Measuring the Impact of Readability Features in Fake News Detection

Roney L. S. Santos\(^1\), Gabriela Wick-Pedro\(^2\), Sidney Leal\(^1\), Oto A. Vale\(^2\), Thiago A. S. Pardo\(^1\), Kalina Bontcheva\(^3\), Carolina Scarton\(^3\)

\(^1\) Interinstitutional Center for Computational Linguistics (NILC), University of São Paulo, São Carlos, Brazil
\(^2\) Department of Languages, Federal University of São Carlos, São Carlos, Brazil
\(^3\) Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

rongeysantos@usp.br, gwpedro@estudante.ufscar.br, sidleal@gmail.com, otovale@ufscar.br, taspardo@icmc.usp.br, \{k.bontcheva, c.scarton\}@sheffield.ac.uk

Abstract

The proliferation of fake news is a current issue that influences a number of important areas of society, such as politics, economy and health. In the Natural Language Processing area, recent initiatives tried to detect fake news in different ways, ranging from language-based approaches to content-based verification. In such approaches, the choice of the features for the classification of fake and true news is one of the most important parts of the process. This paper presents a study on the impact of readability features to detect fake news for the Brazilian Portuguese language. The results show that such features are relevant to the task (achieving, alone, up to 92% classification accuracy) and may improve previous classification results.

**Keywords:** Fake news, Readability Assessment

1. Introduction

The term “Fake News” can be defined as fabricated information that mimics news media content in form but not in organizational process or intent (Lazer et al., 2018), as news articles that are intentionally and verifiably false, and could possibly mislead readers (Allcott and Gentzkow, 2017), and simply as low quality news with intentionally false information. The task of detecting fake news is defined as the prediction of the chances of a particular news article being deceptive (Rubin et al., 2015).

According to Rubin et al. (2015), two main categories of fake news detection methods currently stand out: 1) network approaches, in which network information, such as message metadata or structured knowledge network queries can be harnessed to provide aggregate deception measures; and 2) linguistic approaches, in which the content of deceptive messages is extracted and analyzed to associate language patterns with deception. In network approaches, there are fact-checking (or content verification) models, in which the use of other techniques such as graph theory, complex networks and question answering, being able to link the subjects of an statement that refers to various factual characteristics, such as “is a”, “member of” and “is married to”, among others (Ciampaglia et al., 2015). In the case of linguistic approaches, it is hypothesized that fake news show linguistic clues that make them detectable when compared to true news, since it is believed that the deceiver unconsciously reflects its deceptive behavior and readability issues in the text.

The choice of language clues (i.e., features) that can be used to classify news as true or false vary considerably. Previous work uses n-grams, lexical measures (e.g., density), morphosyntax (temporal references and verbal cues (Volkova et al., 2017)), syntax (complexity and structure), semantics (cognition), psycholinguistics (Pérez-Rosas and Mihalcea, 2015), Pérez-Rosas et al., 2017, stylistic patterns (Potthast et al., 2018), sentiment (Appling et al., 2015), subjectivity (Vikatos et al., 2017), credibility (Castillo et al., 2015), among others.

In this scenario, readability assessment can also provide information for identifying fake news. Readability indicates, according to Dubay (2007), the ease of reading of a text, which is caused by the choice of content, style, structure and organization that meets the prior knowledge, reading ability, interest and motivation of the audience. Readability features usually measure the number of complex words, long words and syllables, the grade level, text cohesion, among others. This way, readability analysis takes into consideration all linguistic levels for computing its features, which we believe that can bring improvements to the tasks of fake news analysis.

This paper aims to explore readability features for distinguishing fake news from true news for Brazilian Portuguese. We show that readability features are, indeed, relevant to the task. Alone, they achieve up to 92% in accuracy, which outperform current results in the area (Monteiro et al., 2018; Faustini and Covões, 2019; Okano and Ruiz, 2019) and shows that such features are discriminatory of fake and true news. In addition, when readability features are incorporated in a previous model, further improvements are achieved, with up to 93% of accuracy.

The remainder of the paper is divided as follows: Section 2 briefly describe the main related work; Section 3 explains our readability features, as well as the tools and resources used to obtain them; whilst Section 4 reports our experiments. Finally, Section 5 concludes this work and suggests future directions.

2. Related Work

Readability assessment is the task of identifying the complexity of a text in order to predict the difficulty a reader of a certain level of literacy will have to understand it. Automatic Readability Assessment has been extensively studied over the last century since the first formulas for selecting reading material in the US education system appeared,
In the last decade, the task of identifying text properties for the English language has been addressed by approaches that evaluate the ease of comprehension of journalistic texts (Dell’Orletta et al., 2011), as well as how much textual genres influence their understanding (Del Orletta et al., 2014). The addition of more robust features, using human cognition as a premise, has also been addressed through syntax features (Vajjala and Meurers, 2016), psycholinguistic features (Howcroft and Demberg, 2017) and eye-tracking features (Singh et al., 2016), with the best approach reaching around 86.62% test accuracy (Gonzalez-Garduño and Søgaard, 2018). For the Portuguese language, some previous works deals with readability assessment between original literary texts and their translations (to English), showing that metrics are affected by the specificities of each language such as Flesch Reading Ease Index (Pasqualini et al., 2011). Other work uses readability features for text simplification (Aluisio et al., 2010), Scarton et al., 2010), text categorization (Branco et al., 2014), teaching (Curtu et al., 2016) and also to infer psycholinguistic features (Santos et al., 2017).

Regarding the use of readability features in fake news detection for the English language, most previous work still uses data from social media (Buntain and Golbeck, 2017; Bourgonje et al., 2017; Reyes and Palafox, 2019) and makes use of detecting deception content in texts from different aspects, as in Pérez-Rosas and Mihalcea (2013) that separate sentences as produced by gender and age. In addition to feature sets such as unigrams, syntax, and semantics, the authors use a feature set with the readability measurements Flesch-Kincaid and Gunning Fog Index to classify sentences and achieve a 64% accuracy in predicting deceptive sentences when applied to sentences related to the deceive age.

In relation to fake news detection in journalistic news, Pérez-Rosas et al. (2017) extract 26 features from news articles that include Flesch Reading Ease Index, Flesch-Kincaid, Gunning Fog Index and what the authors call content features, such as number of complex words, long and short words, word types, among others. When compared to other linguistic features, readability features achieve the best results in a dataset of fake news collected by crowdsourcing with 78% of accuracy and 79% of F-measure when looking at fake news, but when the readability features were analyzed in a dataset collected from tabloid and entertainment-oriented publications, the results drop considerably to just 50% of accuracy.

Potthast et al. (2018) aim to verify how hyperpartisan news texts are written (writing style). Hyperpartisan can be defined as a kind of "news" that is typically extremely onesided, inflammatory, emotional, and often fraught with untruths. In Potthast et al. (2018), hyperpartisan news are compared to fake news. The authors use 10 other readability measures, which included the most common measures, such as Flesch-Kincaid and Gunning Fog Index and more robust ones, such as Coleman-Liau index (Coleman and Llau, 1975) and McAlpine EFLAW Score (McAlpine, 2005), in addition to the most common linguistic measures and metadata such as word frequency, number of external links, number of paragraphs, and the average length of each paragraph. The authors conclude that the writing style does not contribute in general to verify the veracity of the news, with only 61% of F-measure.

For the Portuguese language, Monteiro et al. (2018) introduce the first fake news dataset for this language, called Fake.Br Corpus. From this dataset, the authors extract linguistic features to detect fake news, such as part-of-speech tags, semantic classes (provided by LIWC (Pennebaker et al., 2015) – for Portuguese (Balage Filho et al., 2013)), bag of words and four features proposed by Zhou et al. (2004): pausality (frequency of pauses in a text), emotionness (indication of language expressiveness in a message), uncertainty (indication of uncertainty in a text, measured by the number of modal verbs) and non-immediacy (indication of impersonal sentences, measured by the number of 1st and 2nd pronouns). Using the SVM technique, the authors achieve 89% of accuracy when using a bag of words approach. Features such as pausality and emotionness do not help to significantly improve the results, achieving around 55% accuracy.

With the use of Fake.Br Corpus news, Faustini and Covões (2019) propose to classify politic news as fake or true using an algorithm called DCDistanceOCC, whose main idea is to match the number of features with the number of problem classes. The authors use 14 features, such as proportion of uppercase characters, exclamation and question marks, number of unique words, sentences and characters, part-of-speech tags and proportion of spell errors, achieving 67% accuracy in classification with the DCDistanceOCC algorithm. A wider range of linguistic clues are used by Okano and Ruiz (2019), who analyze the 76 features proposed by Hauch et al. (2015), such as type-token ratio, number of past and present verbs and semantic ones extracted from LIWC as motion, feel and sadness, to the Portuguese language. From this, the authors select features that performed best on the effect size measure (Hedges and Olkin, 2014), such as type-token ratio, quantity of auxiliary, past and present tense verbs. These features are applied to machine learning algorithms such as SVM, Random Forest and Logistic Regression and the authors achieve 75% accuracy in classification of Fake.Br Corpus news.

Despite the efforts already made in the field of readability classification, there are still few approaches that explore readability features and their relation to fake news, especially for the Portuguese language. Previous attempts that focus on the classification of fake news for Portuguese use purely lexical (e.g., bag of words), morphosyntactic and semantic features (usually based on LIWC), without trying readability features.

3. Readability Features

In relation to feature extraction tools, one of the main tools used for extracting readability metrics is Coh-Metrix (McNamara et al., 2014), which was originally developed for the English language and extracts cohesion and coherence metrics from a text. Coh-Metrix version 3.0 implements 106 metrics, grouped into 11 categories, as follows:
A form of type/token ratio that is less sensitive to text size, computed according to the following equation:

\[ N^{V - 0.165} \]  

where \( N \) is the number of tokens and \( V \) is the number of types, in which richer texts produce lower values. The Honoré Statistic (Honoré, 1979) is similar to the Brunet’s Index, but based on the vocabulary of the text. It is computed as follows:

\[ \frac{100 \log N}{1 - \frac{V}{V_1}} \]  

where \( V_1 \) is the number of words in the vocabulary that have unique occurrences. The Dale Chall Formula (Dale and Chall, 1948) combines the amount of unfamiliar (or difficult) words with the average amount of words per sentence. This feature also has an equivalence with school levels and its value is obtained by the equation below:

\[ 0.1579 \left( \frac{DW}{W} \right) + 0.0496 \left( \frac{CW}{S} \right) + 3.6365 \]  

where \( DW \) is the number of difficult words, i.e., unfamiliar words that are not in the basic vocabulary known to fourth graders, \( W \) is the average number of words per sentence and \( S \) is the number of sentences. Finally, the Gunning Fog Index (Gunning, 1968) indicates the readability of the text, specifying the grade level required to understand it. The value of this attribute is obtained by the following equation:

\[ 0.4 \left( \frac{nW}{S} + 100 \left( \frac{CW}{nW} \right) \right) \]  

where \( nW \) is the number of words in document and \( CW \) is the number of complex words in document, indicated by the words that have three or more syllables, in which the higher the metric is, the greater the complexity and grade level required are.

It is worth mentioning that, except for the Flesch Index, the equations related to the features of the Classic category were not adapted for Portuguese, so we used the same equations as in English. However, in the Dale Chall Formula and Gunning Fog Index, the difficult words and complex words for Portuguese were extracted by a list of Brazilian Portuguese simple words (Biderman et al., 2004).

Cohesion features measure the logical-semantic connection between parts of the text, representing the connection of ideas, usually marked by the use of conjunctions, prepositions, adverbs or verbal phrases. In the case of features used in this paper, referent (nouns or pronouns) and content words (nouns, verbs, adjectives and adverbs) characteristics are observed in sentence pairs or adjacent sentence pairs. Sentence pairs are all possible combinations of two sentences of the text, for example, in a 3-sentence text, the sentence pairs analyzed would be: 1-2, 1-3, 2-3. Adjacent sentence pairs are all possible combinations of 2 sentences in sequence, for example, in a 5-sentence text, the combinations analyzed would be: 1-2, 2-3, 3-4 and 4-5.

In relation to Psycholinguistic features, they try to measure the human cognitive processing. In addition to the content words, four features that are widely used in the literature (Graesser et al., 2011; McNamara et al., 2014) were
also extracted, which are: i) concreteness – the level of abstraction associated with the concept a word describes; ii) imageability – the ability of a given word to arouse mental images; iii) familiarity – the frequency of exposure to a word; and iv) age of acquisition – the age at which a given word is learned by a speaker. The features were extracted from the news in the Fake.Br corpus (Monteiro et al., 2018; Silva et al., 2020), which contains aligned 3,600 false and 3,600 true news in plain text individual files, collected in a 2-year time interval, from January 2016 to January 2018. The news are divided in six big categories in relation to their subjects: politics (58%), TV & celebrities (21.4%), society & daily news (17.7%), science & technology (1.5%), economy (0.7%) and religion (0.7%). This dataset was chosen because it is open access and has already been widely used for research of fake news detection for Portuguese (Faustini and Cov˜oes, 2019; Okano and Ruiz, 2019; Monteiro et al., 2019; Gomes Jr and Frizzo, 2019; Cordeiro and Pinheiro, 2019).

As the fake news in the Fake.Br Corpus are generally smaller in size than the true ones, the non-normalized features were not computed. Using non-normalized features might produce biased results that do not contribute to our investigation of the impact of readability features. Due to this fact, 24 features were removed, which includes, for instance, quantity of words, sentences, paragraphs, verbs, adverbs, content words and features that deal with maximum and minimum values – such as maximum quantity of adjectives or minimum number of words per sentence in the news. The subtitles feature, which refers to the intermediate titles throughout the text, was excluded because none of the news contained intermediate titles in their text structure. Thus, the remaining 165 features were computed for each text in the corpus.

### 4. Experiments and Results

Seven experiments were carried out, including all the 165 features as well as only those that we define as readability features (Table 1). The experiments were organized as follows: E01 – experiment with the most important features selected through feature selection (FS) approaches; E02 – experiment with all the 165 features; E03 – experiment with all different features selected by the FS approaches; E04 – experiment with the features shown in Table 1; E05 – experiment with the approach and features used by [Monteiro et al. (2018)](https://scikit-learn.org/stable/index.html) extended with all the features of E02; E06 – experiment with the approach and features used by [Monteiro et al. (2018)](https://scikit-learn.org/stable/index.html) extended with the features of E03; E07 – experiment with the approach and features used by [Monteiro et al. (2018)](https://scikit-learn.org/stable/index.html) extended with the features of E04.

We followed [Monteiro et al. (2018)](https://scikit-learn.org/stable/index.html) and used SVM (Cortes and Vapnik, 1995), since it was the best model and we adopted with the standard parameters of Scikit-learn [Pedregosa et al., 2011]. We computed the traditional evaluation measures of precision, recall, F-measure and general accuracy in a 5-fold cross-validation strategy. In E01 experiment, FS approaches from Scikit-learn were used. We chose five FS approaches: Univariate Selection (US) with ANOVA F-Value (Stable and Wold, 1989), Recursive Feature Selection (RFE) (Guyon et al., 2002), Feature Importance with either Extra Trees Classifier (FI-ETC), Random Forest Classifier (FI-RFC) and Information Gain (IG) (Kent, 1983). These approaches were selected based on the most popular and commonly used techniques in the Machine Learning area (Chandrashekhar and Sainin, 2014; Khalid et al., 2014; Miao and Niu, 2016).

The US approach uses statistical tests to select the features that have the strongest relationship with the output variable. The RFE approach is a method that fits a model and removes the weakest features until the pre-defined number of features is reached. The FI approaches fits either a number of randomized decision trees (FI-ETC) or decision tree classifiers (FI-RFC) on various sub-samples of the dataset and uses the average to improve the predictive accuracy and control over-fitting. Lastly, the IG approach measures the

| Feature Code | Feature Code |
|--------------|--------------|
| imageabilidade | imageabilidade |
| familiarity | familiarity |
| ratio | ratio |
| simplicity | simplicity |
| concreteness | concreteness |
| imageability | imageability |
| acquired | acquired |
| imageability | imageability |

Table 1: Readability features explored in this work
reduction in entropy by splitting a dataset according to a
given value of a random variable. Table 2 shows the 10
most important features that were selected by each FS ap-
proach.
In total, 23 different features were selected by the
five FS approaches. Three features appears in all FS
approaches results: brunet, pronoun diversity and span mean. The first feature is in the Clas-
ic category and it was better detailed in Section 5
The pronoun diversity feature (Morphosyntac-
tic category) measures the ratio of pronoun types and the
number of pronoun tokens in the text. The span mean feature is in Semantic category and it is related to the La-
tent Semantic Analysis (LSA) (Dumais et al., 1988), which
measures the mean of similarity among all pairs of sentence
in texts. Consider the text below as an example to explain this feature:

It was Senator Flávio Arns (PT-PR) who sug-
gested the inclusion of the object among the items
in the uniform of elementary and high school stu-
dents in municipal, state and federal schools. He
defends the measure as a way to protect children
and adolescents from the ills caused by overex-
posure to sunlight. If the idea is approved, stu-
dents will receive two annual sets, complete with
footwear, socks, pants, and t-shirt.

The example has three sentences and two pairs of adjacent
sentences. The LSA similarity between the first and sec-
ond sentences is 0.084, and the similarity between the sec-
ond and third sentences is 0.063. In this case, the average
between these values is 0.0735. Here, we can realize that
there is semantic included in the calculation of this feature,
by the matching of the words in each sentence.

Table 3 shows statistical analysis made about these features
that were selected in all FS approaches. The statistical anal-
ysis was performed using the Scipy implementation of the
Kolmogorov–Smirnov test (Kolmogorov, 1933; Smirnov,
1948), a non-parametric test that quantifies a distance be-
 tween the empirical distribution functions of two samples.
This test was chosen because it is one of the most useful and
general non-parametric methods for comparing two sam-
ples (Corder and Foreman, 2014).

The Kolmogorov–Smirnov test received all values of the
selected attributes separated by their class. For exam-
ple, for the feature brunet, all values of this feature in
true samples were compared with all values of this fea-
ture in fake samples. With the level of significance \(\alpha = 0.01\), we have that the \(p\)-value of the features brunet,
pronoun diversity and span mean is less than \(\alpha\), con-
cluding that the value distributions of the attributes ana-
yzed by each class are different. Because the distributions
are different, the attributes are significant for use in ma-
chine learning approaches.

---

The statistical analysis was performed for all features, so
we could check which features are statistically significant
to be compared with features selected by FS approaches.
Eight features showed no statistical significant difference
because their \(p\)-value are higher than the level of signif-
icance \(\alpha = 0.01\): average number of repeated referents
in the sentence pairs of the text, average number of para-
graphs, standard deviation of number of paragraphs, pro-
portion of verbs in participle in relation to all verbs in the
text, proportion of personal and possessive pronouns in sec-
ond persons relative to all personal and possessive pronouns
in the text, proportion of personal in third persons relative to
all personal and possessive pronouns in the text and propor-
tion of content words with value of imageability between
1 and 2.5. Thus, none of these features were selected by
FS approaches, but one of them, the average number of re-
peated referents in the sentence pairs of the text, was in-
cluded in those features that we defined as readability fea-
tures.

After analysing the most important features, the classifier
performance was verified when applying the features se-
lected by each of the FS approaches. Therefore, tests were
performed with the most important features, starting from
10 to the maximum number of features. Table 4 shows the
accuracy values of the proposed classifiers in the E01 ex-
periment.

Overall, the features selected by the RFE approach had the
best results with 40 features, being equaled by the IG when
used 70 features, and in the rest of the test all FS approaches
had similar accuracy results, as shown in Figure 1. Interest-
ingly, when selected most features (130 and all features),
the results were slightly worse than the best result, suggest-
ing that for the fake news classification task, less features
show to be important.

Figure 1: Variation in the accuracy values for each FS ap-
proach

The E02 was performed with all the 165 features from Coh-
Metrix-Port and Coh-Metrix-Dementia. The E03 was per-
formed with the 23 different features selected by the FS ap-
proaches, shown in Table 2. The next experiment was E04,
which brought together the 17 features that were defined as
directly related to readability, shown in Table 1.

The last three experiments were a combination of one of the
best fake news classification results for Portuguese with the
features used here in this work. In [Monteiro et al., 2018],

| Number of Features | US  | RFE | FI-ETC | FI-RFC | IG  |
|--------------------|-----|-----|--------|--------|-----|
| 10                 | 0.92| 0.92| 0.92   | 0.91   | 0.91|
| 40                 | 0.92| 0.94| 0.91   | 0.91   | 0.92|
| 70                 | 0.92| 0.93| 0.91   | 0.91   | 0.93|
| 100                | 0.92| 0.92| 0.91   | 0.92   | 0.92|
| 130                | 0.91|     |        |        |     |
| 165                | 0.91|     |        |        |     |

Table 4: Experiment 1 (E01) results with variation on number of selected features

with 92% accuracy. The good results in this experiment

can be explained by the readability analysis between fake

and true journalistic news, which turned out differently and

which the readability features could discriminate the types

of news treated here. When added to the features of the

Monteiro et al. (2018) approach (E07), the results were

worse but still close the best results.

To give more support on the results presented in this paper,

a statistical significance analysis of machine learning mod-

els was performed. We used the Friedman test (Friedman,

1937), a non-parametric test that is used to detect signifi-
cant differences in the results of various test experiments.
The testing process is based on a ranking with some eval-
uation measure (accuracy, f-measure, etc.) of the testing
experiments and there is a comparison of the average rank-
ing of each model. The null hypothesis is that all models
have equivalent results and so the rankings must be equal.
As we intend to affirm that the models proposed in this pa-
per are better than the previous ones, it is expected that the
null hypothesis will be rejected.

Ten samples with 200 news from Fake.Br Corpus were

taken, with 100 fake news and 100 true news, and for each
sample the 7 models used in the experiments presented in
this Section were run. The Friedman test implementation
present in Scipy was used and the evaluation mea-
sure applied was the accuracy. The Friedman test result
returned a value (defined as FTV) of 46.734 and a p-value
of 4.39×10⁻⁷.

The FTV value is compared to a critical value (CV), defined
by the chi-squared distribution table (Wilson and Hilferty,
1931). The null hypothesis is rejected if the value found in
FTV is greater than the critical value. For k = 7 (number
of models to compare), N = 10 (number of samples) and
alpha = 0.05 (to compare significance level), the critical
value is 23.209. Like FTV α > CV and p-value < alpha, the
null hypothesis is rejected, allowing the interpretation that
the models have no equivalent results and that the difference
between them (accuracy as well) is significant.

Overall, the results achieved here were promising. Using

Table 2: Selected features according to different FS approaches

Table 3: Statistical analysis of features selected

features used here in this work. In [Monteiro et al., 2018],

the authors used a linguistic-based approach when repro-
ducing the Pérez-Rosas and Mihalcea (2015) approach for
Portuguese, but did not include readability and syntactic
complexity features in their tests. Therefore, experiments
E05, E06 and E07 use the Monteiro et al. (2018) approach
with: i) the addition of all our features of E02 in the E05
experiment; ii) the 27 features of E03 in the E06 experiment;
and finally, iii) the 17 features of E04 in the E07 experiment.
It is important to highlight that the authors truncated
the true news in order to obtain a normalized comparison
with the other proposed features, so, in the experiments
performed here, the same procedures were performed. Results
for all the experiments (except E01) are shown in Table 5.

When all the features were used (E05) and the features of
Monteiro et al. (2018) approach extended with the fea-
tures of E03 (E06), there is an increase in accuracy and
F-measure of both fake and true news when compared to the
best results in Monteiro et al. (2018) linguistic-based
approach. This can be explained by using not only the read-

ability features that were defined in this paper, but all cat-

gories, including syntactic ones, which Pérez-Rosas and
Mihalcea (2015) had suggested in their approach.

When only readability features were used (E04), the results
showed that readability has influence in the classification,
readability features is important to increase the reliability of the language clues that a deceiver leaves when detecting fake news. The increase of classification accuracy may be a good indicator for future research, when more robust features could be proposed and include other features of higher language levels.

5. Conclusions and Future Works

This paper presents studies and experiments concerning the use of readability features in the task of detecting fake news. Although these features are formed by branches of features from other linguistic levels, such as morphological, morphosyntactic, syntactic, and semantic, the robustness of these features could be a differential in identifying different writing styles in fake news.

Although interesting results are achieved when applying only readability features, it is confirmed that better results are achieved when used in combination with features from other language levels, as works in the field have already showed [Perez-Rosas et al., 2017] [Volkova et al., 2017] [Potthast et al., 2018]. The results presented in this paper outperform the previously proposed works for Portuguese, showing that there is still room for improvement in the choice of features for the detection of fake news.

As future work, it is important to deepen the study of syntactic and semantic features, which proved to be useful for the task. In addition, other readability features can be proposed and used in the classification. Moreover, according [Monteiro et al., 2018], the Fake.Br Corpus does not have news with half truths, i.e., news with some actual facts are told in order to give support false facts [Clem, 2017], which are becoming common in current news. This characteristics of the corpus may have impacted in results, being important for achieving good results in the experiments. As future work, we also aim to deal with this complex case, as well as to measure the impact of readability features in other deception content, such as ironic and satirical news and opinion reviews.

More information about this work may be found at OPINANDO project website[9].

6. Acknowledgements

This work was performed during RS and GW-P visits to the University of Sheffield (UK). Their visits were funded by the SoBigData TransNational Access program (EU H2020, grant agreement: 654024). The authors are grateful to CAPES (Finance Code 001) and USP Research Office (PRP 668).

7. Bibliographical References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–36.

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. Proceedings of the NAACL HLT Fifth Workshop on Innovative Use of NLP for Building Educational Applications, pages 1–9.

Appling, D. S., Briscoe, E. J., and Hutto, C. J. (2015). Discriminative models for predicting deception strategies. In Proceedings of the 24th International Conference on World Wide Web, pages 947–952.

Balage Filho, P. P., Pardo, T. A. S., and Aluisio, S. M. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 215–219.

Biderman, M. T. C., Carvalho, C. S., and Pedroso, O. (2004). Meu primeiro livro de palavras: um dicionário ilustrado do português de A a Z. Ática.

Bourgonje, P., Moreno Schneider, J., and Rehm, G. (2017). From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism, pages 84–89.

Branco, A., Rodrigues, J., Costa, F., Silva, J., and Vaz, R. (2014). Rolling out text categorization for language learning assessment supported by language technology. In Computational Processing of the Portuguese Language, pages 256–261.

Brunet, E. (1978). Le Vocabulaire de Jean Giraudoux: structure et évolution : statistique et informatique appliquées à l’étude des textes à partir des données du Trésor de la langue française. Slatkine.

Buntain, C. and Golbeck, J. (2017). Automatically identifying fake news in popular twitter threads. In 2017 IEEE International Conference on Smart Cloud, pages 208–215.

Castillo, C., Mendoza, M., and Poblete, B. (2013). Predicting information credibility in time-sensitive social media. Internet Research, 23(5):560–588.

---

Table 5: All classification experiments results

| Experiment | Precision Fake | Precision True | Recall Fake | Recall True | F-Measure Fake | F-Measure True | Accuracy |
|------------|----------------|----------------|-------------|-------------|----------------|----------------|----------|
| Monteiro et al. (2018) | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 |
| E02 | 0.94 | 0.89 | 0.88 | 0.94 | 0.91 | 0.91 | 0.91 |
| E03 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 |
| E04 | 0.93 | 0.90 | 0.90 | 0.93 | 0.91 | 0.92 | 0.92 |
| E05: Monteiro et al. (2018) + E02 | 0.95 | 0.91 | 0.91 | 0.95 | 0.93 | 0.93 | 0.93 |
| E06: Monteiro et al. (2018) + E03 | 0.93 | 0.91 | 0.91 | 0.93 | 0.93 | 0.93 | 0.93 |
| E07: Monteiro et al. (2018) + E04 | 0.89 | 0.90 | 0.91 | 0.88 | 0.90 | 0.89 | 0.89 |

---

[9]https://sites.google.com/icmc.usp.br/opinando/
Chandrashker, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28.

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

Clem, S. (2017). Post-truth and vices opposed to truth. *Journal of the Society of Christian Ethics*, 37(2):97–116.

Coleman, M. and Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Cordeiro, P. R. and Pinheiro, V. (2019). Um corpus de notícias do twitter e verificação automática de rumores em língua portuguesa. In *Proceedings of the Symposium in Information and Human Language Technology*, pages 219–228.

Corder, G. and Foreman, D. (2014). *Nonparametric Statistics: A Step-by-Step Approach*. Wiley.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cunha, A. L. V. (2015). Coh-metrix-dementia: análise automática de distúrbios de linguagem nas demências utilizando processamento de línguas naturais. Master’s thesis, University of Sao Paulo.

Curto, P., Mamede, N., and Baptista, J. (2016). Assisting european portuguese teaching: Linguistic features extraction and automatic readability classifier. In *Computer Supported Education*, pages 81–96.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Dell’Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of italian texts with a view to text simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Del’Orletta, F., Montemagni, S., and Venturi, G. (2014). Assessing document and sentence readability in less resourced languages and across textual genres. *International Journal of Applied Linguistics (ITL). Special Issue on Readability and Text Simplification*.

dos Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., and Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for Brazilian Portuguese. In *Text, Speech, and Dialogue. TSD 2017. Lecture Notes in Computer Science*, volume 10415, pages 281–289.

Dubay, W. H. (2007). *Smart Language: Readers, Readability, and the Grading of Text*. Impact Information, Costa Mesa, CA.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 281–285.

Faustini, P. and Covêes, T. (2019). Fake news detection using one-class classification. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, pages 592–597.

Flesch, R. (1979). *How to write plain English: a book for lawyers and consumers*. Harper & Row.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Gomes Jr, L. and Frizzo, G. (2019). Fake news and brazilian politics–temporal investigation based on semantic annotations and graph analysis. In *Proceedings of the Brazilian Symposium on Databases*.

Gonzalez-Garduño, A. V. and Sogaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5118–5124.

Grassner, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Gunning, R. (1968). *The technique of clear writing*. McGraw-Hill.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Hauch, V., Blandón-Gitlin, I., Masip, J., and 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.

Hedges, L. V. and Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Honoré, A. M. (1979). Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, 7:172–177.

Howcroft, D. M. and Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, page 958–968.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173.

Khalid, S., Khalil, T., and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378.

Kincaid, J., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). *Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Research Branch report. Chief of Naval Technical Training, Naval Air Station Memphis.

Kolmosgovor, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’Istituto Italiano degli Attuari*, 4:83–91.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and
Zittrain, J. L. (2018). The science of fake news. Science, 359(6380):1094–1096.

Martins, T. B. F., Ghiraldelo, C. M., Nunes, M. G. V., and Oliveira Junior, O. N. (1996). Readability formulas applied to textbooks in brazilian portuguese. Technical report, ICMSC-USP.

Maziero, E. G., Pardo, T. A. S., and Aluíso, S. M., (2008). Ferramenta de Análise Automática de Inteligibilidade de Córpus (AIC). NILC - ICMC - USP. Available at http://www.nilc.icmc.usp.br/nilc/download/NILCCTR0808-MazieroPardo.pdf.

McAlpine, R. (2005). Global English for Global Business, CC Press.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press.

Miao, J. and Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91:919 – 926.

Monteiro, R. A., Santos, R. L. S., Pardo, T. A. S., de Almeida, T. A., Ruiz, E. E. S., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In Computational Processing of the Portuguese Language, pages 324–334.

Monteiro, R., Nogueira, R., and Moser, G. (2019). Desenvolvimento de um sistema para a classificação de fake-news acoplado à etapa de etl de um data warehourse de textos de notícias em língua portuguesa. In Anais da XV Escola Regional de Banco de Dados, pages 131–140.

Okano, E. and Ruiz, E. (2019). Using linguistic cues to detect fake news on the brazilian portuguese parallel corpus fake.br. In Proceedings of the Symposium in Information and Human Language Technology, pages 181–189.

Pasqualini, B. F., Scarton, C. E., and Finatto, M. J. B. (2011). Comparando avaliações de inteligibilidade textual entre originais e traduções de textos literários (comparing textual intelligibility evaluations among literary source texts and their translations) [in Portuguese]. In Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Pennebaker, J., Booth, R., Boyd, R., and Francis, M. (2015). Linguistic inquiry and word count: Liwc 2015. [http://liwc.wpengine.com/] Acesso em 01/12/2018, 13:04.

Pérez-Rosas, V. and Mihalcea, R. (2015). Experiments in open domain deception detection. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1120–1125.

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

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., and 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), pages 231–240.

Reyes, J. and Palafox, L. (2019). Detection of fake news based on readability. In Reunió Internacional de Intel·ligència Artificial y sus Aplicaciones.

Rubin, V. L., Conroy, N. J., and Chen, Y. (2015). Towards news verification: Deception detection methods for news discourse. In Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, pages 5–8.

Santos, L. B., Duran, M. S., Hartmann, N. S., Candido, A., Paetzold, G. H., and Aluisio, S. M. (2017). A lightweight regression method to infer psycholinguistic properties for brazilian portuguese. In Text, Speech, and Dialogue, pages 281–289.

Scarton, C. and Aluisio, S. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. Linguamática, 2(1):45–61.

Scarton, C., Oliveira, M., Candido Jr., A., Gasperin, C., and Aluisio, S. (2010). SIMPLIFICA: a tool for authoring simplified texts in Brazilian Portuguese guided by readability assessments. In Proceedings of the NAACL HLT 2010 Demonstration Session, pages 41–44.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. Expert Systems with Applications, 146:113199.

Singh, A. D., Mehta, P., Husain, S., and Rajkumar, R. (2016). Quantifying sentence complexity based on eye-tracking measures. Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, page 202–212.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics., 19(2):279–281.

Stahle, L. and Wold, S. (1989). Analysis of variance fit of empirical distributions. The Annals of Mathematical Statistics., 19(2):279–281.

Vajjala, S. and Meurers, D. (2016). Readability-based sentence ranking for evaluating text simplification. CoRR - Computer Research Repository.

Vikatos, P., Messias, J., Miranda, M., and Benevenuto, F. (2017). Linguistic diversities of demographic groups in twitter. In Proceedings of the 28th ACM Conference on Hypertext and Social Media, pages 275–284.

Volkova, S., Shaffer, K., Jang, J. Y., and 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), pages 647–653. Association for Computational Linguistics.

Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. proceedings of the National Academy of Sciences of the United States of America, 17(12):684.
maker Jr., J. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–165.