_{it} of the word w_{it} to the sentence s_i. In a summary, u_{it} resulted form a one layer MLP, as seen in Eq. 5. Equation 6 represents the importance of the word, which is calculated using the softmax function to get the normalized importance weight α_{it}. Equation 7 refers to the output of the word attention layer, which is the sentence vector s_i, calculated as the weighted sum of the words annotation.

The following equations summarize both processes:

$$x_{it} = W_e w_{it} \quad (1)$$

$$\overrightarrow{h_{it}} = \overrightarrow{\text{GRU}}(x_{it}), \quad (2)$$

$$\overleftarrow{h_{it}} = \overleftarrow{\text{GRU}}(x_{it}), \quad (3)$$

$$h_{it} = [\overrightarrow{h_{it}}, \overleftarrow{h_{it}}] \quad (4)$$

$$u_{it} = \tanh(W_s h_i + b_s) \quad (5)$$

$$\alpha_{it} = \frac{\exp(u_{it}^\top u_s)}{\sum_t \exp(u_{it}^\top u_s)} \quad (6)$$

$$s_i = \sum_t \alpha_{it} h_{it} \quad (7)$$

Recalling that $t \in [1, T]$ for all equations above.

Sentence encoder. Similarly to the word encoder layer, we have the sentence encoder, which uses an analogous mechanism. Here we have sentence $s_i, i \in [1, L]$, for L the max number of sentences in a document. Equations 8, 9 and 10 represent the sentence encoder mechanism and, similarly, with word encoder. Here h_i stands for the information of the whole document centered around s_i.

Sentence attention. A similar attention mechanism was implemented to reward the sentences that are more important to the classification. Equations 11, 13 can summarize the sentence attention layer.

$$\overrightarrow{h_i} = \overrightarrow{\text{GRU}}(s_i), i \in [1, L] \quad (8)$$

$$\overleftarrow{h_i} = \overleftarrow{\text{GRU}}(s_i), i \in [L, 1] \quad (9)$$

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \quad (10)$$

$$u_i = \tanh(W_s h_i + b_s) \quad (11)$$

$$\alpha_i = \frac{\exp(u_i^\top u_s)}{\sum_i \exp(u_i^\top u_s)} \quad (12)$$

$$v = \sum_i \alpha_i h_i \quad (13)$$

Document classification. To classify the document we apply the document vector v through a softmax function as shown in Eq. 14, where W_c is the weight vector and the b_c is the bias vector.

$$p = \text{softmax}(W_c v + b_c) \quad (14)$$

In our work, we used the categorical cross-entropy, Eq. 15, as loss function. There the loss, L, is calculated as the sum of the entropy of all classes M, where y_c is the class label, and p_c is the predicted label.

$$L = -\sum_{c=1}^{M} y_c \log p_c \quad (15)$$

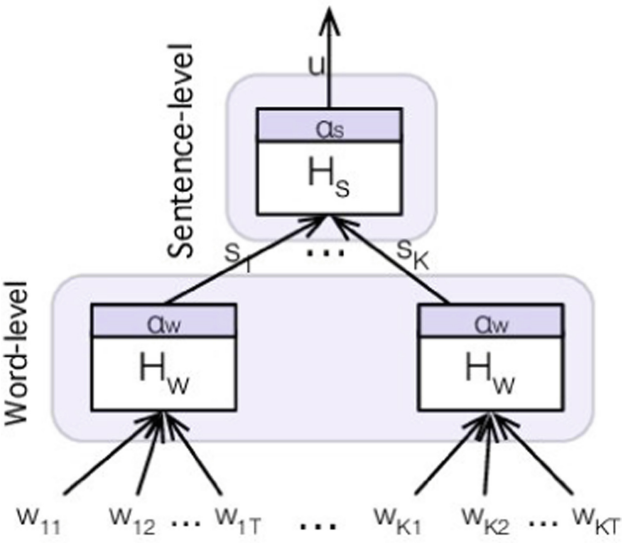

Fig. 1. General architecture of hierarchical attention neural networks for modeling documents. Reprinted from Pappas Popescu-Belis [10]

3.3 Model Configuration

Initially, we performed the preprocessing of the text, converting it to lowercase and removing or not the stop words(NLTK stop words). We then split the documents into sentences using a sentence tokenizer function (NLTK [2]). Each sentence was also split into tokens using the Keras [3] tokenizer.

In the word embedding layer, we used Hartmann et al. [6] pre-trained Global Vectors (GloVe). Actually trained in a large Portuguese multi-genre corpus. In our experiments, we trained models utilizing GloVe with different dimensions ($d = \{100, 300, 600\}$).

For training, we used a batch size of 400, using the Keras Adam optimizer using the default parameters. The models were trained using this configuration during 100 epochs or until the loss did not decrease by at least 10^{-3} for ten successive epochs.

To evaluate the performance of the Hierarchical Attention Network in the task of detection of fake news, we used the same methodology used by Monteiro et al. [9], running the algorithm in two setups: using full texts and using truncated texts. We used the 5-fold cross-validation to evaluate the algorithm calculating the precision, recall, and F-measure metrics for each class, as well as overall accuracy.

4 Results

We performed our experiments using the full texts and the truncated texts. We chose the truncated datasets due to the difference in the number of tokens between the two classes. The average number of tokens in real news is $1,268.5$, and in fake news is 216.1.

In Table 2, we show the results obtained using the full-text dataset, where in the first column we present the setup used where GloVe x are the results obtained using GloVe of x dimensions and SW we removed all stop words. Looking at the results, we can observe that we obtained almost the same result using all the different configurations we proposed: Removing or not the stop words and using different numbers of dimensions of the GloVe model.

Using this model in the full texts, we obtained 97% of accuracy. Considering the significant difference in the number of tokens between the two classes to make this classification task more manageable, we checked the performance of this model in the truncated texts too.

In Table 3, we present the results using the truncated texts. There we can observe that the results obtained using HAN were close, or slightly better, than the results obtained by Monteiro et al. [9], when they mainly used bag of words to obtain an accuracy of 89% in both, fake and real class. One advantage of this model, to be pursued during future studies, is that we can extract an attention map from the attention layers showing the importance of each sentence and each word.

Table 2. Results of the HAN applied on full texts.

	Precision		Recall		F-score		Accuracy
	Fake	True	Fake	True	Fake	True	
GloVe 100	0.9655	0.9695	**0.9694**	0.9653	0.9674	0.9673	0.9674
GloVe 100 SW	0.9696	0.9532	0.9522	0.9700	0.9608	0.9615	0.9611
GloVe 300	0.9666	0.9622	0.9619	0.9667	0.9642	0.9644	0.9643
GloVe 300 SW	0.9672	0.9640	0.9639	0.9672	0.9655	0.9656	0.9656
GloVe 600	**0.9706**	**0.9696**	**0.9694**	**0.9706**	**0.9700**	**0.9700**	**0.9700**
GloVe 600 SW	0.9655	0.9632	0.9631	0.9656	0.9643	0.9643	0.9643

Table 3. Results of the HAN applied on truncated texts.

	Precision		Recall		F-score		Accuracy
	Fake	True	Fake	True	Fake	True	
GloVe 100	0.8966	**0.9038**	**0.9044**	0.8956	0.9004	0.8996	0.9000
GloVe 100 SW	0.8801	0.8884	0.8892	0.8769	0.8840	0.8819	0.8831
GloVe 300	0.9035	0.9018	0.9017	0.9036	0.9026	0.9027	0.9026
GloVe 300 SW	0.8808	0.8983	0.9003	0.8764	0.8899	0.8865	0.8883
GloVe 600	**0.9094**	0.9026	0.9014	**0.9103**	**0.9053**	**0.9064**	**0.9058**
GloVe 600 SW	0.8999	0.8990	0.8989	0.9000	0.8994	0.8995	0.8994

5 Conclusion

We believe the enormous spread of fake news came along with the increasing popularity of social media. In this article, we explored the problem of fake news detection by reviewing the hierarchical attention networks applied to texts written in Brazilian Portuguese. Previously the HAN model has been successfully applied to sentiment analysis [19], where the attention layers were able to select qualitatively informative words and sentences. We expected a similar behavior for this model when applied to distinguish between fake and real news, which did not happen for this corpus. Besides the narrow margin of contrast between fake and real news presented by this deep learning model, this architecture poses a high computing cost to obtain similar results as other comparable lower-cost architectures such as machine learning models [9].

For future work, we want to use the metadata provided in Fake.Br corpus to improve the classification results and also explore the attention map extracted from these texts to verify the importance of each word and sentence to the classification of fake and true news, we also want to use this model in another classification task as well as sentiment analysis in Portuguese texts.

Acknowledgements. This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process number 2018/03129-8.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**(2), 211–236 (2017)
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc., Sebastopol (2009)
3. Chollet, F., et al.: Keras (2015). https://keras.io
4. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015)
5. Guo, H., Cao, J., Zhang, Y., Guo, J., Li, J.: Rumor detection with hierarchical social attention network. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, pp. 943–951. ACM, New York (2018). https://doi.org/10.1145/3269206.3271709.https://doi.acm.org/10.1145/3269206.3271709
6. Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., Aluisio, S.: Portuguese word embeddings: evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025 (2017)
7. Kershner, J.W.: Elements of News Writing. Allyn and Bacon, Boston (2004)
8. Lazer, D.M., et al.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
9. 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. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 324–334. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99722-3_33
10. Pappas, N., Popescu-Belis, A.: Multilingual hierarchical attention networks for document classification. arXiv preprint arXiv:1707.00896 (2017)
11. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R.: Automatic detection of fake news. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3391–3401. Association for Computational Linguistics, Santa Fe, August 2018. https://www.aclweb.org/anthology/C18-1287
12. Rubin, V.L., Conroy, N.J., Chen, Y.: Towards news verification: deception detection methods for news discourse. In: Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium. Grand Hyatt, Kauai (2015). https://doi.org/10.13140/2.1.4822.8166
13. Ruiz, E.E.S., Okano, E.Y.: Using linguistic cues to detect fake news on the Brazilian Portuguese parallel corpus Fake.BR. In: Proceedings of the 12th Brazilian Symposium in Information and Human Language Technology (2019)
14. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
15. Singhania, S., Fernandez, N., Rao, S.: 3HAN: a deep neural network for fake news detection. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (eds.) ICONIP 2017. LNCS, vol. 10635, pp. 572–581. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70096-0_59
16. Tandoc Jr., E.C., Lim, Z.W., Ling, R.: Defining "fake news": a typology of scholarly definitions. Digit. Journal. **6**(2), 137–153 (2018)
17. Wakefield, A.J.: MMR vaccination and autism. Lancet **354**(9182), 949–950 (1999)

18. Yang, F., Mukherjee, A., Dragut, E.: Satirical news detection and analysis using attention mechanism and linguistic features. arXiv preprint arXiv:1709.01189 (2017)
19. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)