Expert Systems With Applications 146 (2020) 113199

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Towards automatically filtering fake news in Portuguese

Renato M. Silva[a], Roney L.S. Santos[b], Tiago A. Almeida[a,*], Thiago A.S. Pardo[b]

[a] Department of Computer Science, Federal University of São Carlos, Sorocaba, Brazil
[b] Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil

ARTICLE INFO

Article history:
Received 22 April 2019
Revised 2 October 2019
Accepted 8 January 2020
Available online 14 January 2020

Keywords:
Fake news
Text categorization
Natural language processing
Machine learning
Corpus construction

ABSTRACT

In the last years, the popularity of smartphones and social networks has been contributing to the spread of fake news. Through these electronic media, this type of news can deceive thousands of people in a short time and cause great harm to individuals, companies, or society. Fake news has the potential to change a political scenario, to contribute to the spread of diseases, and even to cause deaths. Despite the efforts of several studies on fake news detection, most of them only cover English language news. There is a lack of labeled datasets of fake news in other languages and, moreover, important questions still remain open. For example, there is no consensus on what are the best classification strategies and sets of features to be used for automatic fake news detection. To answer this and other important open questions, we present a new public and real dataset of labeled true and fake news in Portuguese, and we perform a comprehensive analysis of machine learning methods for fake news detection. The experiments were performed using different sets of features and employing different types of classification methods. A careful analysis of the results provided sufficient evidence to respond appropriately to the open questions. The various evaluated scenarios and the drawn conclusions from the results shed light on the potentiality of the methods and on the challenges that fake news detection presents.

1. Introduction

Deception is a kind of information that is intentionally produced and transmitted in order to create a false impression or conclusion (Burgoon, Buller, Guerrero, Afifi, & Feldman, 1996). Nowadays, the most dangerous type of deception, the fake news, tries to mimic the content reported by the official press. Fake news is different from news where the source is unsure or has not performed a thorough search on the subject, which is called misinformation, because it is purposely released to deceive people (Lazer et al., 2018). As a consequence, these news may be misleading or even harmful, especially when they are disconnected from their origins and original contexts (Rubin, 2014).

Today, social networks and instant messaging applications allow deceptive content to reach a number of people that was impossible before the Internet era. Due of their appealing nature, fake news spreads rapidly (Vosoughi, Roy, & Aral, 2018) influencing people's perceptions about various subjects, from news stories about alleged scientific studies that confirm half-truths to statements by politicians and celebrities that are distorted and act like a fire in the timelines of social networks. In this way, fake news have not only influenced political elections around the world, but also caused problems in public healthy (*e.g.*, by spreading conspiracies about vaccination campaign) and human tragedies (as the public lynchings and people doing justice with their own hands).

To make things worse, it is important to highlight the human difficulty of detecting not only fake news, but deceptive content in general. Research on this fact has already shown that humans can unsatisfactorily separate true news from fake ones (Charles F. Bond & DePaulo, 2006; George & Keane, 2006), reaching between 50% and 63% success depending on what is considered deceptive (Rubin & Conroy, 2011).

In such scenario, efforts to deal with fake news have arisen. Communication agencies have been giving support to fact-checking websites and companies with great digital appeal (*e.g.*, Facebook) are trying to educate their users. The academy has made efforts to combat fake news by studying how fake news spread, whether the statements made in the written language are true from the automatic verification of the facts, and how users behave. Some studies in Natural Language Processing (NLP) have also explored the linguistic features that might help detecting fake news.

The attempts to detect fake news include theoretical (Duran, Hall, McCarthy, & McNamara, 2010; Hauch, Blandn-Gitlin, Masip, & Sporer, 2015; Zhou & Zhang, 2008) and practical (Appling, Briscoe, & Hutto, 2015; Pérez-Rosas & Mihalcea, 2015; Rubin, Conroy, Chen,

* Corresponding author.
E-mail addresses: renatoms@dt.fee.unicamp.br (R.M. Silva), roneysantos@usp.br (R.L.S. Santos), talmeida@ufscar.br (T.A. Almeida), taspardo@icmc.usp.br (T.A.S. Pardo).

https://doi.org/10.1016/j.eswa.2020.113199
0957-4174/© 2020 Elsevier Ltd. All rights reserved.

& Cornwell, 2016) NLP approaches. According to Hauch, Masip, Blandón-Gitlin, and Sporer (2012), the automation of deceptive content detection is attractive for at least two reasons: i) such systems can be more objective than human judges, who are prone to biases (Levine, Park, & McCornack, 1999); and ii) online judgments of multiple cues from videos or audios can overwhelm the judge and lead to delays and errors. Therefore, NLP-based applications try to use linguistic patterns that serve to detect whether information is fake or not. However, much of the difficulty in such NLP-based research lines resides in the fact that it is language dependent and there are very few available corpora to develop and test the systems, mainly if we consider non-English languages.

To help filling this gap, we have recently presented preliminary data regarding a new dataset of labeled fake news written in Portuguese, called Fake.Br Corpus (Monteiro et al., 2018). In such paper, we basically have explained the data acquisition and labelling processes and run some preliminary classification algorithms.

In this paper, we have significantly extended the previous work by first introducing in details the manually built reference corpus with true and fake news, which was made publicly available in order to foster research and advances in the area. Then, we report experiments with machine learning techniques (using varied strategies, such as ensemble and stacking) on different sets of features (linguistic-based features and distributive and distributed text representations). We show that we significantly outperformed previous results recently reported in the literature over the same corpus, and we provide proper answers for the following important research questions that still remain open:

- *Q1: What are the best current methods for automatic detection of fake news?*
- *Q2: What is the best feature set for fake news classification?*
- *Q3: What is the impact of different classification strategies (e.g., ensemble and stacking) for fake news detection?*
- *Q4: Can the size of the texts influence the results of the classification?*

The remainder of this paper is organized as follows. Next section presents the main related work in the area. Section 3 details the Fake.Br Corpus. Sections 4 and 5 report our experiments and the obtained results. Conclusions and guidelines for future work are presented in Section 6.

2. Related work

In the NLP and related areas, the task of deceptive content detection has seen some important efforts and produced promising results, mainly motivated by the devastating nature of fake news.

Formally, in definitional terms, according to Rubin, Chen, and Conroy (2015), three main types of deception can be identified: (i) deception for humor purposes, making use of sarcasm and irony for producing parodies and satires; (ii) fake content to deceive people and spread misinformation; and (iii) the non-confirmed information that is publicly accepted – the rumors. Fake news usually fit in the second type.

Zhou (2005) has broadly defined a range of behaviors that people show when they are consciously generating or disseminating deceptive content, strategically or not. Such behaviours are organized through a taxonomy, reproduced in Fig. 1.

The taxonomy has two main groups: indicators of verbal and nonverbal language. Verbal indicators are directly related to the spoken or written content and language, whereas nonverbal cues focus on accessory features that are exhibited while a person is producing content. The verbal indicators are divided into two subgroups: linguistics-based and content-based. Linguistics-based verbal indicators include the attributes of the language, such as grammatical classes, semantics, spelling errors, and content diversity. NLP initiatives have mainly focused on this kind of features. Content-based verbal indicators focus on what is being transmitted, *i.e.*, on identifying the meaning of content sent to the user. The fact-checking models are in this subgroup. Nonverbal behaviors can be grouped further according to the source of the behavior as paralinguistic or proxemic-kinesic features. Paralinguistic attributes refer to properties that do not directly refer to the content of the speech, including tone and filled pauses, typing traces, participation behavior in a discourse, and so on. The proxemic-kinesic features describe a person's body postures, facial expressions, eye movements, and so on. They are usually associated with face-to-face communication.

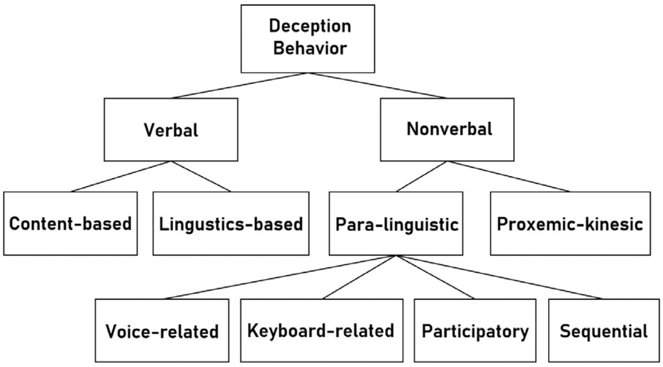


Fig. 1. Taxonomoy of deception behavior (Zhou, 2005).

The use of textual features to indicate potentially misleading content was studied in a variety of modalities and contexts presented in the literature (Burgoon et al., 1996; Pennebaker, Mehl, & Niederhoffer, 2003). Methods for identifying deceptive content have been explored, using varied features, as cited by Conroy, Rubin, and Chen (2015) and systematically organized by Zhou and Zhang (2008). There are relevant related work in the field of instant messengers, e-mails, chat rooms, social networks, and journalistic news. Zhou, Burgoon, Twitchell, Qin, and Nunamaker (2004a) propose to look at the amount of verbs and modifiers (adjectives and adverbs), text complexity (average sentence length and average word length), pausality (rate of occurrence of punctuation marks in the sentences), uncertainty (number of modal verbs and passive voice), non-immediacy (number of personal pronouns), expressivity (number of adverbs and adjectives in relation to nouns and verbs), diversity, and informality features. Pérez-Rosas and Mihalcea (2014), Pérez-Rosas and Mihalcea (2015), and Pérez-Rosas, Kleinberg, Lefevre, and Mihalcea (2017) evaluate the performance of machine learning classifiers using bag of words, part of speech tags, syntactic information, readability metrics, and word semantic classes. Of special interest to us, Monteiro et al. (2018) test some of these features for the corpus that we introduce here, producing promising results.

There are also other efforts to identify deception. For instance, Appling et al. (2015) look for indications of falsification, exaggeration, omission, and deception in declarations in social networks, evaluating the following hints in the texts: lies, contradictions, distortions, phrase modifiers, superlatives, lack of information, half truths, subject change, irrelevant information, and misconception. Potthast, Kiesel, Reinartz, Bevendorff, and Stein (2018) use writing style patterns to detect hyperpartisan news, *i.e.*, a type of "news" that is extremely one-sided, inflammatory, emotional, and often full of untruths, in connection to fake news. Volkova, Shaffer, Jang, and Hodas (2017) propose cues related to verbs (covering assertive, factive, implicative, and report verbs), subjectivity cues (polarity of words) and psycholinguistic cues (*e.g.*, factual data and personal pronouns).

Regarding fact-checking, the researches focus mostly on structuring the information for further analysis. Thorne and Vlachos (2018) consider that a frequent entry for fact-checking approaches is a triple (subject, predicate, and object), with the justification that this type of input facilitates the fact checking in structured databases (and in some cases when they are semi-structured). Ciampaglia et al. (2015) use the concept of knowledge graphs filled with infoboxes from Wikipedia: when given a new information, it is assumed to be true if the predicate of the statement exists as an edge in the graph, or if there is a shortest path connecting the related nodes. Although most of the efforts in this line use the classification of the statement on a scale between fake and true, other initiatives highlight alternative ways of checking content, such as verifying whether the statement is common sense (Angeli & Manning, 2014; Habernal, Wachsmuth, Gurevych, & Stein, 2018), a rumor (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018), or a clickbait (Chakraborty, Paranjape, Kakarla, & Ganguly, 2016; Potthast et al., 2018), and if the title of the text is related to its content (Chesney, Liakata, Poesio, & Purver, 2017). Rashkin, Choi, Jang, Volkova, and Choi (2017) evaluate the reliability of news articles, classifying them as reliable, hoax, satire, or advertisement. At the sentential level, Hassan et al. (2015) modeled a classifier that categorized sentences from presidential debates into three categories: non-factual sentence (opinions, beliefs, and declarations), unimportant factual sentence (factual, but not check-worthy), and check-worthy factual sentence (factual claims that are true).

Although the task is recent, some corpora with different types of deception have been created. For example, Pérez-Rosas and Mihalcea (2014) introduce three datasets on popular topics (abortion, death penalty, and feelings about friendships) with 100 deceptive and 100 truthful sentences. Rubin et al. (2016) build two datasets of satirical and true news for the domains of civics, science, business, and “soft” news, summing up 240 texts. Pérez-Rosas et al. (2017) collect two datasets about celebrities: the first one was collected from the web (with 100 fake and 100 true news), and the other emulates journalistic writing style (with 240 fake news). The Emergent (Ferreira & Vlachos, 2016) and LIAR (Wang, 2017) are also well-known corpora for the English language. There are also some available datasets in Dutch (Verhoeven & Daelemans, 2014), Chinese (Zhang, Wei, Tan, & Zheng, 2009), and Italian (Fornaciari & Poesio, 2013) languages. The cited corpora were constructed in different ways. Most of them were manually collected, searching for the fake and true news (or, sometimes, not the full texts, but only parts of them) in websites, in a time consuming and laborious approach. Other corpora used crowdsourcing to collect the texts, using Amazon Mechanical Turk or proprietary online platforms, having to deal with issues of reliability and spontaneity of the data and willingness of online users to contribute.

It is also worthy citing some recent international efforts for building datasets and performing scientific contests in related tasks, such as the ones of CLEF 2019 (in the ProtestNews and CheckThat! evaluation tracks)[1] and SemEval 2019 (in the Hyperpartisan News Detection track).[2]

Despite recent efforts, there is still few real, public, and labeled collections of fake news, especially for non-English languages. Such datasets are essential for machine learning workflows, such as feature extraction and analysis, as well as training and testing of different filtering approaches.

To help fill this important gap, we report our efforts to build a reference corpus of aligned true and fake news – the Fake.Br Corpus – that may subsidize the research efforts in the area, specially for the Portuguese language, which is the native language of the authors of this paper. To the best of our knowledge, this is the first corpus of such nature for this language. Differently from most of the corpora cited before, the corpus we present here was built with a mixed approach: we have manual steps, but we also employ automatic processes to speed up the corpus construction, resulting in a semi-automatic strategy; our manual steps were also performed to favor reliability. The corpus and the related processes to build it are described in what follows.

3. The Fake.Br Corpus

Creating a corpus with the potential to be a benchmark is a challenging task with several project decisions underlying it. Hovy and Lavid (2010), who are reference authors in the area, cite some important research questions that anyone working with corpus should pay attention, which include issues related to selecting the texts to compose the corpus, determining the phenomenon of interest to annotate, performing the annotation (which, in turn, requires selecting the annotators and, if necessary, the annotation interface, as well as to constantly follow and evaluate the annotation work), and distributing the corpus. Depending on the corpus purpose, each step must be appropriately adapted.

Besides the general guidelines in the area for corpus construction, specific directions do exist for building corpora of deceptive content. Rubin et al. (2015) suggests that: the corpus should have truthful texts and their corresponding deceptive versions (which, according to the authors, is challenging), in order to allow finding patterns and regularities in “positive and negative instances”; the texts in the corpus should be in plain text format (simplifying the posterior NLP tasks); the texts should show similar number of words (to avoid bias in learning)[3]; the whole corpus should belong to a specific time interval (as language is alive and writing style changes in time, what might bring problems for the corpus intended purposes); and the corpus should keep the related metadata information (*e.g.*, the URL of the news, the authors, publication date, and number of comments and visualizations) because it can be useful for fact checking algorithms.

We have followed the above steps and directions to create our corpus for the final purpose of fake news classification. For such purpose, our corpus is composed of aligned true and fake news written in (Brazilian) Portuguese. For the alignment, we mean that, for each fake news, we collected a corresponding true news, which, if not explicitly denying the fake news, is topically related (which is the most common case).

To find the appropriate texts to compose the corpus was a challenging task. We searched the web for the available fake news, which were manually checked to guarantee that they had deceptive content. The manual verification was important to ensure the data quality and, therefore, the reliability of the resulting corpus. The selected fake news were then used in a semi-automatic process to look for their corresponding true versions in the web.

The availability of the deceptive news and their corresponding true versions is very important for machine learning tasks (which require positive and negative instances for the learning success) and linguistic investigations, which look for textual patterns and their contexts of usage for language description.

Our resulting corpus has 7200 news (3600 fake and 3600 legitimate news) in plain text individual files. For each fake news, we tried to collect a corresponding true news with similar text size. However, in general, the true news in the corpus are longer

[1] http://clef2019.clef-initiative.eu/index.php?page=Pages/labs_info.html.

[2] https://pan.webis.de/semeval19/semeval19-web/.

[3] If the texts have very different sizes, normalization (such as text truncation) may be performed.

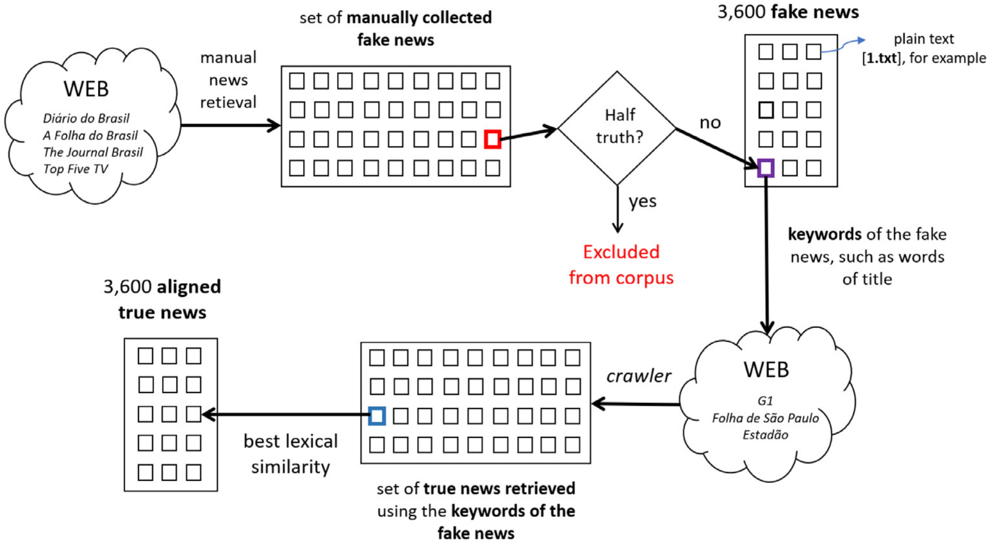


Fig. 2. Process of building the Fake.Br Corpus.

than their corresponding fake news versions. Most of the news we collected were published between January of 2016 and January of 2018. In addition to the plain text of the news, we also saved all the available metadata information.

The general schema of the process of collecting news for the Fake.Br Corpus is shown in Fig. 2. The whole process (including the analysis of the news) took approximately three months — from December 2017 to February 2018 — to be fully accomplished.

The first step was collecting the fake news. We initially looked for sites and blogs that post dubious news. According to the Monitor Tool of the Political Debate in the Digital Media, from the Research Group on Public Policies for Access to Information,[4] some characteristics of layout and content may help to identify a site that reproduces false content, which are listed below:

- The author of the news is not cited;
- The titles of the news are sensationalist, *i.e.*, they lead the user to click on it for curiosity;
- The news contains grammatical and agreement errors, as well as adjectives, such as “coup” and “thief”, among others of strong sense;
- The news is written in a way that has many uppercase letters, with multiple exclamation or question marks, since this type of text is often not written in newspaper essays;
- The news does not indicate when the fact happened, not containing other sources and references;
- The site does not have a page that identifies its administrators or journalists in charge of the news. When there is, in some cases, the “Who We Are” page does not allow to identify who is responsible for the site and its content;
- The site has a polluted and sometimes confusing layout, which makes it to look like big news sites, showing credibility to users who do not quite understand what is shown.

In a manual search on the web, we identified four sites with the characteristics presented above: *DiÃ!‘rio do Brasil*, *The Folha do Brasil*, *The Jornal Brasil* and *Top Five TV*. We manually checked the news to prevent collecting news that contains half-truths. Therefore, we only selected[5] completely fake news to compose the corpus. Two people were involved in this task. The checking step was supported by online news portals, such as *Agncia Lupa*,[6] *Fato ou Fake*,[7] *Aos Fatos*,[8] and *Boatos.org*,[9] that perform fact-checking for news in Portuguese, listing and commenting each one. It is important to highlight that, as this checking step was mostly mechanical, i.e., looking for the fake news in the online portals and verifying the comments about their content, it made no sense to compute agreement annotation measures in this case.

Once we had the fake news, we used a web crawler to collect the true news from webpages of some prestigious news agencies in Brazil, such as *G1*, *Folha de SÃ£o Paulo*, and *EstadÃ£o*. To perform the search, we used some keywords extracted from the fake news, such as the nouns and verbs of the titles, and the most frequent words (after removing stopwords). This process resulted in the retrieval of 40,000 news. After, we used the cosine lexical similarity measure (Salton & McGill, 1986) to select one corresponding true news for each fake news previously collected. The equation to compute the cosine similarity (*cos*) is shown below:

$$cos(\vec{f},\vec{v}) = \frac{\vec{f}\cdot\vec{v}}{|\vec{f}||\vec{v}|} = \frac{\sum_{i=1}^{n} f_i v_i}{\sqrt{\sum_{i=1}^{n} f_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \quad (1)$$

where $\vec{f}$ represents the fake news and $\vec{v}$ the true news. The two news are converted into vectors by some vector representation (*e.g.*, bag of words). The cosine similarity value is a number in the interval [0,1], where the value 0 indicates that the vectors are completely different and the value 1 indicates that the vectors are completely similar.

Finally, we have also manually checked the selected true news in order to guarantee that they were topically related to their fake versions. The same two people that checked the fake news were in charge of checking the topically relatedness of the true news (which is also a straightforward process, without need of agreement measurement). Table 1 shows some examples of aligned true and fake news in the corpus.

The news in the corpus may be categorized in the following topics: economy, science & technology, society & daily news, politics, religion, and TV & celebrities. We manually assigned the news to the topics that they were associated to in the sites they were

[4] Available at http://gpopai.usp.br/monitor/070616/, accessed on Mar 22, 2019.

[5] No automatic crawling was done, neither any keywords were used to search/filter the news in the sites.

[6] https://piaui.folha.uol.com.br/lupa/.

[7] https://g1.globo.com/fato-ou-fake/.

[8] https://aosfatos.org/.

[9] https://www.boatos.org/.

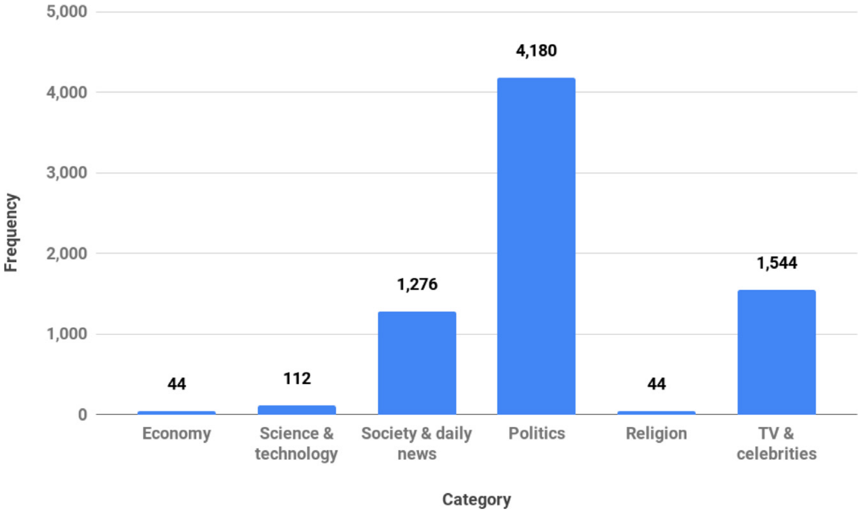


Fig. 3. Frequency of news by category in the Fake.Br Corpus.

Table 1
True and fake news: examples from the corpus.

Fake	True
Polos magnéticos da Terra podem se inverter e causar colapso mundial: "A Terra ficaria inabitável". Aos menos 100 mil pessoas morreriam por ano pela alta nos níveis de radiação espacial." Se o campo magnético continuar a diminuir e os polos magnéticos se inverterem, a Terra pode acabar como Marte – um local seco, árido e incapaz de preservar a vida.	*Inversão dos polos magnéticos da Terra pode ocorrer mais rápido do que o previsto. Segundo afirmações, essas ocorrências são, a princípio, indistinguíveis das verdadeiras mudanças nos polos. Apesar dessas reversões não representarem qualquer ameaça áhumanidade, os especialistas alertam que poderão gerar falhas nos satélites que orbitam a Terra.*
Temer avisa que vai vetar a lei anti-Uber. Mesmo com a aprovação dos deputados federais a lei que dificulta o trabalho do UBER no Brasil poderá ser vetada pelo presidente Michel Temer. A equipe do presidente Michel Temer diz esperar que as emendas consideradas prejudiciais ao serviço de transporte Uber – empresa que conecta motoristas particulares a passageiros – e similares sejam alteradas ou derrubadas pela base aliada no Senado.	*Prefeitura de SP flexibiliza futuras regras para motoristas de aplicativos. Ás vésperas do início da vigência das novas regras para aplicativos de transporte em São Paulo, a gestão João Doria (PSDB) decidiu flexibilizar nesta sexta-feira (5) alguns pontos da regulação e adiou o prazo para que motoristas e aplicativos se preparem antes de serem fiscalizados.*

Table 2
Basic analysis of the Fake.Br Corpus.

Features	True news	Fake news
Average number of tokens	1268.5	216.1
Average number of types (without punctuation symbols and numbers)	494.1	119.2
Average size of words (in characters)	4.8	4.8
Type-token ratio	0.47	0.68
Average number of sentences	54.8	12.7
Average size of sentences (in words)	21.1	15.3
Average number of verbs (normalized by tokens)	13.4	14.3
Average number of nouns (normalized by tokens)	24.6	24.5
Average number of adjectives (normalized by tokens)	4.4	4.1
Average number of adverbs (normalized by tokens)	4.0	3.7
Average number of pronouns (normalized by tokens)	5.2	5.0
Average number of stopwords (normalized by tokens)	32.8	31.0
Proportion of texts with spelling errors	3.0	36.0

collected from. We also make this information available in the corpus distribution (as we comment later). The distribution of news by category is shown in Fig. 3. We can see that politics is the most frequent topic.

We show in Table 2 an analysis of the news in relation to some traditional NLP features that are based on the number of types, tokens, sentences, verbs, adjectives, and other components of the sentences.

It is perceptible that the true news are much larger in size than the fake news, in number of tokens, words, terms and characters, hurting one of the recommendations proposed by Rubin et al. (2015), which can be a problem to machine learning algorithms because this characteristic can bias the classification.

We can see in Table 2 that the number of nouns, adjectives, adverbs, and pronouns in the true news is higher than in fake news. On the other hand, the fake news, in general, have more spelling errors (36% of fake news has some type of spelling error against only 3% of the true news)[10].

We have also computed the linguistic features proposed by Zhou et al. (2004a) (see Table 3). The pausality feature indicates the occurrence of pauses in the text, which is computed as the

[10] To find spelling errors, we have (i) used the ENELVO text normalization tool (Bertaglia & Nunes, 2016) (which is a state of the art tool for Portuguese) to automatically correct the texts and (ii) compared the original and corrected versions of the texts to detect texts that had to be corrected.

Table 3
Features of Zhou et al. (2004a) computed for the Fake.Br Corpus.

Features	True news	Fake news
Average pausality	3.04	2.46
Average emotiveness	0.21	0.20
Average uncertainty	2.11	2.39
Average non-immediacy	0.235	0.249

number of punctuation marks over the number of sentences. Emotiveness measures the language expressiveness (Zhou, Twitchell, Qin, Burgoon, & Nunamaker, 2003), calculated as the sum of the number of adjectives and adverbs divided by the sum of the number of nouns and verbs. Uncertainty is based on the occurrences of modal verbs and the use of passive voice. The non-immediacy feature is based on the frequency of use of the 1st and 2nd pronouns.

To offer a general view of the most important corpora in the literature (that we cited in the previous section) and the similarities and differences in relation to our Fake.Br corpus, we show in Table 4 a synthetic comparative view of the corpora. One may see that our corpus is among the largest ones (after the corpora of Rashkin et al. (2017) and Wang (2017)); considering the ones with aligned texts, our corpus is the biggest one by a large margin.

Finally, it is important to highlight that the Fake.Br Corpus is publicly available[11].

In what follows, we present a set of experiments performed to check if well-known text categorization techniques can be successfully employed to automatically detect fake news in Portuguese language. For this, different text representation techniques, features and text categorization approaches have been combined to provide robust results that can be used as a baseline for future comparisons.

4. Experiments

The experiments were diligently designed to find proper answers for the open research questions presented at the end of Section 1. For this, we performed experiments using the following linguistic based-features (Monteiro et al., 2018; Zhou, Burgoon, Twitchell, Qin, & Nunamaker, 2004b): pausality, emotiveness, uncertainty over the number of verbs of the news, non-immediacy, diversity, average size of the sentences, average size of the words, number of spelling errors. The four first features were introduced in the previous section. Diversity is computed as the total number of different content terms over the number of content terms (*i.e.*, it is a more refined version of the type-token ratio). All the features are properly normalized.

Each sample was also represented in three different ways: with the traditional bag-of-words (BoW) and with two distributed text representations using the state-of-the-art Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) techniques. In the experiments with BoW, we used the TF-IDF (term frequency-inverse document frequency) technique to adjust the weights of the tokens of each document. For Word2Vec and FastText, we used the pre-trained word vectors proposed in Hartmann et al. (2017). The models used to generate these vectors were trained with Portuguese language documents from 17 datasets of different domains, totalizing 1,395,926,282 tokens. For both Word2Vec and FastText, we use vectors with 300 dimensions trained with the Skip-Gram approach (Hartmann et al., 2017). For each document of the Fake.Br Corpus, we obtained the pre-trained vector for each word and then we compute their average (Joulin, Grave, Bojanowski, & Mikolov, 2017).

4.1. Preprocessing

In the experiments with the linguistic-based features, we applied the Z-score normalization using information from the training examples.

Before generating the feature vectors with BoW, Word2Vec or FastText, all instances of our dataset were converted to lowercase. Then, numerals, URLs, and emails were normalized into the dummy features ‘0’, ‘URL’, and ‘EMAIL’, respectively. After that, we tokenize the documents based on whitespaces and punctuation marks. Fig. 4 presents word clouds to visually summarize the relative frequency of tokens obtained after the preprocessing. As we can see, many frequent words in the true news also occur in fake news, which may hinder the identification of the news class.

4.2. Methods

We performed experiments with the following established classification methods: logistic regression (LR) (Yu, Huang, & Lin, 2011), support vector machines (SVM) (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995), decision trees (DT) (Breiman, Friedman, Olshen, & Stone, 1984), random forest (RF) (Breiman, 2001), bootstrap aggregating (bagging) (Breiman, 1996), and adaptive boosting (AdaBoost) (Freund & Schapire, 1996).

We used the implementations of all methods from `scikit-learn` library (Pedregosa et al., 2011). The experiments with SVM were evaluated using a linear kernel because its computational cost is lower than RBF and polynomial. Moreover, as the performance of SVM, RF, Bagging, and AdaBoost can be highly affected by the choice of parameters, we performed a grid search using hold-out cross-validation to find the best values for their main parameters. For the regularization parameter of SVM, the following range of values were analyzed: $\{2^{-5}, 2^{-3}, 2^{-1} \ldots, 2^{15}\}$. For the number of estimators used in RF, bagging, and AdaBoost, the following range of values were analyzed: {10, 30, 50, ..., 110}. For the other methods, we set their parameters to the default values.

4.3. Performance measures

To compare the results, we employed the following well-known performance measures for spam and other misleading content (Silva, Alberto, Almeida, & Yamakami, 2017):

- Legitimate news blocked rate (LBR) or false positive rate: proportion of legitimate news incorrectly labeled as fake news (the lower, the better);
- Fake news caught rate (FCR) or recall: proportion of fake news correctly identified (the higher, the better);
- Fake news precision rate (FPR): proportion of news classified as fake and that truly belong to the fake class;
- F-measure: harmonic average of the FCR and FPR.

5. Results

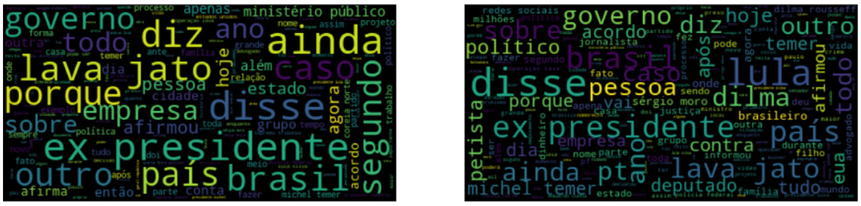
We performed experiments with linguistic-based features and with features generated by varied text representation techniques (BoW, Word2Vec, and FastText).

As the legitimate news are often longer than fake news (see Section 3), we evaluated the hypothesis that the classifiers can be biased by the size of the text. If this hypothesis is true, conclusions based on the results obtained with the full texts may be wrong because the classifiers can present overestimated performance. In

[11] At https://github.com/roneysco/Fake.br-Corpus and in the OPINANDO project webpage at https://sites.google.com/icmc.usp.br/opinando/.

Table 4
A synthetic view of the corpora in the literature.

Reference work	Name of the corpus	Language	Type of deception	Number of true texts	Number of deceptive texts	Topics of the texts	Aligned texts?	Specific time interval?	Available metadata?	Construction mode
Zhang et al. (2009)	-	Chinese	Rumor	131	187	Sports, Entertainment, Social life	No	Yes (2001-2008)	None	Semi-automatic
Fornaciari and Poesio (2013)	DECOUR	Italian	Fake News	1202	945	People testimony (calumny and false testimony)	No	No	Time of hearings, Time stamps	Manual
Pérez-Rosas and Mihalcea (2014)	-	English and Spanish	Fake News	100	100	Abortion, Death penalty, Best friend	No	No	None	Crowdsourcing
Verhoeven and Daelemans (2014)	CSI	Dutch	Fake News	270	270	Musicians, Food chains, Books, Smartphones, Movies	No	Yes (2012-2013)	Age, Gender, Region of origin, Personality, Sexual orientation	Manual
Vlachos and Riedel (2014)	-	English	Fake News	135	86	Politics and public life	No	No	Date, Author, Link	Manual
Ferreira and Vlachos (2016)	Emergent	English	Rumor	1,237	395	World and national U.S. news, Technology	No	No	None	Manual
Rubin et al. (2016)	-	English	Humorous Fakes	240	240	Civics, Science, Business, “Soft” news	Yes	Yes (2016)	None	Manual
Pérez-Rosas et al. (2017)	-	English	Fake News	340	340	Sports, Business, Entertainment, Politics, Technology, Education	Yes	No	None	Manual and crowdsourcing
Rashkin et al. (2017)	-	English	Fake News and Humorous Fakes	13,995	60,481	U.S. news and world reports	No	No	None	Automatic
Wang (2017)	LIAR	English	Fake News	~4600	~8,200	Economy, Health care, Taxes, Federal budget, Education, Jobs, State budget, Candidates biography	No	Yes (2007-2016)	Speaker affiliations	Manual
Our corpus	Fake.Br	Brazilian Portuguese	Fake News	3600	3,600	Politics, Economy, TV & celebrities, Society & daily news, Science & technology, Religion	Yes	Yes (2016-2018)	Link, Date, Number of links	Semi-automatic



(a) *True news.*

(b) *Fake news.*

Fig. 4. Word clouds representing the relative frequency of the tokens.

Table 5
Scores obtained by each method in the experiments with the linguistic-based features.

	LBR	FCR	FPR	F-measure
RF	**0.060**	**0.941**	**0.940**	**0.941**
Bagging	0.067	0.935	0.933	0.934
AdaBoost	0.080	0.929	0.921	0.925
SVM	0.081	0.931	0.920	0.925
LR	0.081	0.928	0.921	0.924
NB	0.135	0.938	0.875	0.905
DT	0.099	0.902	0.901	0.901

real world applications, they could easily be tricked by long fake news. To evaluate this hypothesis and answer the research question Q4 presented at the end of Section 1, we compared the results obtained using full texts with the results obtained using truncated ones.

In what follows, we report the results of experiments considering different settings.

5.1. Results obtained with the linguistic-based features

Table 5 shows the results obtained with the linguistic-based features. The results are sorted by F-measure and bold values indicate the best scores. The scores are presented as a grayscale heat map, where the better the score for a given method, the darker the cell color.

All methods obtained an F-measure above 0.9, which indicates that the linguistic-based features are sufficiently informative to detect more than 90% of the fake news (FCR).

RF obtained the best result for all the four performance measures. It was able to detect more than 94% of fake news with the price of wrong blocking 6% of true news. On the other hand, NB achieved the worst LBR (0.135) and FPR (0.875), while DT obtained the worst FCR (0.902) and F-measure (0.901).

5.2. Results obtained with features generated by text representation techniques using full texts

In this section, we present the results of the experiments with the three text representation techniques previously described: BoW, Word2Vec, and FastText. The full text of the documents was used, that is, we did not use any truncation process.

For BoW, we performed the following experiments:

1. Stopwords were not removed and stemming was not applied (Table 6a);
2. Stopwords were removed (Table 6b);
3. Stopwords were removed and stemming was applied (Table 6c);
4. Stopwords were removed and information gain technique was used for selecting the best 1,000 features (Table 6d);

Table 6 synthesizes our results for the BoW variations and for the Word2Vec and FastText methods. In each subtable of Table 6, bold values indicate the best scores. Moreover, the scores are presented as a grayscale heat map, where the better the score for a given method, the darker the cell color.

The results indicate that removing stopwords and applying stemming did not improve the performance of the classification methods. In the experiments with the straightforward BoW, LR (the best overall classifier) was able to detect about 98% of fake news with the price of wrongly blocking 3.8% o true news. In the experiment with stopwords removing, the rate of fake news detected by LR was about the same, but the best rate of true news wrongly blocked increased to 4.6%. After applying stemming, the rate of true news wrongly blocked by LR increased to 5.0%. Therefore, there is evidence that these preprocessing techniques can remove features that are important for the fake news classification, as well as in some other text classification tasks such as spam detection (Méndez, Iglesias, Fdez-Riverola, Díaz, & Corchado, 2006). The scores shown in Table 6d indicate that feature selection was also not effective.

The results with BoW were better than those obtained with Word2Vec and FastText. For example, the best F-measure obtained with BoW was 0.971, while in the experiments with Word2Vec and FastText, the best F-measure was 0.893 and 0.897, respectively (a difference of more than 7%). Fake news, in general, contains noise such as abbreviations, slang, and misspelled words. The Word2Vec and FastText models used to generate the word vectors were trained with documents from Wikipedia, Google News, and other sources that, in general, contain well-written, low-noise text. Therefore, we believe that these models do not generate representative vectors for fake news. Probably, if the word embedding models had been trained with noisy documents, the results would have been better since some studies recommend training distributed representation models with a corpus composed by text with the same characteristics of the application domain (Lochter, Pires, Bossolani, Yamakami, & Almeida, 2018).

The scores in the experiments with BoW (Table 6) were also better than those obtained with the linguistic-based features (Table 5). For example, the best overall method in the experiments with BoW obtained a LBR of 3.8%, while the LBR obtained by the best overall method in the experiment with linguistic-based features was 6%. However, the dimensionality of the BoW-based representation is very higher than the dimensionality of the representation based on linguistic features. Therefore, in devices with low computational resources, a fake news filter based on linguistic features may be more advantageous.

Regarding the classification methods, it is clear that logistic regression obtained the best score in most of the experiments with the BoW-based representation, being able to detect, on average, 97% of fake news with the price of wrongly blocking, on average, 6% of true news. In the experiments with the distributive text representation techniques (Word2Vec and FastText), RF achieved the best results. On the other hand, DT and NB obtained the worst FCR and F-measure in all the experiments.

Table 6
Scores obtained by each method in the experiments with the full texts.

(a) BoW.

	LBR	FCR	FPR	F-measure
LR	0.038	0.978	**0.963**	**0.971**
SVM	0.043	**0.979**	0.958	0.968
AdaBoost	0.044	0.965	0.956	0.961
Bagging	0.036	0.950	**0.963**	0.956
RF	0.059	0.969	0.943	0.956
DT	0.062	0.933	0.938	0.935
NB	**0.016**	0.278	0.946	0.429

(b) BoW – *stopwords*.

	LBR	FCR	FPR	F-measure
LR	0.046	**0.980**	0.955	**0.967**
SVM	0.050	0.979	0.951	0.965
AdaBoost	0.045	0.965	**0.956**	0.960
RF	0.069	0.978	0.934	0.955
Bagging	0.051	0.936	0.948	0.942
DT	0.101	0.907	0.900	0.903
NB	**0.016**	0.278	0.947	0.430

(c) BoW – *stopwords* and *stemming*.

	LBR	FCR	FPR	F-measure
LR	0.050	**0.974**	**0.951**	**0.963**
SVM	0.055	0.972	0.947	0.959
AdaBoost	0.051	0.959	0.949	0.955
RF	0.075	0.971	0.928	0.949
Bagging	0.057	0.934	0.943	0.939
DT	0.100	0.903	0.900	0.901
NB	**0.047**	0.489	0.912	0.636

(d) BoW – *stopwords* and *feature selection*.

	LBR	FCR	FPR	F-measure
RF	0.049	**0.966**	0.952	**0.959**
AdaBoost	0.054	0.958	0.946	0.952
Bagging	0.064	0.948	0.937	0.943
SVM	0.088	0.954	0.916	0.934
LR	0.092	0.952	0.912	0.931
DT	0.124	0.908	0.880	0.893
NB	**0.012**	0.764	**0.984**	0.861

(e) Word2Vec.

	LBR	FCR	FPR	F-measure
RF	0.130	**0.912**	0.876	**0.893**
Bagging	0.139	0.896	0.865	0.880
SVM	**0.092**	0.837	**0.902**	0.868
AdaBoost	0.123	0.854	0.874	0.864
LR	0.116	0.789	0.872	0.828
NB	0.149	0.745	0.834	0.787
DT	0.218	0.727	0.769	0.748

(f) FastText.

	LBR	FCR	FPR	F-measure
RF	0.122	**0.913**	0.882	**0.897**
Bagging	0.133	0.895	0.871	0.883
AdaBoost	0.123	0.863	0.876	0.870
SVM	**0.086**	0.832	**0.907**	0.868
LR	0.106	0.803	0.884	0.841
NB	0.161	0.762	0.826	0.793
DT	0.216	0.731	0.772	0.751

5.3. Results obtained with features generated by text representation techniques using truncated texts

In this section, we show in Table 7 the results of the same experiments presented in the previous section but with truncated texts (limited to 200 tokens).

As in the experiments with the full texts, the results in Table 7 indicate that removing stopwords, applying stemming, and performing feature selection did not improve the results with the truncated news. The best F-measure with BoW was 0.937, but it decreased to 0.924 after removing stopwords, it decreased to 0.920 after applying stemming, and it decreased to 0.898 after applying feature selection. The drop in scores was also observed for all three other performance measures. As we discuss in Section 5.2, we believe that these techniques remove important features for fake news detection.

The results in the experiments with Word2Vec and FastText were inferior to those obtained in experiments with BoW. For example, the best rate of fake news detected in the experiments with Word2Vec and FastText was 11% lower in comparison to the best result of the experiments with BoW. At the same time, in the experiments with BoW, the best classifier wrongly blocked 83% fewer true news than the best classifier of the experiments with Word2Vec and FastText. These results reinforce the hypothesis raised in the previous section that the word embedding models generated vectors of low quality because they were trained with well-written texts. Fake news have noises (*e.g.*, misspelled words and slangs) and, therefore, we believe that models trained with both well-written documents and noisy documents could generate more representative vectors. Unfortunately, we did not find any public model of word embeddings trained with a corpus composed of well-written and noisy Portuguese language documents.

We show in the previous section that the results obtained with the full texts were higher than those obtained with the linguistic features. However, the same performance was not observed in the experiments with truncated texts. The FCR and F-measure obtained in these experiments, for all textual representation techniques, were inferior to the results obtained with the linguistic-based features. For example, the best FCR and F-measure in the experiments with truncated texts were, respectively, 0.937 and 0.932, while the best FCR and F-measure obtained with linguistic-based features were both 0.941. If we analyze the LBR and FPR, we can see that the analysis of the results is different, since the values of these two performance measures were better in the experiments with the truncated texts. The best LBR and FPR in the experiments with truncated texts were, respectively, 0.057 and 0.943, while the best LBR and FPR obtained with linguistic-based features were, respectively, 0.060 and 0.940.

The great difference between the results obtained with the full texts and the truncated texts confirms our hypothesis that the classifiers are biased by the size of the text. Therefore, we recommend that studies that investigate fake news evaluate the classification methods based on the truncated texts because experiments with full texts can present overestimated results. It is important to look for classification methods that use other characteristics of the documents to identify their classes because the size of the text (number of terms) can be easily manipulated by fake news writers.

In the experiments with truncated texts, as well as in the experiments with full texts, LR obtained the best scores in most experiments with BoW. On the other hand, SVM obtained the best results in the experiments with Word2Vec and FastText. NB and DT, as in the previous experiments, obtained the lowest results. For example, in the experiment with FastText, DT has detected less than 67% of fake news and wrongly blocked more than 31% of true news.

Given that the best results considering all the experiments were obtained using BoW and linguistic-based features, we raised the hypothesis that combining the predictions using these features can improve the overall performance. So, in the following two subsections, to evaluate this hypothesis and answer the research question Q3 presented at the end of Section 1, we present an ensemble and

Table 7
Scores obtained by each method in the experiments with the truncated texts.

(a) BoW.

	LBR	FCR	FPR	F-measure
LR	**0.057**	**0.932**	**0.943**	**0.937**
SVM	0.060	0.930	0.939	0.935
AdaBoost	0.093	0.907	0.907	0.907
RF	0.058	0.872	0.937	0.903
Bagging	0.120	0.858	0.878	0.868
DT	0.189	0.801	0.809	0.805
NB	0.252	0.685	0.731	0.707

(b) BoW – *stopwords*.

	LBR	FCR	FPR	F-measure
LR	**0.078**	**0.926**	**0.922**	**0.924**
SVM	0.081	0.925	0.919	0.922
RF	0.104	0.893	0.897	0.895
AdaBoost	0.122	0.894	0.880	0.887
Bagging	0.167	0.876	0.840	0.858
DT	0.224	0.801	0.782	0.791
NB	0.253	0.685	0.730	0.707

(c) BoW – *stopwords* and *stemming*.

	LBR	FCR	FPR	F-measure
LR	**0.083**	**0.923**	**0.918**	**0.920**
SVM	0.086	0.920	0.915	0.917
RF	0.093	0.886	0.906	0.896
AdaBoost	0.119	0.884	0.882	0.883
Bagging	0.145	0.846	0.855	0.850
DT	0.227	0.777	0.774	0.775
NB	0.392	0.736	0.653	0.692

(d) BoW – *stopwords* and *feature selection*.

	LBR	FCR	FPR	F-measure
SVM	**0.100**	**0.896**	**0.900**	**0.898**
LR	0.103	**0.896**	0.897	0.896
AdaBoost	0.125	0.891	0.878	0.884
RF	0.135	0.894	0.869	0.881
Bagging	0.156	0.883	0.851	0.866
NB	0.155	0.809	0.840	0.824
DT	0.214	0.800	0.789	0.795

(e) Word2Vec.

	LBR	FCR	FPR	F-measure
SVM	**0.143**	**0.833**	**0.854**	**0.843**
LR	0.168	0.797	0.827	0.812
Bagging	0.177	0.777	0.815	0.796
RF	0.178	0.774	0.813	0.793
AdaBoost	0.211	0.773	0.786	0.779
NB	0.205	0.651	0.762	0.701
DT	0.316	0.677	0.682	0.679

(f) FastText.

	LBR	FCR	FPR	F-measure
SVM	**0.138**	**0.833**	**0.858**	**0.845**
LR	0.151	0.802	0.842	0.821
RF	0.172	0.786	0.821	0.803
Bagging	0.173	0.777	0.818	0.797
AdaBoost	0.198	0.781	0.798	0.789
NB	0.209	0.656	0.759	0.703
DT	0.312	0.664	0.680	0.671

Table 8
Results obtained by the ensemble approach.

(a) Ensemble – BoW (full text) + linguistic features.

	LBR	FCR	FPR	F-measure
LR	0.036	**0.976**	**0.964**	**0.971**
SVM	0.037	0.971	**0.964**	0.967
AdaBoost	0.040	0.961	0.960	0.961
Bagging	0.041	0.959	0.959	0.959
RF	0.053	0.962	0.947	0.954
DT	0.099	0.902	0.901	0.901
NB	**0.016**	0.298	0.951	0.454

(b) Ensemble – BoW (truncated text) + linguistic features.

	LBR	FCR	FPR	F-measure
LR	**0.024**	**0.954**	**0.976**	**0.965**
SVM	0.025	0.949	0.975	0.961
RF	0.041	0.946	0.958	0.952
Bagging	0.035	0.933	0.964	0.949
AdaBoost	0.046	0.940	0.954	0.947
DT	0.099	0.902	0.901	0.901
NB	0.224	0.690	0.755	0.721

a stacking approach to automatically combine the predictions of both representations.

5.4. Ensemble of predictions using different sets of features

For a given test document, if the class predicted by the classifier trained with BoW is different from the class predicted by the classifier trained with linguistic-based features, the class with the highest probability is chosen. The results obtained by this approach are presented in Table 8.

It is clear that the ensemble approach was not effective in the classification of the full texts. However, in the experiments with the truncated texts, the results were higher than those obtained with both BoW and linguistic-based features. Moreover, the best ensemble approach was the one that combined the predictions obtained with the LR. In the experiment with full text, the ensemble of LR wrongly blocked only 3.6% of true news, at the same time that it was able to detect more than 97.6% of fake news. In the experiment with truncated text, the ensemble of LR was able to detect 95.4% of fake news with the price of wrong blocking only 2.4% of true news.

5.5. Stacking of classifiers trained with different sets of features

In this section, we propose a stacking approach that uses a meta-classifier trained with the probabilities given by two individual classifiers. The first one is the LR trained with the linguistic-based features, and the second one is the LR trained with BoW-based feature vectors. Fig. 5 presents an overview diagram of this approach.

As shown in Fig. 5, in the training stage, each training example is represented by two feature vectors: FS_1 (the vector based on linguistic features) and FS_2 (the vector based on BoW). All feature vectors are presented to the module of *transformation of the training set*. This module performs n rounds of training and classification, where n is the number of examples in the training set. In each round, it creates two predictive models using LR, one for each set of feature vectors. In the jth round, the jth training example is classified by the two models trained with the other examples. Then, a new feature vector is created with two dimensions, where the ith element of the vector is the probability of the example being a fake news given by the ith predictive model. The new feature vectors generated by the module of *transformation of*

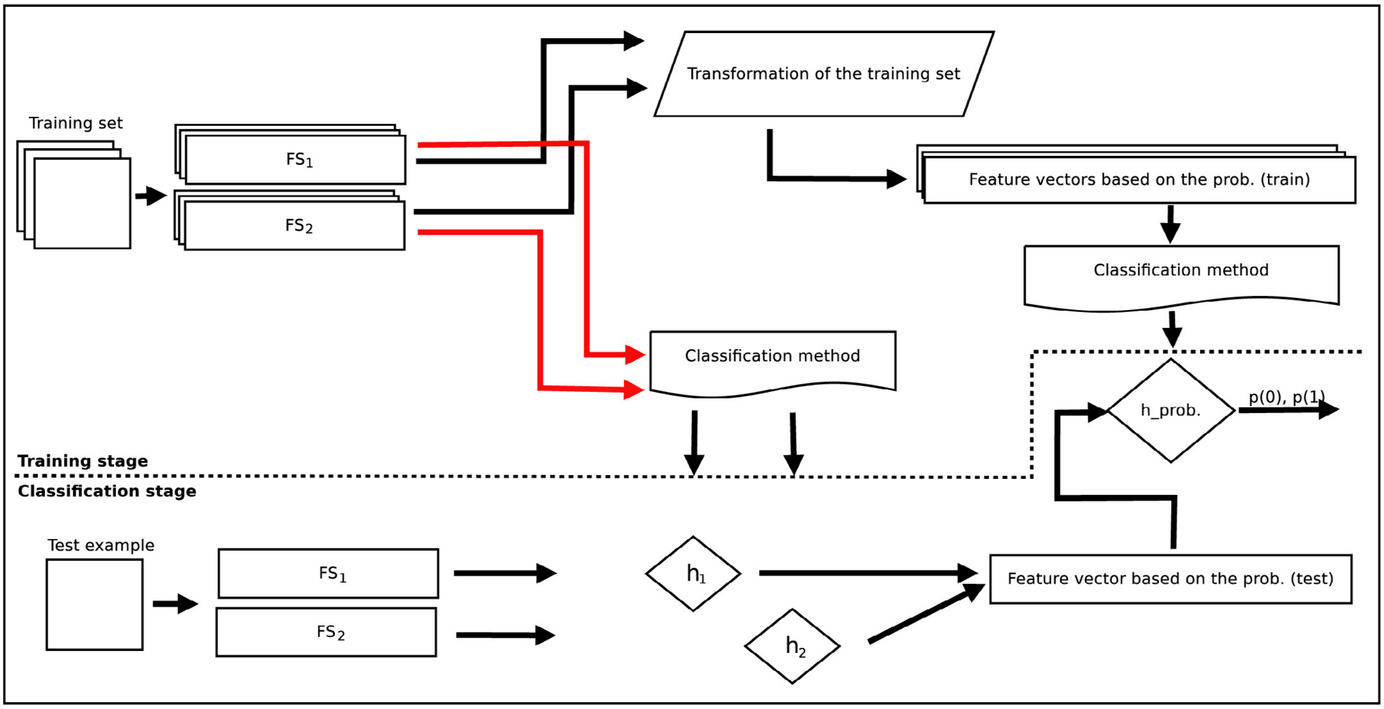


Fig. 5. Overview diagram of the stacking approach.

Table 9
Results obtained by the stacking approach.

	LBR	FCR	FPR	F-measure
Stacking – BoW (full text) + ling. feature	0.036	0.978	0.964	0.971
Stacking – BoW (truncated text) + ling. feature	0.030	0.959	0.970	0.964

the training set are used to train another classification method (LR) that generates a meta-classifier *h_prob*.

In the test stage, an unseen example is also represented by the two feature vectors (FS_1 and FS_2). The *i*th feature vector is presented to the predictive model h_i. Then, a new feature vector is created, where the *i*th element is the probability of the example being a fake news given by the *i*th model. This new feature vector is classified by the meta-classifier *h_prob* that returns the value of $p(0)$ (probability of the example being a legitimate news) and $p(1)$ (probability of the example being a fake news).

We use the LR method in the stacking approach because it is fast and obtained good results in the previous experiments. The results obtained by this approach are shown in Table 9.

In the experiment with full texts, the stacking approach was able to detect 97.8% of fake news with the price of wrong blocking only 3.6% of true news, being superior to the best results obtained with BoW and linguistic-based features. In the experiment with truncated texts, the stacking approach wrongly blocked only 3% of true news, at the same time that it was able to detect 95.9% of fake news, which is a superior performance to that obtained with BoW and linguistic-based features individually. We can also note that the results obtained by the ensemble approach are similar to the score obtained by the stacking approach. For example, the best F-measure of the ensemble approach in the experiment with full text is equal to that of the stacking approach (0.971). In the experiment with truncated text, the difference between the best F-measure obtained by the ensemble approach and the stacking approach was only 0.001.

5.6. Comparison with previous approaches

In this section, we present a comparison between the results of this study and the results obtained in Monteiro et al. (2018). Table 10 summarizes the best results we have obtained. Since Monteiro et al. (2018) have performed experiments only with truncated texts, our results with full texts are not shown in this table.

Table 10 shows that previous results in the literature obtained on the Fake.Br corpus (for truncated texts) are inferior to the ones we present in this study. Moreover, the linguistic features extracted from Fake.Br corpus performed very poorly in the study of Monteiro et al. (2018), achieving an F-measure of 0.550. This big difference is probably because the following reasons:

- Monteiro et al. (2018) have used only the following linguistic-based features: pausality, emotiveness, uncertainty, and non-immediacy. As we describe in Section 4, besides the features used by Monteiro et al. (2018), we used the following additional features: diversity, average size of the sentences, average size of the words, and number of spelling errors.
- They have not normalized the linguistic-based features, which may have affected the performance of the method they used (linear SVM). On the other hand, we applied the Z-score normalization, since we observed that the range of values of the linguistic features varies widely.
- They have not performed grid-search to find the best regularization parameter of SVM.

Table 10
Comparison between our best results and the results of previous approaches.

	LBR	FCR	FPR	F-measure
Ling. features	0.060	0.941	0.940	0.941
BoW (trunc. text)	0.057	0.932	0.943	0.937
Ensemble – BoW (trunc. text) + ling. features	**0.024**	0.954	**0.976**	**0.965**
Stacking – BoW (trunc. text) + ling. features	0.030	**0.959**	0.970	0.964
Monteiro et al. (2018) (Ling. features)	–	0.53	0.57	0.55
Monteiro et al. (2018) (BoW)	–	0.89	0.88	0.88
Monteiro et al. (2018) (POS tags + semantic classes + BoW)	–	0.89	0.88	0.89

The adapted approach proposed by Pérez-Rosas and Mihalcea (2015) (*i.e.*, BoW, POS tags and semantic classes features with a SVM classifier) – the last line in Table 10 – results in an F-measure of 0.89 in the best case. A straightforward BoW solution achieved an F-measure of 0.88. Our best result (namely, 0.965 with the ensemble approach) outperforms the best performance reported on the same dataset, improving the results in 8.4%. This great difference may have been because (i) Monteiro et al. (2018) have not performed grid-search to find the best regularization parameter of SVM, (ii) they have used the binary term weighting scheme representing the text, and (iii) Monteiro et al. (2018) truncated the longer texts (considering number of words) to the size of the corresponding counterparts.

The differences between the results obtained in this paper and the results presented in previous approaches show that small changes in the experimental protocol can improve performance in fake news detection and change the conclusions about this challenging classification task.

6. Conclusions

Fake news can cause major problems for humanity, mainly in areas like political, economy, health, and security. Although this is a problem that society has been facing for several centuries, the volume of these messages has been increasing in a frightening way with the advances of instant messaging and social networks. In this paper, we presented a comprehensive analysis of a novel fake news collection in order to find the best features or combination of features and the best machine learning methods to be used for the automatic detection of fake news. Our experiments have been carefully designed and the results can help answer the following research questions:

- *Q1: What are the best current methods for automatic detection of fake news?*
  To answer this question, we compared the performance of the following widely used machine learning methods: LR, SVM, AdaBoost, RF, Bagging, DT, and NB. None of these methods was superior to the others in all experiments. However, the methods that obtained the best results in most of the evaluated scenarios were LR, SVM, and RF. On the other hand, NB and DT, in general, obtained the lowest results.
- *Q2: What is the best feature set for fake news classification?*
  We performed experiments with linguistic-based features and features generated by text representation techniques (BoW, Word2Vec, and FastText). Surprisingly, the results using BoW, in general, outperformed the results obtained using linguistic-based features and even the results obtained by the state-of-the-art Word2Vec and FastText.
- *Q3: What is the impact of different classification strategies for fake news detection?*
  We combined the results obtained with BoW with the results obtained with linguistic-based features using ensemble and stacking of classifiers. The results obtained by the ensemble and the stacking approach outperformed the scores obtained by the individual classifiers, which demonstrated that the combination of the results obtained using the two sets of features is beneficial for detecting fake news.
- *Q4: Can the size of the texts influence the results of the classification?*
  In the previous analysis of the proposed collection, we noted that the average size of true news is higher than fake news. Therefore, we performed experiments with full texts and with truncated ones to check if there is a difference in the results. In general, the results obtained in the experiments with the full texts were higher than the ones obtained with the truncated texts. Then, we believe there may be a bias in the dataset in relation to the text size and, therefore, the results with the truncated texts probably best represent the results that would be obtained in a real-world application. Classifiers trained with full texts can easily be tricked by people who write fake news if they write longer fake texts.

In future research, we intend to investigate fake news detection using text representation techniques that generate sentence embeddings (*e.g.*, Doc2Vec and Sent2Vec). The challenge of using this type of technique is that no public pre-trained model is available in Portuguese. Therefore, a large repository of documents in Portuguese and great computational power is required to train the sentence embedding models to be used in fake news detection.

We also intend to investigate fake news classification using word embedding models trained with a corpus composed not only of well-written texts but also with noisy language documents, such as documents extracted from Twitter or other social networks.

Finally, we aim to study other types of deception news, such as half-truth and news with satirical content.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit authorship contribution statement

Renato M. Silva: Methodology, Software, Investigation, Writing - original draft. **Roney L.S. Santos:** Methodology, Software, Investigation, Data curation, Writing - original draft. **Tiago A. Almeida:** Conceptualization, Validation, Writing - review & editing, Supervision. **Thiago A.S. Pardo:** Conceptualization, Validation, Data curation, Writing - review & editing, Supervision.

Acknowledgments

We gratefully acknowledge the support provided by São Paulo Research Foundation (FAPESP; grants #2017/09387-6 and #2018/02146-6), the Coordination for the Improvement of Higher Education Personnel (CAPES; Finance Code 001), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Research Office (PRP N. 668) of the University of São Paulo.

References

Angeli, G., & Manning, C. D. (2014). Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 534–545).

Appling, D. S., Briscoe, E. J., & Hutto, C. J. (2015). Discriminative models for predicting deception strategies. In *Proceedings of the 24th international conference on world wide web* (pp. 947–952).

Bertaglia, T. F. C., & Nunes, M. d. G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd workshop on noisy user-generated text* (pp. 112–120).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual acm workshop on computational learning theory (colt'92)* (pp. 144–152). Pittsburgh, PA, USA: ACM.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, California, USA: Wadsworth International Group.

Burgoon, J. K., Buller, D. B., Guerrero, L. K., Afifi, W. A., & Feldman, C. M. (1996). Interpersonal deception: Xii. information management dimensions underlying deceptive and truthful messages. *Communication Monographs, 63*(1), 50–69.

Chakraborty, A., Paranjape, B., Kakarla, S., & Ganguly, N. (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 ieee/acm international conference on advances in social networks analysis and mining (asonam)* (pp. 9–16).

Charles F. Bond, J., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234.

Chesney, S., Liakata, M., Poesio, M., & Purver, M. (2017). Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 emnlp workshop: Natural language processing meets journalism* (pp. 56–61).

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS one, 10*(6), e0128193.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). 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).

Cortes, C., & Vapnik, V. N. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. doi:10.1007/BF00994018.

Duran, N. D., Hall, C., McCarthy, P. M., & McNamara, D. S. (2010). The linguistic correlates of conversational deceprion: comparing natural language processing technologies. *Applied Psycholinguistics, 31*(3), 439–462.

Ferreira, W., & Vlachos, A. (2016). 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.

Fornaciari, T., & Poesio, M. (2013). Automatic deception detection in italian court cases. *Artificial Intelligence and Law, 21*(3), 303–340.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th international conference on machine learning (icml'96)* (pp. 148–156). Bari, Italy: Morgan Kaufmann.

George, J. F., & Keane, B. T. (2006). Deception detection by third party observers. *Paper presented at the deception detection symposium, 39th annual hawaii international conference on system sciences*.

Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2018). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1930–1940).

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., & Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of symposium in information and human language technology* (pp. 122–131).

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., et al. (2015). The quest to automate fact-checking. *World*.

Hauch, V., Blandn-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review, 19*(4), 307–342.

Hauch, V., Masip, J., Blandón-Gitlin, I., & Sporer, S. L. (2012). Linguistic cues to deception assessed by computer programs: A meta-analysis. In *Proceedings of the workshop on computational approaches to deception detection* (pp. 1–4).

Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies, 22*, 13–36.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers* (pp. 427–431). Association for Computational Linguistics.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., & Menczer, F. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs, 66*(2), 125–144.

Lochter, J. V., Pires, P. R., Bossolani, C., Yamakami, A., & Almeida, T. A. (2018). Evaluating the impact of corpora used to train distributed text representation models for noisy and short texts. In *2018 international joint conference on neural networks (ijcnn)* (pp. 1–8). doi:10.1109/IJCNN.2018.8489355.

Méndez, J. R., Iglesias, E. L., Fdez-Riverola, F., Díaz, F., & Corchado, J. M. (2006). Tokenising, stemming and stopword removal on anti-spam filtering domain. In *Proceedings of the 11th spanish association conference on current topics in artificial intelligence (caepia'05)* (pp. 449–458). Santiago de Compostela, Spain: Springer-Verlag.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems (nips'13)* (pp. 3111–3119). Lake Tahoe, Nevada, USA: Curran Associates Inc.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., & Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *13th international conference on computational processing of the portuguese language (propor'2018)* (pp. 324–334). Canela, Rio Grande do Sul, Brazil: Springer International Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*(1), 547–577.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. *CoRR, abs/1708.07104*.

Pérez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural deception detection. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 440–445).

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1120–1125).

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 231–240).

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931–2937).

Rubin, V. L. (2014). Talip perspectives, guest editorial commentary: Pragmatic and cultural considerations for deception detection in asian languages, *13*(2), 10:1–10:8.

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology, 52*(1), 1–4.

Rubin, V. L., & Conroy, N. J. (2011). Challenges in automated deception detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology, 48*(1), 1–4.

Rubin, V. L., Conroy, N. J., Chen, Y., & Cornwell, S. (2016). 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).

Salton, G., & McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Silva, R. M., Alberto, T. C., Almeida, T. A., & Yamakami, A. (2017). Towards filtering undesired short text messages using an online learning approach with semantic indexing. *Expert Systems with Applications, 83*, 314–325. doi:10.1016/j.eswa.2017.04.055.

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th international conference on computational linguistics* (pp. 3346–3359).

Verhoeven, B., & Daelemans, W. (2014). Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the international conference on language resources and evaluation (lrec)*.

Vlachos, A., & Riedel, S. (2014). Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (pp. 18–22). Baltimore, MD, USA: Association for Computational Linguistics. doi:10.3115/v1/W14-2508.

Volkova, S., Shaffer, K., Jang, J. Y., & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 647–653). Association for Computational Linguistics.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

Wang, W. Y. (2017). "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*.

Yu, H.-F., Huang, F.-L., & Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning, 85*(1-2), 41–75. doi:10.1007/s10994-010-5221-8.

Zhang, H., Wei, S., Tan, H., & Zheng, J. (2009). Deception detection based on svm for chinese text in cmc. In *International conference on information technology: New generations* (pp. 481–486).

Zhou, L. (2005). An empirical investigation of deception behavior in instant messaging. *IEEE Transactions on Professional Communication, 48*(2), 147–160.

Zhou, L., Burgoon, J., Twitchell, D., Qin, T., & Nunamaker Jr, J. (2004a). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139–165.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker Jr, J. F. (2004b). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems, 20*(4), 139–166. doi:10.1080/07421222.2004.11045779.

Zhou, L., Twitchell, D. P., Qin, T., Burgoon, J. K., & Nunamaker, J. F. (2003). 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*.

Zhou, L., & Zhang, D. (2008). 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.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. ACM Computing Survey, *51*, 32:1–32:36.