Fake News Classification and Topic Modeling in Brazilian Portuguese

Maik Paixão
Departamento de Computação
Universidade Federal Rural de Pernambuco
Recife, Brazil
maik.paixao@ufrpe.br

Rinaldo Lima
Departamento de Computação
Universidade Federal Rural de Pernambuco
Recife, Brazil
rinaldo.jose@ufrpe.br

Bernard Espinasse
LIS-UMR CNRS
Aix-Marseille Université
Marseille, France
bernard.espinasse@lis-lab.fr

WI-IAT 2020

***Abstract*—All over the world, people receive daily news on many subjects through web-based information sharing platforms such as social networks. However, some of such news are false (fake) with the potential to deceive them. Thus, the automatic detection of false news is a major issue and is gaining careful attention from the scientific community. In this paper, we present experimental analysis using both supervised and unsupervised learning on the Fake.Br corpus, a fake news dataset in Brazilian Portuguese. We propose a classification method for fake news detection based on distinct types of features, and deep learning supervised algorithms. Our best classification model achieved F1 scores up to 96% and was compared with other non-deep learning classifiers. Furthermore, we provide a complementary analysis of the same dataset by performing topic modeling based on both uni-grams and bi-grams.**

***Index Terms*—fake news detection, topic modeling, machine learning**

I. Introduction

People receive daily news on many topics every day by means of web-based information sharing platforms. However, some of this information can be false (fake) having the potential to deceive them and even affecting their lives in many aspects. According to [14], fake news can be any content that is not truthful and is generated to convince or deceive their readers to believe in something that is not true. The term ”Fake News” was popularized mainly in the 2016 USA elections according to some political experts [2]. Since then, the automatic detection of fake news has been the subject of many studies [1], [8], [13], [4].

To mitigate this problem, this paper aims at providing two experimental analyses based on supervised and unsupervised learning settings on the the Fake.Br corpus, a fake news dataset in Brazilian Portuguese. We propose a classification method for fake news detection based on both distinct types of features and supervised machine learning algorithms. The classification results obtained by our models are compared with those ones found in related work based on non-deep learning classifiers evaluated on the same dataset. Furthermore, we provide a complementary analysis on Fake.br dataset by performing topic modeling based on uni-gram and bi-gram representation of the documents. This analysis provides not only the topics present in both real and fake news but also novel insights.

The contributions of this paper are three-fold:

- (i) it provides classical and deep learning-based classification models on fake news data for the Brazilian Portuguese which relies on distinct types of features (TF-IDF, POS tags, psycholinguistics, and word embeddings). Our best classifier achieves state-of-the-art performance on this dataset.
- (ii) a novel topic modeling analysis on the Fake.Br dataset based on uni-grams and bi-grams that highlights some major aspects differing fake news from real ones.
- (iii) the implementation of the proposed method for Fake news detection in Brazilian Portuguese publicly available for research proposes [1].

The rest of this paper is organized as follows: Section 2 describes related work. Section 3 presents the proposed supervised method for fake news classification. Section 4 presents our experimental evaluation. Section 5 provides another analytical perspective of the Fake.Br corpus based on topic modeling using both uni-grams and bi-grams document representations. Section 6 concludes the paper and points out future work.

II. Related Work

Many works for fake news detection based on supervised classification have already been proposed in the literature. They typically employ word-level feature extraction methods, including BOW, N-grams, syntactic, and linguistic features [13], [1], [4], [8].

Reference [16] shows that computational methods can be employed to both recognize fake news and exploit the differences in writing style, language, and sentiment. The author describes that even though features like N-gram and POS tags are effective in classifying Fake News, they are less useful in capturing deeper syntactic, and semantic features.

The authors in [9] present an analysis of the correlation between the performance of the machine learning classifiers and the length of the fake news. The investigation using the Naive Bayes classifier concluded that the accuracy of the model is proportional to the average length of the news texts,

[1] https://bit.ly/36ZgHE4

suggesting that an increase in the length of the articles can lead to an increase in the classifier performance.

Another work [18] relies on Deep Learning achieving state-of-the-art performance in the fake news classification task. It presents a fake news detection method based on a Convolutional Neural Network (CNN) trained on dataset providing on image and text achieving a f-measure score of 0.92.

Monteiro et al. (2018) [13] introduced Fake.Br, a benchmark corpus for fake news detection in Brazilian Portuguese language. The authors best SVM classification model employs many types of features achieving 0.89 accuracy. No parameter optimization was reported.

Our work differs from [13] in the sense that we conduct a more in-depth qualitative and quantitative analysis on the same dataset. Indeed, we investigate the performance of several machine learning algorithms as well as hyperparameter optimization. To the best of our knowledge, we are the first to provide both new insights and discussions concerning the topics found in the Fake.Br dataset using the unsupervised Topic Modelling technique.

III. Proposed Method for Detecting Fake News

The architecture of the proposed method for fake news classification is depicted in Fig. 1. This pipeline process is divided into four basic steps: (i) text preprocessing, (ii) feature extraction, (iii) model generation (training), and (iv) prediction.

Firstly, the news dataset is preprocessed by natural language processing tools performing tokenization, stopwords removal, and lemmatization. Then, the feature extraction step creates both relevant features from the input texts and their labels, representing them as feature vectors. Finally, the feature vectors are given as training input data to supervised machine learning algorithms that build a predictive model. The last component performs the prediction on unlabeled news to detect whether is real of fake news.

In the remainder of this section, the components shown in Fig. 1 are described in more detail.

A. Text Preprocessing

We rely on Spacy [2] for performing the following tasks (in this order): *tokenization*, lowercase, *stopwords removal*, and *lemmatization*. Such operations are commonly employed to remove inflectional affixes of words, as well as, to improve the overall classification results since distinct word derivations of a given word are converted to only one term in the feature vectors.

B. Features Extraction

The feature extraction step in our work was inspired by [13] [18] [11]. The following features were used: Bag-of-Words (BOW), Term Frequency-Inverse Term Frequency (TF-IDF), Part-of-Speech tags (POS-TAG), (psycho)-linguistic features (LIWC), and Word Embeddings. In addition, both uni-gram and bi-gram tokens are generated as dimensions in the feature vectors. For the sake of space, we describe some of the aforementioned features next.

Part-of-Speech Tagging. POS Tagging is the process of mapping a word to its syntactic function in a sentence (e.g., nouns, verbs, etc.) based on both its lexical elements and context. There are several reasons to use POS tags as features for fake news classification as pointed out in [13] [16]. For each news, the normalized frequency of each POS tags provided by Spacy is used.

Pylinguistic. The Pylinguistics [6] open source tool that provides up to 25 functional understandability metrics for Portuguese texts was used in this work. Such metrics include Noun Incidence, Adverb Incidence, Text Readability, Lexical Diversty, Content Diversity, among others. Yet, it is well known that texts differ in degrees of complexity. For instance, the scientific journalistic domain exhibit some typical properties, such as relative abstractness, technicality, and informational density while the journalistic genre, targeting to general public audience usually presents a higher incidence of nouns and verbs that would decrease the comprehension difficulty of a document [6].

Linguistic Inquiry and Word Count (LIWC). LIWC [7] is the Brazilian Portuguese version of the lexicon in the Linguistic Inquiry and Word Count tool. It performs text analyses calculating the degree to which people use different categories of words in many types of documents. The Pt-Br LIWC dictionary has 127,149 entries, each one assigned to one or more classes.

Word Embedding (WE). To feed our neural networks, we adopted the same token embedding input layer implemented in the Keras API [3] and adopted by [18]. In this API, the *TextToSequence* function tokenizes and represents the set of all distinct words (vocabulary) as indexes. Thus, a sentence is represented as a sequence of indexes. Next, the *EmbeddingLayer* function transforms the index sequence to word embeddings (300d) initialized with random weights and adjusted during training phase via backpropagation. In fact, we had experimented with other pretrained WE including Glove and Word2Vec, but their results were not satisfactory.

C. Classification: Model Generation and Prediction

This work adopts the supervised machine learning approach to classify Fake News. This approach comprises two phases: (i) *model generation*, in which a labeled dataset is used as input to a supervised machine learning algorithm for building the classification models that comprise the patterns extracted from the input dataset. Typically, such algorithms can be fine-tuned to a given dataset by a previous step of parameter optimization. We adopted the Grid Search technique for parameter optimization. (ii) *prediction*, in which the built classification model is applied on new (unlabeled) examples to infer its class. This work focuses on binary classification, i.e., the classification models are trained on examples from two distinct classes.

[2] https://spacy.io/models/pt

[3] https://keras.io/api/preprocessing/text/

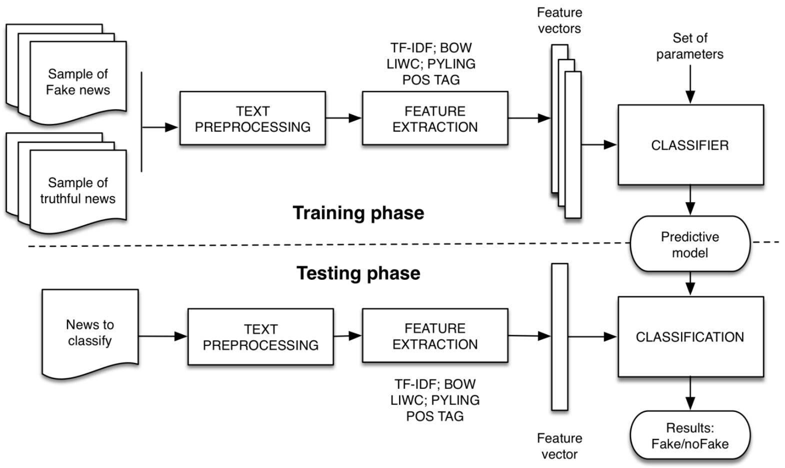


Fig. 1. Fake News Detection Architecture. Rectangles represent processes while Ovals represent intermediate results.

IV. Experimental Evaluation on Fake News Classification

This section presents our first contribution concerning an effective method for classifying fake news. Our method is based not only on several types of features but also performs a hyperparameter optimization step.

A. Experimental Setting

Dataset Preprocessing and Statistics. We used a normalized version of the Fake.Br dataset generated following the same procedure found in [3]. This normalization process deals with a significant size difference between real and fake news in Fake.Br corpus. Such text size differences would cause biased results because the size of the real news texts was significantly longer than the fake news ones. This normalization is accomplished by truncating all the news documents to obtain their new versions with approximately the same size. According to [13], such a normalization process enables a more robust evaluation methodology as well as more reliable classification results. Table I shows the fake news distribution according to its six categories. The great majority of the news texts concerns ”Politics” (58%), followed by ”TV and celebrities” (21%). For more details statistics about the Fake.Br dataset, refer to [13].

TABLE I
Category distribution of the news in the Fake.Br corpus.

Category	Number of samples	%
Politics	4180	58.0
TV Shows	1544	21.4
Daily News	1276	17.7
Technology	112	1.5
Economy	44	0.7
Religion	44	0.7
Total	**7200**	100

Evaluation Methodology and Metrics For a fair comparison with related work, we follow the same evaluation methodology presented in [13]. The classification models are evaluated using 5-fold cross-validation using the metrics accuracy (A), Precision (P), Recall (R), and F1-score (F1).

Selected Machine Learning Algorithms. The following supervised machine learning algorithms were selected: Decision Tree, Random Forest, Naive Bayes, Linear Support Vector Machine, Logistic Regression (LR), eXtreme Gradient Boosting , and Convolutional Neural Networks (Deep Learning). We adopted the implementation of these algorithms available in scikit-learn [4], XGBoost[5], and Keras [6] libraries.

Optimization of (hyper)parameters. Table III summarizes all the tested (hyper)parameters. The best ones are in bold. We have chosen the most sensible parameters that could impact most on the classification results. Particularly for the CNN classifier, we tried different hidden layer dimensions [100, 200, 300] dimensions (see Table II). Next section presents our results and discussions.

TABLE II
CNN model architecture including its layers, shape and number of parameters. Our network used one convolution layer with 100 feature maps and kernel size of **3** [13]

Layers	Shape
embedding	(200, 300)
conv1d	(198, 100)
pooling1d	(66, 100)
flatten	(6600
dense	(1)

B. Results and Discussion

In the rest of this section, we present the results of two experiments. The first one consists in the evaluation of all

[4] https://scikit-learn.org/

[5] https://github.com/dmlc/xgboost

[6] https://keras.io/api

TABLE III
COMPARATIVE ANALYSIS OF THE BEST RESULTS OF OUR WORK AND [13]

Algorithm	Hyperparameter	Values
Linear SVM	param_c	0.1, 1, 10, **100**
	max_iter	500, **1000**
Linear Regression	param_c	0.1, 1, 10, **100**
	max_iter	**500**, 1000, 2000
Naive Bayes	binarize	**0.2**, 0.4, 0.6
	alpha	0.1, 0.3, **0.5**
Decision Tree	max_depth	10, 20, 30, **40**
	min_split	2, 3, 5, **10**
Random Forest	max_depth	10, 20, 30, **40**
	n_estimators	10, 50, **100**
XGBoost	max_depth	4, 5, **6**
	n_estimators	200, 300, **400**
CNN	batch_size	25, 50, **100**
	epochs	10, **20**, 30

possible combinations of the selected features and classifiers presented earlier. The second experiment aims at performing a fair comparison between our work and [3], which is, to the best of our knowledge, the only work that used the Fake.Br dataset so far.

Comparison with classification algorithms. Table IV summarizes the best results among all possible combinations of feature types (T = TF-IDF, L = Linguistics, P = POS tags, Py = Pylinguistic) and classifiers. Differently from work [13] that did not employ TF-IDF-based features, the best classification results in our work were obtained using this type of feature. In fact, four of seven models have some combination of features integrating TF-IDF. The CNN classification model based on Word Embedding features has been the best alternative to classify fake news in our experiments. Indeed, it achieved up to 0.95 accuracy and 96.37 F1-score, which outperforms all the other classifiers. This represents an increase of more than 5% when compared to our optimized Linear LSVM model. Both linear classifiers (SVM and Logistic Regression) were the best classification models among the non-deep learning classifiers. Surprisingly, the combination (Naive Bayes model + BOW + POS) yielded practically the same performance score compared to a more sophisticated supervised learning algorithm such as Random Forest. This is another evidence showing that Naive Bayes classifier is still a good baseline classifier to be taken into consideration in comparative evaluations against other more complex supervised machine learning algorithms.

TABLE IV
THE BEST CLASSIFICATION RESULTS OBTAINED FOR EACH OF THE SIX CLASSIFIERS.

Algorithm	Features	A	P	R	F1
CNN	WE	0.9474	0.9345	0.9934a	0.9637
SVM	T + L + P	0.9164	0.9071	0.9134	0.9102
Regression	T + P	0.9137	0.9069	0.9220	0.9144
XGBoost	T + Py	0.8974	0.9081	0.8891	0.8985
Random F.	BOW	0.8745	0.8673	0.8800	0.8736
Naive B.	BOW + P	0.8684	0.8848	0.8567	0.8705
Decision T.	T + Py	0.7860	0.7699	0.7955	0.7824

Comparison with Related Work. Table V presents the results of the best classifier proposed in [13] that employed the Linear SVM classier and a set of features including BOW, POS tag, LIWC, and more linguistic-based features (emotiveness, uncertainty, etc.). Using the same classifier, but including TF-IDF-based features, our model was slightly superior considering all the evaluation metrics. This suggests that the majority of the linguistic-based features used in [13] might introduce some noise that negatively impacts the classifier performance on the Fake.Br dataset. For this reason, we decided not to use such kind of features in our work. Finally, the classifier based on CNN combined with our customized word embedding outperforms all the others.

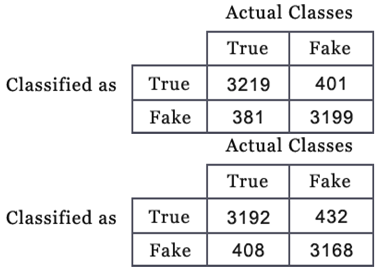
Fig. 2 depicts a comparison of two confusion matrices corresponding to the best classification models using BOW features in both works. The authors in [13] claim that misclassifying real news is more harmful than not detecting some fake news. Our XGBoost classifier mitigates precisely this problem since the misclassifications corresponding to false positives have considerably decreased. It was able to reduce true negative errors as well.

TABLE V
COMPARATIVE ANALYSIS OF THE BEST RESULTS BETWEEN OUR WORK AND [13]

Algorithm	Features	A	P	R	F1
CNN	WE	0.9474	0.9345	0.9934	0.9637
SVM (ours)	T + L + P	0.9164	0.9071	0.9134	0.9102
SVM [13]	All	0.8913	0.8975	0.8988	0.8934

V. DISCOVERING TOPICS IN FAKE.BR DATASET

This section provides qualitative analysis of real and fake news of the Fake.BR dataset. For that, we perform topic modeling, an unsupervised machine learning technique which find the most relevant topics. This analysis allowed us to obtain an insightful analysis distinguishing between real and fake news. We employed the Latent Dirichlet Allocation algorithm [5] for the automatic discovery of topics in an unsupervised manner. For the best of our knowledge, this is the first topic modeling analysis performed on the Fake.BR dataset.



		Actual Classes	
		True	Fake
Classified as	True	3219	401
	Fake	381	3199

		Actual Classes	
		True	Fake
Classified as	True	3192	432
	Fake	408	3168

Fig. 2. Comparison of the confusion matrices between our BoW-XGBoost classifier (above) and the BoW-LSVM classifier from [3].

A. Optimal number of topics

For determining the optimal number of topics in our dataset, we employed the coherence model pipeline in [15] which produces topic models based on segmentation, probability estimation, confirmation measure, and aggregation. The coherence measure describes the quality of a set of words fitted to some topics, i.e., a higher coherence score correlates to topics with more comprehensibility with regard to its constituent words. The optimal number of topics found for our dataset was 6.

B. Distribution of word probability of the topics in news.

Figures 3 and 4 depict the distribution of the top ten words (highest probability) in all the real and fake news documents, respectively. Such distributions are rooted in the distributional hypothesis, i.e., it is assumed that words occurring in similar documents have similar meanings. Therefore, it is acceptable to claim that documents with a similar set of words fit within the same topics. [5] It can be seen that the Topic 1 in Fig. 3 and Topic 2 in Fig. 4 are similar as they share several terms except for a few words and the order in which they appear. Both denote topics concerning justice trials.

Reference [9] discusses how unreliable sources of news misinterpret quotes and citations to make them as believable as possible. According to that study, words like ”said” are highly frequent in fake news. Interestingly, our topic modeling analysis corroborates this finding in the sense that the word ”said” (its lemma ”say”) are present in three of the six topics in Fig. 4. Indeed, the lemma ”say” is the first, second, and third word more frequent in Topics 5, 3, and 2, respectively. This shows that taking the sum of all its probability, this lemma is among the most frequent in all the documents of the Fake.Br dataset. On the contrary, the same lemma is found only once in Topic 0 in Fig. 3 (real news). Furthermore, a project[7] led by researcher Fatemeh T. Asr at Simon Fraser University in Canada states that, on average, fake news often employ words related to sex and death. Indeed, this is evidenced by our topic modeling analysis (Fig. 5) shows that Topic 4 has the term ”death” or other terms related to it (illness and cancer). However, none of such words are present in any of the topics in real news.

VI. Conclusion and Future Work

This work presented two distinct machine learning tasks concerning the empirical evaluation of many classifiers models for the fake news detection and topic modeling, respectively. Our feature engineering step combined with state-of-the-art supervised machine learning algorithms achieved superior performance in fake news classification compared to related work on the same dataset. In addition, we provided a novel detailed analysis of the dominant topics found in the Fake.Br dataset with the aim of highlighting the major difference between real and fake news. As future work, we intend to improve the current work by testing it on larger datasets (scalability analysis), and dimensionality-reduction techniques like UMAP [10]. Furthermore, based on the encouraging results described in recent studies using Deep Learning to classify texts [3], [12], we intend to make use of WE trained with Hierarchical Attention Network (HAN) [17] as an attempt to obtain even better classification models. Finally, we also aim to validate our classification models using other fake news datasets in different languages.

[7] https://bit.ly/36MOrWL

References

[1] Hadeer Ahmed, Issa Traore, and Sherif Saad. *Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques*. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. 2017.

[2] Hunt Allcott and Matthew Gentzkow. *Social Media and Fake News in the 2016 Election*. Journal of Economic Perspectives. 2017.

[3] Fabrício Benevenuto, Julio C. S. Reis, and Adriano Veloso. *Supervised Learning for Fake News Detection*. Intelligent Systems, IEEE. 2019.

[4] Gaurav Bhatt et al. *Combining Neural, Statistical and External Features for Fake News Stance Identification*. 9th International Workshop on Modeling Social Media. 2018.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent Dirichlet Allocation*. Journal of Machine Learning Research. 2003.

[6] Suya Castilhos et al. *Pylinguistics: an open source library for readability assessment of texts written in Portuguese*. Information Systems Magazine of FSMA. 2016.

[7] Pedro P. Balage Filho, Thiago A.S Pardo, and Sandra M. Aluísio. *An Evaluation of the Brazilian Portuguese LIWC Dictionary for SentimentAnalysis*. 9th Brazilian Symposium in Information and Human Language Technology. 2013.

[8] Mykhailo Granik and Volodymyr Mesyura. *Fake News Detection using Naive Bayes classifier*. IEEE First Ukraine Conference on Electrical and Computer Engineering. 2017.

[9] Junaed Younus Khan et al. *A Benchmark Study on Machine Learning Methods for Fake News Detection*. arXiv preprin arXiv:1905.04749v1. 2019.

[10] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv preprin arXiv:1802.03426. 2018.

[11] Tomas Mikolov, Ilya Sutskever, and Kai Chen. “Distributed Representations of Words and Phrases and their Compositionality”. In: (2013).

[12] Shervin Minaee et al. *Deep Learning Based Text Classification: A Comprehensive Review*. CoRR. 2020.

[13] Rafael A. Monteiro et al. *Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results*. Computational Processing of the Portuguese Language. 2018.

[14] Shivam B. Parikh and Pradeep K. Atrey. *Media-Rich Fake News Detection: A Survey*. IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). 2018.

[15] Michael Roder, Andreas Both, and Alexander Hinneburg. *Exploring the Space of Topic Coherence Measures*. 8th ACM International Conference on Web Search and Data Mining. 2015.

[16] Karishma Sharma, Feng Qian, and He Jiang. *Combating Fake News: A Survey on Identification and Mitigation Techniques*. arXiv preprin arXiv:1901.06437. 2019.

[17] Koustuv Sinha, Jackie C.K. Cheung, and Derek Ruths. *A Hierarchical Neural Attention-based Text Classifier*. Conference on Empirical Methods in Natural Language Processing. 2018.

[18] Yang Yang et al. *TI-CNN: Convolutional Neural Networks for Fake News Detection*. CoRR. 2018.

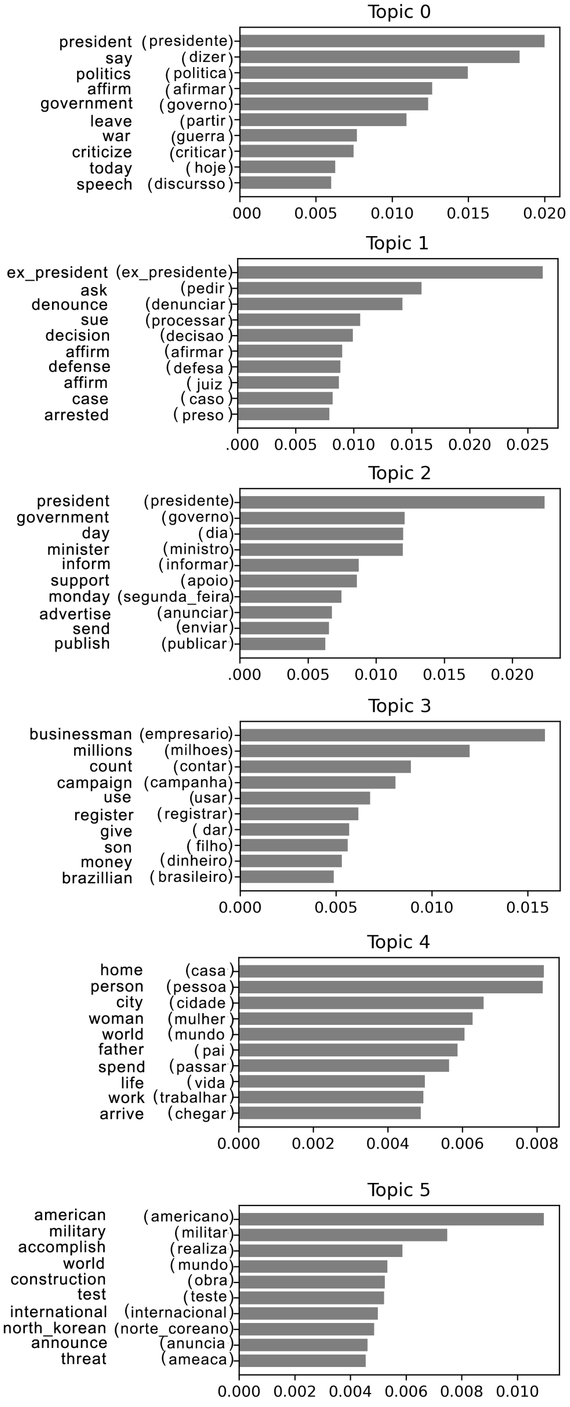


Fig. 3. Word-Topic probability distributions in real news.

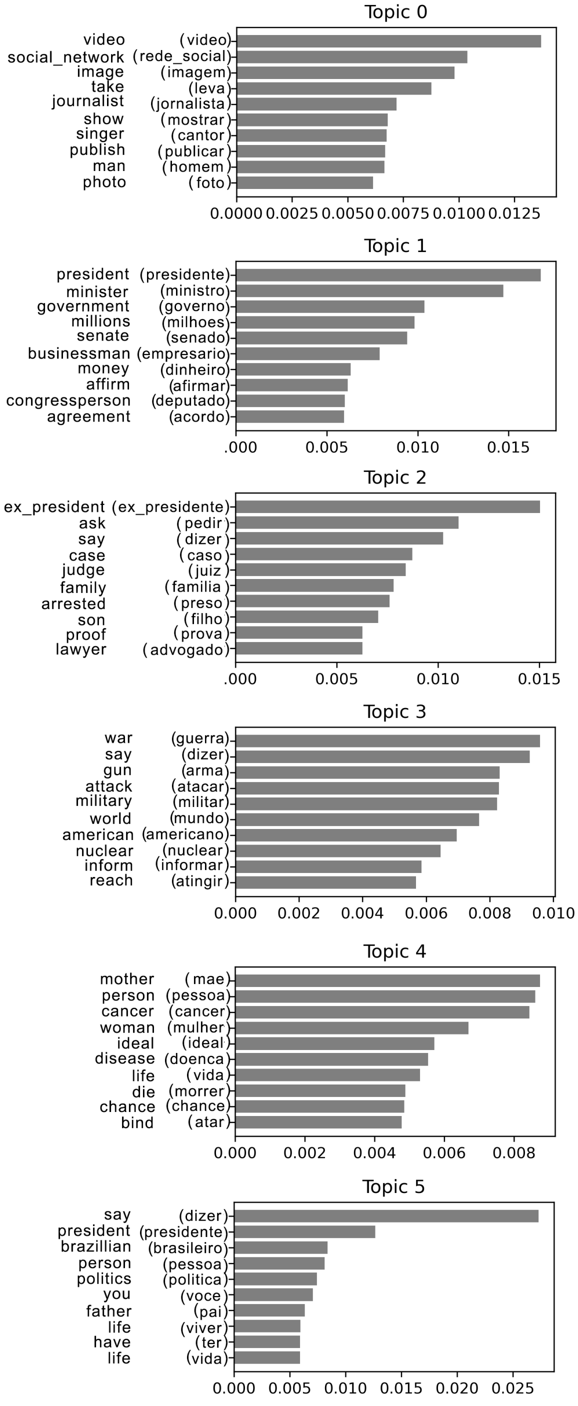


Fig. 4. Word-Topic probability distributions in fake news.