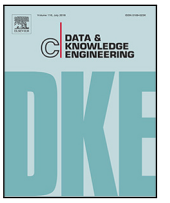
Data & Knowledge Engineering 147 (2023) 102207

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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



An empiric validation of linguistic features in machine learning models for fake news detection

Eduardo Puraivan [a,b,*], René Venegas [c], Fabián Riquelme [b]

[a] Escuela de Ciencias, Universidad Viña del Mar, Chile
[b] Escuela de Ingeniería Informática, Universidad de Valparaíso, Chile
[c] Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso, Chile

ARTICLE INFO

Keywords:
Fake news
Mass media
Natural language processing
Linguistic features
Machine learning

ABSTRACT

The diffusion of fake news is a growing problem with a high and negative social impact. There are several approaches to address the detection of fake news. This work focuses on a hybrid approach based on functional linguistic features and machine learning. There are several recent works with this approach. However, there are no clear guidelines on which linguistic features are most appropriate nor how to justify their use. Furthermore, many classification results are modest compared to recent advances in natural language processing. Our proposal considers 88 features organized in surface information, part of speech, discursive characteristics, and readability indices. On a 42 677 news database, we show that the classification results outperform previous work, even outperforming state-of-the-art techniques such as BERT, reaching 99.99% accuracy. A proper selection of linguistic features is crucial for interpretability as well as the performance of the models. In this sense, our proposal contributes to the intentional selection of linguistic features, overcoming current technical issues. We identified 32 features that show differences between the type of news. The results are highly competitive in the classification and simple to implement and interpret.

1. Introduction

The study of fake news is a hot topic presenting various open problems and technological challenges, approached from different scientific disciplines.

Fake news is false or misleading news, whose main objective is misinformation or manipulation of information. The design and elaboration processes of fake news require structuring the text, packaging, determining its scope, among other aspects. The motivation of those who create fake news is frequently related to economic and socio-political issues. Some people may inadvertently share wrong information, while others will do so on purpose. Regarding the type of content, fake news includes satires or parodies, deceptive content, impostor content, and fabricated content, each of these types being more dangerous than the previous one [1]. There are many other taxonomies of fake news. Among different possible classifications, clickbait, propaganda, satires and parodies, hoaxes, name-theft, and framing, are particular types of fake news [2].

It is known that previous exposure to some news with strong ideas increases the perception of credibility in the face of future fake news that reuses those strong ideas. Therefore, repetition in editorial lines for particular purposes can be a source of biased information and fake news that quickly becomes viral [3]. Furthermore, the appeal to emotions in the generation of fake news increases the probability they are more accepted and disseminated by people [4]. Fake news can be spread through various mass

* Corresponding author at: Escuela de Ingeniería Informática, Universidad de Valparaíso, Chile.
E-mail addresses: eduardo.puraivan@postgrado.uv.cl (E. Puraivan), rene.venegas@pucv.cl (R. Venegas), fabian.riquelme@uv.cl (F. Riquelme).

https://doi.org/10.1016/j.datak.2023.102207
Received 24 July 2022; Received in revised form 25 July 2023; Accepted 28 July 2023
Available online 2 August 2023

media, including newspapers, websites, television, radio, among others. Usually, fake news is installed in these environments and then goes viral through online social networks, through massive interaction between actors, who can collaborate consciously or unconsciously to their diffusion [5].

While fake news is not new, the explosive development of technology and the mass media since the beginning of the 21st century has sharply increased its impact and speed of spread [6]. This expansive development has made fake news a global concern. Currently, the phenomenon is closely related to ideological polarization [7], the generation of hate speech [8], astroturfing [9], post-truth [10,11], and the installation of populisms that deteriorate democratic processes [12]. In addition to the socio-political impacts, there are risks to global health [13] and high economic impacts, since fake news can even affect financial markets [14]. For example, during the COVID-19 Pandemic, fake news has had a negative and important effect, especially around misinformation related to vaccines and health measures [15]. Fake news has not only contributed to misinformation but generated hate speech against certain races and immigrant groups. In addition, fake news produces a negative impact on people's health, as news dangerous to physical integrity has been disseminated, as possible cures and tips on how to fight the virus [16]. According to the World Health Organization, "We're not just fighting an epidemic; we're fighting an infodemic" [17]. In this sense, a closely related problem that has also generated much research is trust prediction [18].

Despite the growing efforts from research, technology, and public policies to detect and prevent the spread of fake news early, the same technological revolution has allowed its diversification. Thus, the processes for generating fake news are increasingly sophisticated and hard to detect [19].

There is currently much research on the automatic detection of fake news [20–22]. In the most common strategy, namely the supervised learning, a classifier is trained with numerous previously labeled news (e.g., as real or false) so that the classifier continues to classify the next news, following patterns learned during the training process. Despite the various very optimistic results obtained with this approach [23], several technical and methodological issues have not already been fully resolved [24]. In addition, hybrid approaches that combine machine learning techniques with natural language processing have also been defined. In particular, it is known that linguistic features can help identify patterns that distinguish fake news from real news in some specific contexts [25]. However, these studies are still scarce, the linguistic features used are few, and their use is not entirely justified. Thus, it is necessary to identify which features are most useful for each context, for which it is necessary to validate them statistically through experimentation.

In this work, we experiment with 88 linguistic features on a dataset of 42 677 news labeled as either real or fake. The linguistic features are grouped into surface information, part of speech (POS), discursive characteristics, and readability indices. From all of them, 56 features are avoided because of their many missing values, low variability, or little significance. The latter left us with 32 features that significantly differ between real and fake news. We use principal component analysis (PCA) to identify the linguistic features with the most significant contribution in the first components. After that, we conduct a trade-off study, by testing different machine learning models on the same context or dataset. Through a k-fold cross-validation, the results show a high average accuracy, specially with XGBoost classifier. Thus, our research pursues the following objectives: (1) Propose a comprehensive list of linguistic features capable of being computed and organized in well-defined categories to extract linguistic information from the news text, and (2) Empirically validate the proposed linguistic features for developing news classification models, showing the way of interpretation in exploratory and explanatory stages.

The paper continues as follows. Section 2 briefly describes different uses of linguistic features to improve the automatic detection of fake news and the main approaches used today to deal with this problem. Section 3 introduces some relevant aspects for the work related to linguistic analysis techniques and the development of machine learning models. Section 4 presents the proposed solution, including methodological aspects, implementation, and dataset. Section 5 shows and discusses the results obtained. Finally, Section 6 ends with the main conclusions of this work and some proposals for future work.

2. Related work

Among the numerous existing surveys on the subject, we can distinguish at least five methods for fake news detection [2]:

- *Natural language processing (NLP)*, that utilizes content analysis techniques and linguistic features to automatically detect fake news [26];
- *Machine learning (ML)*, that uses statistical and computational methods for automatic classification. The main supervised ML techniques include support vector machines (SVM), logistic regression (LR), random forests (RF), naive Bayes (NB), Gaussian naive Bayes (GNB), and k-nearest neighbors (KNN) [27,28];
- *Experts Facts-Checker* and *Crowdsourced* approaches, that involve experts verifying news manually or crowdsourcing efforts to verify fake news through online collaboration, and
- *Hybrid* approach, that combines various methods to achieve a balance between classification costs and accuracy [29–32].

Our aim is to explore a combination of NLP and ML techniques, focusing on significant linguistic features and supervised learning, while avoiding deep learning methods [33] to reduce computational costs and increase interpretability. Below we present a brief review about the use of linguistic features for fake news classification models.

2.1. Linguistic features in fake news detection

The linguistic analysis seeks to objectively measure different *linguistic features* to describe how a speech community uses language.

Several works show the potential of linguistic features in fake news detection models. In [26], features organized in n-grams, punctuation, psycho-linguistic, readability, and syntax are used, achieving an accuracy of 76% in the best cases. Considering most recent works, in [28], the features used are grouped into linguistic patterns, emotions, sentiment, personality traits, communication style, and readability. The authors experiment with different models. The performance (accuracy) of the different systems when one group of features is used are as follows: BERT-LR (58%), Emotion-GNB (66%), Sentiment-RF (59%), LIWC-RF (58%), Personality-RF (62%), Readability-RF (58%), CommunicationStyle-RF (62%). The performance of the different systems on the fact-checkers detection task are as follows: BERT-LR (59%), USE-LR (79%), LSTM-embeddings (56%), CNN-embeddings (53%), CheckerOrSpreader (62%). In [34] a hybrid approach is proposed that combines features of content, social context, and knowledge. As linguistic features, the authors use the number of words, reading ease, lexical diversity, and sentiment. Several models are developed, but only using linguistic features achieve an accuracy of 89.40% through XGBoost. Of all the models developed, the best performance is obtained with the "Linguistic+Fac-verification" model, achieving an accuracy of 94.40% with RF. On the other hand, in [35] different feature extraction techniques named tf-idf, hash-vectorizer, and count-vectorizer are adopted. The proposed method achieved the highest of 72.8% accuracy using an RF classifier and applying the tf-idf algorithm. Also, in [36] the authors created a sequential neural network based on linguistic features that distinguish between fake and real news with 86% accuracy. In [37], through NB and combining linguistic features with structural variables on the Twitter network, a precision of 99.79% is achieved. In [38], a clickbait and linguistic feature-based scheme is used, achieving an accuracy of 77% with the XGBoost classifier.

Besides the above, there are also exploratory studies using linguistic features to characterize text. In [25] an analytical study of the news language is proposed to investigate and identify linguistic features and their contribution to data analysis, to detect, and differentiate between fake and authentic news. In [39], the author perform a forensic linguistic analysis of fake news published in English and Portuguese.

2.2. Limitations

In the results mentioned above, even combining linguistic features with other features and using novel techniques such as XGBoost or BERT, few classifications exceed 90% accuracy. Also, difficulties are observed in selecting and classifying the appropriate linguistic features for each case. To illustrate the latter, we have preferred to analyze some of the previous works.

In [36] a linguistic model is proposed to find the content properties that generate language-guided features. The model extracts syntactic, grammatical, sentimental, and readability features from certain news. The features used are motivated by previous literature. The syntactic features were chosen since they are the most frequently used. For syntactic analysis, they use features related to the number of characters, number of words, number of words in the title, number of stop words, number of words in uppercase, and word density (i.e., number of occurrences of keywords according to the total text). However, none of these linguistic features are syntax indicators. Although the results obtained are positive, they are not very interpretable. Therefore, an adequate classification of the features could allow a previous selection that helps to conduct the analysis.

In [37], the authors consider tweets written in English concerning the events in Hong Kong and a well-defined method for detecting fake news using linguistic and network features. They use ten linguistic features: number of syllables, the average number of syllables, average number of words in the sentence, Flesh-Kincaid, number of long words, number of long and short sentences, number of sentences, number of words, and the ratio between adverbs and adjectives. However, a categorization of features is not established. Most are textual surface features (statistics on syllables, sentences, and words). The most sophisticated are the Flesh-Kincaid (readability) and the ratio between adverbs and adjectives. In short, they use few linguistic features, none of them is adequately justified, and there is no classification of the features used.

Finally, at [25] the authors carry out an analytical study of the language of news to investigate and identify linguistic features and their contribution to data analysis to detect, filter, and differentiate between fake and authentic news. They analyze 40 news items in total under a qualitative approach (unlike the previous works, predictive models are not developed here). The authors use 16 features organized in three main categories (lexical, grammatical, and syntactic features) manually labeled. Regarding linguistic features, the following are declared: frequency of personal pronouns, proper names, adverbs, stative verbs, infinitive verbs, passive voice, reported speech, comparative adjectives, superlative adjectives, modal verbs, citations, conjunctions, long sentences, interrogations, and negations. In general, the description of results does not maintain the idea of the three main categories. By including passive voice, reported speech and quotes are already in unreported categories: passive voice is syntactic, while reported speech and quotes are discursive features.

In summary, few classifications exceed 90% accuracy, and the linguistic features are usually chosen without a clear justification (possibly because they are the most commonly used or available in widely used libraries). We think an adequate categorization of the linguistic features could help their intentional selection. A hasty selection can generate confusion at the methodological level and lead to difficulties in the interpretation of the results. In this work, we use a classification into four categories, described in Section 3.1.

3. Preliminaries

In this section, known and relevant aspects to be considered in this work are described. In particular, we focus on linguistic analysis techniques and machine learning models development.

3.1. Linguistic analysis

For the linguistic analysis, we focus on the lexical-grammar and discursive level of the texts, using automated analysis over different texts data. The linguistic features addressed in this study can be classified into four types:

- *Surface information.* It includes quantitative values from descriptive statistics (mainly totals and means) on paragraphs, sentences, words, and characters.
- *Part of speech (POS).* It consists of POS-tag statistics, including descriptive measurements (total, mean, min, max, std, median) and representation of the frequency (idf, tf-idf) of the morphological categories in the text (adjectives, adpositions, adverbs, pronouns, punctuation marks, verbs, among others).
- *Discursive characteristics.* It integrates a customized configuration of the Simple Natural Language Processing (SiNLP) tool [40], which allows calculating the frequency (i.e., summation of occurrences on a particular ad hoc dictionary) of morphological and discursive categories. The morphological categories include determiners, demonstratives, personal pronouns, adverbs, articles, prepositions, and negations. Among the functional ones are discursive markers, connectors of various types, lexicons for different kinds of academic texts, words of impolite use or insults. These kinds of tools also include the possibility of considering lexicons associated with positive and negative feelings, as well as lexicons related to the linguistic and psychological dimensions of the Linguistic Inquiry and Word Count (LIWC) program [41]. Note that SiNLP also includes some basic surface information features, such as the number of words, number of paragraphs, and number of sentences.
- *Readability indices.* It consists of different readability indices that return a numeric value for each text according to its readability level. These values usually refer to the people's educational level capable of understanding the text. These indices have usually been developed for texts written in English. For this work, we use the Flesch Reading Ease index (FRE) [42] and the Flesch Kincaid Grade Level index (FKG) [43] since they are widely used readability formulas to assess the approximate reading grade level of a text.

Table A.7 presents the 88 features initially considered in this work.

3.2. Multivariate analysis and machine learning techniques

Principal component analysis (PCA) [44,45] is a technique used to describe a dataset in terms of new uncorrelated variables (namely, the components). The components are linear combinations of the original variables. By construction, they are orthogonal with each other. Furthermore, the components are ordered by the amount of original variance they describe, so the first components accumulate most of the variance (information). Thus, this technique helps reduce the dimensionality of the dataset. In other words, PCA can be interpreted as a transformation of the original axes into new dimensions. PCA provides several advantages [46]: (1) extract the most relevant information from the dataset; (2) compress the size of the dataset by keeping only this relevant information; (3) simplify the description of the dataset, and (4) analyze the structure of the observations and the variables.

Cross-validation techniques are used to estimate the accuracy of classification and regression models, seeking to overcome the classic underfitting and overfitting problems [24]. A widely used method is k-fold cross-validation, where the training set is partitioned into "folds" to use one percentage for training and another for validation, e.g., in a proportion of 90%–10%, respectively [47].

Machine learning and deep learning models have shown excellent performance on different application problems [23]. However, despite their excellent performance, some models have scarce value for the industry due to their low interpretability [48]. Such is often the case, for example, with neural networks [49]. Therefore, to achieve transparent and reproducible predictions, it is necessary to develop interpretable machine learning models [50]. In addition, a solution does not necessarily seek to generate an immediate transfer to diverse contexts. The models need to be validated in a controlled experiment and then generalized to other contexts by considering the sensitivity and trade-off analyses [51].

In this work, after selecting a set of linguistic features, we aim to test the performance of different models in the same context, i.e., make a trade-off (a database as fixed context). Once we have satisfactory results for controlled experiments, it will be possible to advance to sensitivity tests, that is, to test the models in different contexts. In the context of fake news detection, note that despite the promising results reported in laboratory studies, the existing models are still immature to move to the sensitivity stage. Moreover, the available datasets still need to be improved, and many have errors, such as those reported before [24].

4. Experimentation

4.1. Methodology

This work continues ideas developed in [52]. A general solution scheme is shown in Fig. 1. From the news texts, it is proposed to calculate linguistic features, organized in surface information, part of speech, discursive characteristics, and readability indices (see details in Section 3.1 and Table A.7). The results can be used to develop news classification models.

Fig. 2 presents three recognizable analysis steps:

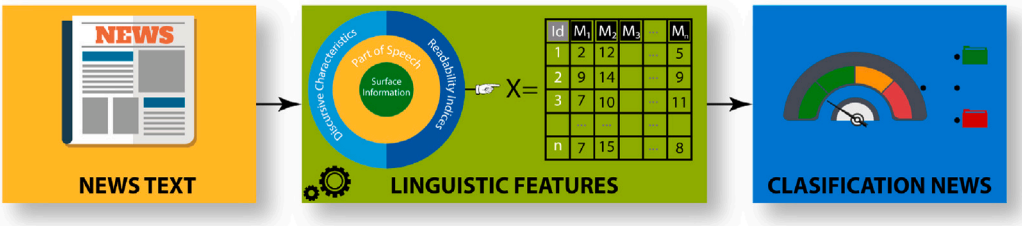


Fig. 1. Work methodology used in this work.

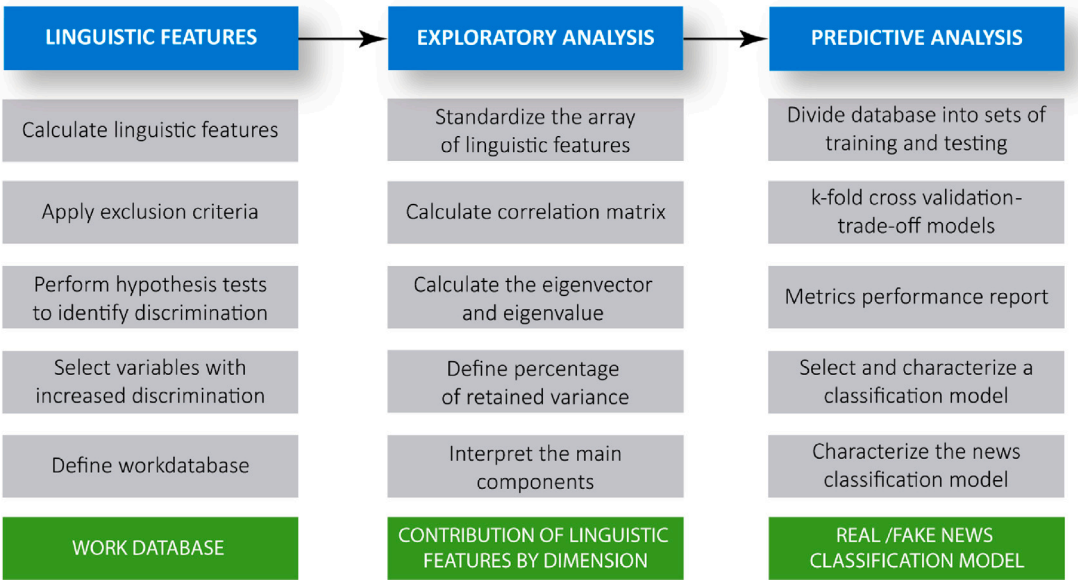


Fig. 2. Analysis methodology used in this work.

- *Linguistic features.* We calculate all the linguistic features shown in Table A.7. To do this, we use a platform that includes different modules of analysis, created in Python and with an online interface named DescApp.[1] The POS features use spaCy[2] as tagger and parser. Then we exclude all those features that do not provide information (without variability or interpretability). After that, we perform hypothesis testing to identify the linguistic features that show significant differences for real and fake news and select those variables with increased discrimination. The output of this step is the working database.
- *Exploratory analysis.* We use principal component analysis (PCA) for data transformation and features extraction. It requires scaling the linguistic features and calculating a correlation matrix. Then eigenvalues and eigenvectors are obtained to identify the proportion of variance explained by each linguistic feature. The output of this step is the characterization of the PCA dimensions in terms of linguistic features.
- *Predictive analysis.* The original data are divided into training and tests. Finally, through k-folder cross-validation, different models (trade-offs) are developed. The output at this stage is a collection of classification models for real and fake news, along with their respective performances.

We used the R programming language [53] for machine learning techniques and statistical methods on a computer with macOS (Catalina), a 2 GHz Intel Core i5 quad-core processor with 16 GB of RAM. For randomization, `set.seed(123)` was used throughout the whole process. For the principal component analysis (PCA), we used the `factoextra` package [54].

[1] http://158.251.88.54:8891
[2] www.spacy.io

Table 1
Mean, standard deviation, and median of the linguistic features considered.

ID	Linguistic feature	Fake news			Real news		
		mean	std	median	mean	std	median
1	Letters per word	4.87	0.38	4.85	5.07	0.25	5.06
2	Number of types	251.52	160.87	224	223.44	130.96	217
3	Number of words	458.99	406.85	381	390.98	272.70	364
4	TTR	0.59	0.08	0.59	0.63	0.10	0.61
5	Adjective mean	4.17	2.23	3.67	3.48	1.33	3.33
6	Adverb total	25.44	23.22	22	14.38	13.08	11
7	Pronoun total	19.68	18.13	16	9.86	10.83	7
8	Verb total	89.32	72.95	77	71.48	52.97	64
9	Verb 3rd person singular present	15.10	13.40	13	8.04	7.52	6
10	Adverbs	4.11	4.26	3	2.99	3.12	2
11	Comparison	4.25	4.46	3	3.33	3.39	2
12	Connectives	19.94	19.83	16	15.79	12.69	14
13	Connectors add information	0.16	0.54	0	0.19	0.51	0
14	Connectors comparison	0.87	1.31	0	0.23	0.59	0
15	Connectors contrast	2.69	2.97	2	2.11	2.35	1
16	Connectors explain	0.67	1.08	0	0.33	0.69	0
17	Connectors expressing facts actually	0.38	0.75	0	0.07	0.29	0
18	Demonstratives	10.34	10.11	8	5.39	5.08	4
19	Discourse markers	0.82	1.36	0	0.39	0.75	0
20	Future	3.85	4.36	3	4.25	4.44	3
21	LIWC psychological dimension	34.90	32.78	29	29.16	22.19	26
22	Negations	3.19	3.50	2	2.25	2.44	2
23	Negative words	14.43	16.30	11	10.77	9.63	9
24	Positive words	13.40	13.47	11	10.13	9.84	7
25	Novelty	1.16	2.12	1	1.23	1.85	1
26	Number of pronouns	22.07	19.93	18	10.17	11.53	7
27	Pronouns	5.65	7.22	4	2.17	3	1
28	First person pronouns	6.02	8.90	4	1.97	3.98	1
29	Second person pronouns	2.26	4.21	1	0.28	1	0
30	Third person pronouns	13.79	11.92	11	7.92	8.95	5
31	FKG	10.24	2.99	10.20	11.96	2.32	11.80
32	FRE	56.53	11.38	57.40	49.14	10.66	49.75

4.2. Datasets

As a case study, we consider the ISOT Fake News Dataset,[3] previously used in research to detect spam and fake news [55,56]. This dataset contains press news between 2015 and 2017. It contains two different files:

- `True.csv` file. It contains over 21 417 real-news articles collected from the Reuter.com website.
- `Fake.csv` file. It contains over 23 481 fake news articles collected from different websites.

Each article contains the following metadata: article title, article text or body, article type (political, worldwide, among others), and publication date. The metadata used in this study is the text of the articles and their previous classification as either real news or fake news.

The ISOT Fake News Dataset is clean since it does not require extensive data preprocessing. However, there are few instances (i.e., news) without text body and others in which the body only contains video URLs or very reduced text. In some cases, the text of the news is not informative. After processing the original dataset and considering the 88 linguistic features mentioned in Section 3.1, we decided to include as an inclusion criterion that the number of words for each news must be greater than or equal to 50. This criterion allows working with texts that effectively report information. Thus, the job database contains 42 677 news, of which 21 586 are fake news (median of 381 words), and 21 091 are real news (median of 364 words). Under the described criteria, approximately 5% of the data were excluded.

5. Results and discussion

The results obtained for the data analysis are described below, in terms of the three steps described in Section 4.1: linguistic features, features extraction, and classification models.

[3] www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset/metadata

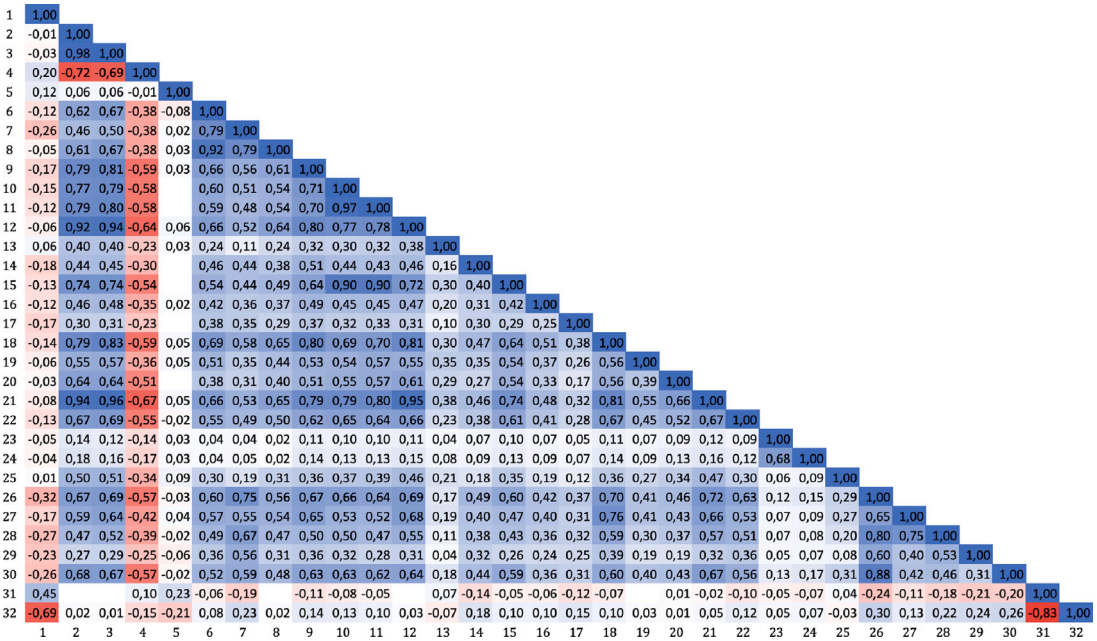


Fig. 3. Correlations between linguistic features.

5.1. Linguistic features

First, of the 88 initial linguistic features, all those for which there were missing values were excluded. Note that can be missing values for various reasons. For example, if the feature Symbol total equals 0 (see Table A.7), then the feature Symbol mean would be N/A (because of a division by zero). Because the distribution of the features is unknown (and these distributions are beyond the scope of this work), we avoid performing interpolations. On the other hand, those features having little variability (standard deviation close to zero) were discarded. For example, in Swear words, more than 98% of the values are zero, and the rest is 1.

Thus, 53 linguistic features were avoided, leaving 35 for experimentation (the ones highlighted in gray in Table A.7). After performing a non-parametric Mann–Whitney U test for each feature, with a reliability of 99%, the list was reduced to 32 linguistic features. These features are the ones that show a significant difference in their average values depending on the news is real or fake. The features Number of words per sentence, Connectors emphasis, and Nouns for research were not significant. Table 1 shows the mean, standard deviation (std), and median values for each feature and each type of news (real or fake).

Note that there are atypical data (evidenced to a certain extent in the standard deviations). However, as this is representative of a particular form of writing, they have been preserved. The treatment of atypical data is widely discussed in the literature. Our decision is consistent with that suggested by Hair et al. [57], that is, "if they represent a segment of the population, they should be retained to ensure their generality to the population as a whole".

5.2. Exploratory analysis

PCA is used to extract features as an exploratory phase of the data. As shown in Fig. 3, note that some linguistic features have a high positive or negative correlation. White cells denote not statistically significant results (i.e., p-value > 0.05).

As explained in Section 3.2, from the original variables (linguistic features), PCA allows obtaining new variables (components), uncorrelated with each other, and linear combinations of the original variables. The first components are those that accumulate the highest variance (information).

As the original variables assume values in different ranges (i.e., they have different variability), it is suggested to perform scaling [58]; in this case, a standard scaling. Subsequently, the correlation matrix is calculated, and then the eigenvalues and the variances. These values are shown in Table 2. For this case, the Kaiser criterion [59] shows that five components must be retained. However, since the eigenvalue of component 6 is 0.97, very close to 1.0, it would be appropriate to consider it as well.

Fig. 4 shows a two-dimensional projection of the variables. Each vector on the plane represents a different variable. Dimension 1 is projected on axis X, and Dimension 2 on axis Y. Both the angle and length of the vectors provide relevant information about the variables. Regarding the angles, the more parallel the vector is to some axis, the greater is its contribution to that dimension. For instance, both TTR and Negations variables contribute more to Dimension 1 than to Dimension 2. In addition, the angles between two vectors represent the correlation between both variables: small angles show a high correlation; angles close to 180°, negative correlation; and right angles or close to 90°, indicate that the correlation between both variables is close to zero. For example,

Table 2
Eigenvalue and variances for each dimension.

Dimension	Eigenvalue	Variance percentage	Cumulative variance percentage
1	14.23	44.47	44.47
2	2.92	9.12	53.58
3	1.82	5.68	59.26
4	1.47	4.58	63.85
5	1.04	3.25	67.10
6	0.97	3.03	70.13
7	0.89	2.78	72.91
8	0.87	2.71	75.62
9	0.76	2.38	78.01
10	0.74	2.31	80.32
11	0.68	2.12	82.44
12	0.66	2.07	84.51
13	0.64	1.99	86.49
14	0.57	1.77	88.27
15	0.54	1.68	89.95
16	0.52	1.61	91.56
17	0.49	1.54	93.10
18	0.41	1.28	94.38
19	0.35	1.11	95.49
20	0.32	0.99	96.47
21	0.26	0.81	97.29
22	0.19	0.60	97.89
23	0.13	0.42	98.31
24	0.12	0.37	98.68
25	0.11	0.34	99.02
26	0.10	0.32	99.34
27	0.07	0.23	99.56
28	0.06	0.19	99.75
29	0.04	0.12	99.88
30	0.03	0.09	99.96
31	0.01	0.04	100.00
32	0.00	0.00	100.00

the FKG variable is highly correlated with Letters per word, has a high inverse correlation with FRE, and a low correlation with Negations. Regarding the length of the vector, the longer the length, the better the information representation by said variable. Thus, for instance, the FKG variable gives more information than the Negative words.

Just as the vectors in Fig. 4 represent the variables, the scores in Fig. 5 represent the observations. The closest scores represent similar observations. Note that a separability of the data is observed. For a better data interpretation, the news are labeled by colors; real news are presented in blue and fake news in red. Observe that most extreme data (far from the center of the biplot) are fake news.

Table 3 shows the first six components of the rotation matrix. Based on loads of the variables in each component, it is possible to perform a high-level characterization, considering the sign difference. The first two components are discussed below.

- *Component 1*. It distinguishes between a "psychosocial-affective communication" and an "information-dense communication" (based on the sign change of the component's loads). The discourses that develop psychosocial-affective communication contain expressions of subjectivity and affectivity. They are cohesive, extensive, and varied, developing comparisons and negations, temporal-spatial references, and third persons. The features with the highest weights that characterize this component are the following: LIWC psychological dimension, Number of words, Connectives, Number of types, Demonstratives, Verb 3rd person singular present, Adverbs, Comparison, Number of pronouns, Connector contrast, Adverb total, Negations, Third person pronouns, Pronouns. On the other hand, information-dense communication develops repetitions of long words. Furthermore, the texts' style requires complete general education levels to be read (TTR, Letters per word, FKG).
- *Component 2*. It distinguishes between "informative-projective communication" and an "appellative-evaluative communication" (based on the sign change of the component's loads). The informative-projective communication is developed through a cohesive discourse, providing information characterized by the producer, with projection to the future. It also presents a vast number, variety, and extension of words (FKG, Letters per word, Adjective mean, Connectors add information, Number of types, Number of words, Connectives, and Future). Appellative-evaluative communication is developed through personal

Table 3
Rotation matrix for the different linguistic features.

Linguistic feature	Dimension					
	1	2	3	4	5	6
Letters per word	−0.05	0.42	0.11	−0.17	0.01	0.22
Number of types	0.24	0.15	−0.10	0.08	0.07	0.01
Number of words	0.25	0.14	−0.05	0.06	0.05	0.01
TTR	−0.18	−0.02	0.18	−0.13	−0.24	−0.01
Adjective mean	0.00	0.19	0.02	−0.19	0.17	−0.84
Adverb total	0.21	−0.02	0.25	−0.16	−0.31	0.04
Pronoun total	0.19	−0.20	0.27	−0.23	−0.08	0.14
Verb total	0.19	0.02	0.27	−0.16	−0.25	0.12
Verb 3rd person singular present	0.23	0.01	−0.00	0.01	−0.06	−0.07
Adverbs	0.23	0.05	−0.08	0.16	−0.04	0.05
Comparison	0.23	0.08	−0.09	0.17	−0.06	0.05
Connectives	0.24	0.12	−0.03	0.04	0.06	−0.00
Connectors add information	0.10	0.18	−0.08	0.13	−0.36	−0.03
Connectors comparison	0.15	−0.09	0.05	−0.07	−0.20	−0.17
Connectors contrast	0.21	0.07	−0.10	0.19	−0.06	0.06
Connectors explain	0.15	−0.00	0.01	−0.00	−0.14	−0.11
Connectors expressing facts actually	0.11	−0.11	0.07	−0.10	−0.29	−0.23
Demonstratives	0.23	0.03	0.04	−0.05	0.01	−0.05
Discourse markers	0.16	0.09	−0.00	0.07	−0.35	−0.10
Future	0.17	0.12	−0.11	0.12	0.18	0.12
LIWC psychological dimension	0.25	0.11	−0.04	0.05	0.08	0.01
Negations	0.20	0.00	−0.01	0.03	0.12	0.10
Negative words	0.04	−0.02	−0.51	−0.47	−0.11	0.07
Positive words	0.05	−0.01	−0.53	−0.44	−0.09	0.07
Novelty	0.12	0.15	−0.09	0.10	0.13	−0.15
Number of pronouns	0.22	−0.18	0.04	−0.10	0.23	0.03
Pronouns	0.19	−0.06	0.13	−0.16	0.17	−0.07
First person pronouns	0.18	−0.18	0.16	−0.20	0.29	0.01
Second person pronouns	0.12	−0.24	0.16	−0.23	0.20	0.08
Third person pronouns	0.20	−0.10	−0.08	0.04	0.11	0.01
FKG	−0.03	0.45	0.17	−0.22	0.10	0.05
FRE	0.05	−0.49	−0.18	0.24	−0.08	−0.13

appeal and relatively simple texts to read with the expression of opinions (FRE, Second person pronouns, Pronoun total, Pronouns, Number of pronouns, first person pronouns, Positive words, and Negative words).

5.3. Predictive analysis

We develop different news classification models. To test the different models, the work database is divided into training (80%) and testing (20%) data, that is, 34 142 news for training (17 269 fake and 16 873 real) and 8535 for testing (4317 fake and 4218 real). We consider k-fold cross-validation with $k = 10$ samples on different classification models.

We develop a first model based on explainable machine learning techniques. Specifically, we use the XGBoost classifier (with a cross-validation process repeated 'nrounds' = 400 times). The 32 features enter the model without transformation. This model achieves the best performance in our experiment, with an accuracy of 99.99%. Table 4 indicates the most important linguistic features according to the XGBoost classifier. The information 'gain' represents the importance of the feature in the model. We show all the features which obtained a gain ≥ 0.01. The 'cover' is the relative number of observations related to each characteristic, and the 'frequency' is the percentage that represents the relative number of times a feature occurs.

Second, we consider different classification models: support vector machine (SVM) with radial and linear base, decision tree, and naive Bayes. The models considered are popular in the literature for classification tasks because of their ease of implementation and excellent computational performance [60]. The input variables of the models are all PCA components (PCA is applied to standardized data). Table 5 shows the mean accuracy and standard deviation (in brackets) for each model. In the tests performed, consider that even with 10 PCA components, an accuracy of 0.92 (sd = 0.0055) is achieved using SVM (radial).

5.3.1. News classification example

As we have already mentioned, PCA generates new artificial variables from a reduction in the dimensionality of the dataset. Therefore, to interpret the results obtained, in practice, it is not enough to refer to these artificial variables, but rather it seems appropriate to use an example of specific fake news as evidence.

As an illustration, we describe a fake news well-classified through the SVM model. The news was originally published on Fox News on August 23, 2015, and then updated on December 20 of that year. Here the original news was processed before its update:

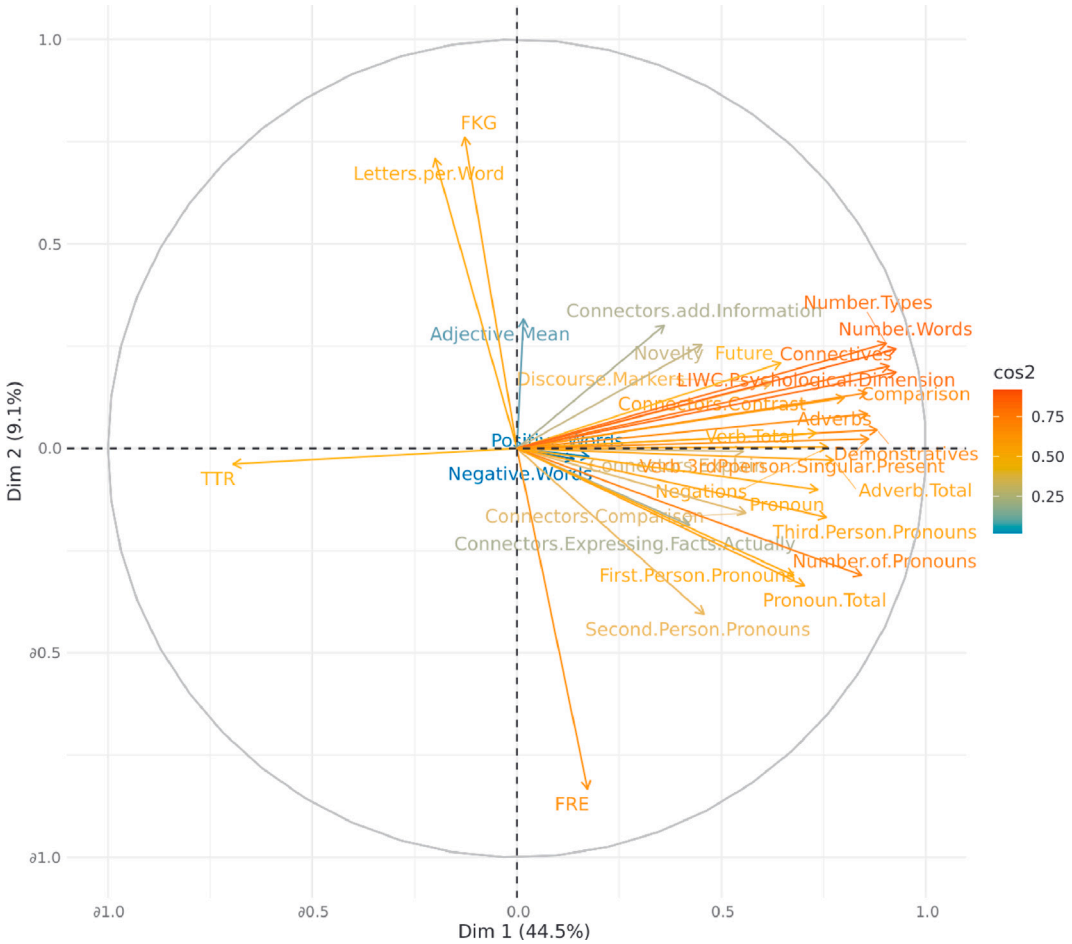


Fig. 4. Two-dimensional projection of the variables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Linguistic features with gain ≥ 0.01 for the XGBoost classifier.

	Linguistic feature	Gain	Cover	Frequency
1	Second person pronouns	0.21	0.01	0.01
2	Pronoun total	0.09	0.03	0.05
3	Adverb total	0.09	0.04	0.04
4	Adjective mean	0.09	0.06	0.06
5	Verb total	0.08	0.06	0.07
6	Number of pronouns	0.06	0.03	0.04
7	Number of words	0.06	0.06	0.04
8	Number of types	0.05	0.06	0.05
9	Verb 3rd person singular present	0.05	0.03	0.03
10	FKG	0.04	0.04	0.03
11	Pronoun	0.03	0.02	0.02
12	FRE	0.02	0.06	0.04
13	Letters per words	0.02	0.11	0.07
14	Connectors comparison	0.02	0.01	0.01
15	Positive words	0.02	0.02	0.04
16	TTR	0.01	0.12	0.06
17	Demonstratives	0.01	0.02	0.03
18	Future	0.01	0.02	0.03
19	LIWC psychological dimension	0.01	0.03	0.03
20	First person pronouns	0.01	0.01	0.02
21	Connectives	0.01	0.03	0.04

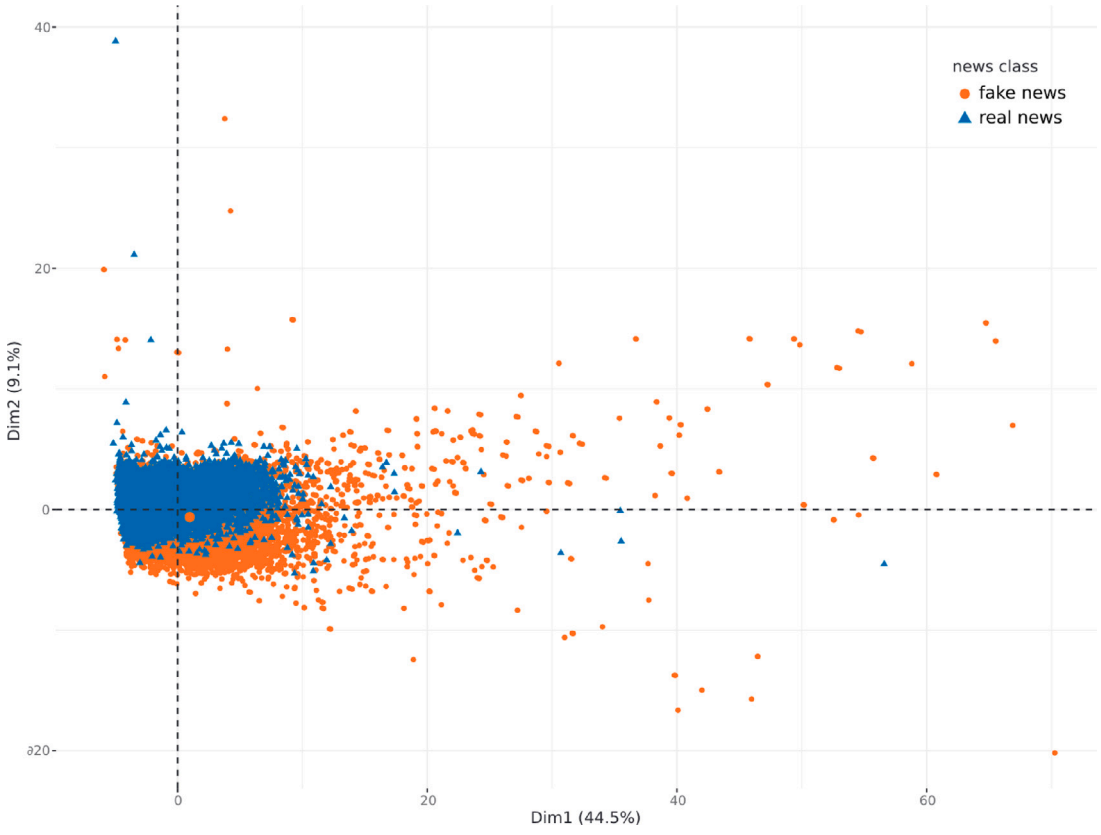


Fig. 5. Two-dimensional projection of the observations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Accuracy (and standard deviation) for different models on 10-fold cross-validation.

XGBoost	SVM(radial)	SVM(linear)	Decision tree	Naive Bayes
0.9999 (0.0002)	0.9633 (0.0044)	0.8826 (0.0390)	0.6772 (0.0142)	0.6660 (0.0091)

"Hillary may have gotten away with lying to the public for decades, but what she underestimated this time around was the enormous power of Chicago thugs, Valerie Jarrett and Barack Hussein Obama. If the two of them decide they don't want you running for office, you likely won t stand a chance. The tens of thousands of emails on Hillary Clinton's private server from when she was secretary of state could also be on a second device or server, according to news reports. The FBI now has the only confirmed private server, as part of a Justice Department probe to determine whether it sent of received classified information for Clinton when she was the country's top diplomat from 2009 to 2013. Platte River Networks, which managed Clinton's server and private email network after she left the State Department, has indicated it transfer —or migrated— emails from the original server in 2013, according to The Washington Examiner. However, Clinton, the front-running Democratic presidential candidate, has suggested that she gave the department 55,000 pages of official emails and deleted roughly 30,000 personal ones in January, which raises the possibility they were culled from a second device. Neither a Clinton spokesman nor an attorney for the Colorado-based Platte River Networks returned an Examiner's request for comment, the news-gathering agency reported Saturday. The DailyMail.com on Aug. 14 was among the first to report the possibility of a second server. The FBI took the server last week, after a U.S. Intelligence Community inspector general reportedly found two Clinton emails that included sensitive information, then asked the FBI to further investigate. Platte River Networks has told news agencies that the server, now in New Jersey, has been wiped clean. But forensics experts still might be able to recover some information. There have been reports that some of the emails that Clinton turned over included classified information. Clinton maintains that she neither sent nor received classified data, which suggests the missives might have been marked after the fact as classified or with some other top-secret classification. The emails that Clinton gave to the State Department were on multiple storage devices. A Clinton lawyer turned over at least one thumb drive that reportedly included copies of the emails that his client has already given to the federal government. Clinton has maintained that she has done nothing wrong or illegal and says she will cooperate fully with the non-criminal investigations".–Via: FOX News, 23 August 2015

Table 6
Accuracy (and standard deviation) for different models on 10-fold cross-validation. Additional experiments in the COVID-19 Fake News Dataset.

XGBoost	SVM(radial)	SVM(lineal)	Decision tree	Naive Bayes
0.8956 (0.1173)	0.8744 (0.0146)	0.7825 (0.0231)	0.7450 (0.0265)	0.6838 (0.0238)

After its correct classification, we can interpret the following. The news is recognized as fake fundamentally because it has a greater tendency towards psychosocial-affective and appellative-evaluative communication, with a high presence of positive and varied words that qualify things and objects. In addition, high temporal-spatial reference and pronouns in the second and third person are observed. For the following linguistic features, values closer to fake news than real news are observed (based on the difference between the value of the news feature and the averages according to news type): Connectors add information, Connectors explain, Demonstratives, Positive words, Second person pronouns, Third person pronouns, TTR, Adjective mean, Pronoun total, Adverb total.

5.4. Complementary experiments

We have previously reported the good performance of classifiers based on linguistic features on the ISOT Fake News Dataset. Note that the texts in that database are long, averaging 425 words. The results are promising, especially using XGBoost (See Table 5). In addition, we have considered another sample of 4000 tweets from the COVID-19 Fake News Dataset[4] to show that the models also work with short texts. This database contains tweets of 27 average words tagged as true or fake. We used the 32 linguistic features from the previous experiment and considered cross-validation under the same conditions. The results are shown in Table 6.

Although the Twitter data has a limitation in the text length, which could affect the information reported by the linguistic features, this type of data could be complemented with those provided by the social network structure [61]. Both inputs can help improve the classification.

6. Conclusions and future work

Although linguistic features have been used successfully in models for classifying news into real and fake, there are currently no clear criteria for which features are more suitable for each context. A justified selection of linguistic features can help conduct the analysis and for the interpretability of the results.

In this paper, we propose a comprehensive list of 88 linguistic features, which can be computed and organized into well-defined categories (surface information, part of speech, discursive characteristics, and readability indices), which allow extracting linguistic features from the news text. The idea is to overcome the main issues described in Section 2.2. Organizing the linguistic features into categories allows to speed up the statistical tests necessary to evaluate if those features show significant differences according to the type of news (classification levels). This selection of linguistic features has implications for the parsimony of the model (selection of variables), efficient use of computational resources, and, most importantly, obtaining more consistent interpretations.

For the experimentation, we use the database considered in [55,56]. In Section 5.2, we performed a descriptive analysis (mean, median, standard deviation) of the selected linguistic features and reported significant correlations. PCA allowed us to identify and interpret the first two components, providing several advantages in this exploratory stage. At the graphic level, it allowed seeing the separability of the data according to the type of news. It also allowed identifying the weight of the features in the components. Finally, PCA allowed a high-level linguistic interpretation. In particular, the first component distinguishes between "psychosocial-affective communication" (with the highest weights in LIWC psychological dimension and Number of words) and "information-dense communication" (with the highest weights in TTR and Letters per word). The second component distinguishes between an "informative-projective communication" (FKG, Letters per word) and an "appellative-evaluative communication" (FRE and Second person pronouns). This analysis highlights the importance of an a priori classification of linguistic features, allowing an ad-hoc interpretation in the functional use of linguistics.

By observing the performance of different models using the 32 selected linguistic features, we show that the XGBoost classifier and SVM achieve an accuracy of 99.99% and 96.33%, respectively. Also, the XGBoost classifier allowed us to identify the most important linguistic features. The highest 'gain' scores are for Pronoun total, Adverb total, and Adjective mean. Note that our models exceed the results reported so far (92% accuracy in the best case with SVM). The most recent studies using this database are [62,63]. In [62] the best results are obtained with statistical features for SVM (78.1% accuracy) and with contextual features for decision tree (96.2% accuracy). On the other hand, using BERT, they achieve 96.9% accuracy, and with Stacked GRU (Glove NLP Technique), 99.1% accuracy. Finally, in [63], the authors achieve an accuracy of 99.96% using BERT and Funnel Transformer under deep learning methods.

Our proposal is simple to implement and interpret. Our results far outperform most previous results (see Section 2.2) and are equal to the highest reported precisions. Our contribution also includes listing well-defined linguistic features in categories that help to overcome current problems. In addition, we provide detailed interpretations of the linguistic features we use. These results

[4] https://paperswithcode.com/dataset/covid-19-fake-news-dataset

Table A.7
All linguistic features initially considered in this work. Features initially selected for experiments are highlighted in gray. The three features marked with * do not provide a significant difference between real and fake news, and thus were also avoided for the experiments.

Group	Linguistic feature	Description
Surface information	Letters per word	Mean number of letters in the words of a text
	Number of paragraphs	Num. of paragraphs build by *n* sentences
	Number of sentences	Num. of sentences build by *n* words
	Number of types	Num. of different words in a text
	Number of words	Num. of words (or tokens) in a text
	*Number of words per sentence	Mean number of words in the sentences of a text
	TTR	Type Token Ratio: relationship between the number of types and number of tokens. A high TTR means a large amount of lexical variation. A low TTR means little lexical variation.
Part of speech (POS)	Adjective mean	Mean number of adjectives in a text
	Adjective total	Num. of adjectives in a text
	Adposition mean	Mean number of prepositions and postpositions in a text
	Adposition total	Num. of prepositions and postpositions in a text
	Adverb mean	Mean number of adverbs in a text
	Adverb total	Num. of adverbs in a text
	Coordinating conjunction mean	Mean number of conjunctions in a text
	Coordinating conjunction total	Num. of conjunctions in a text
	Determiner mean	Mean number of determiners in a text
	Determiner total	Num. of determiners in a text
	Noun mean	Mean number of nouns in a text
	Noun total	Num. of nouns in a text
	Numeral mean	Mean number of words used to represent numbers in a text
	Numeral total	Num. of words used to represent numbers in a text
	Pronoun mean	Mean number of pronouns in a text
	Pronoun total	Num. of pronouns in a text
	Proper noun mean	Mean number of proper nouns (i.e., names) in a text
	Proper noun total	Num. of proper nouns (i.e., names) in a text
	Punctuation mean	Mean number of punctuation marks in a text
	Punctuation total	Num. of punctuation marks in a text
	Symbol mean	Mean number of symbols in a text
	Symbol total	Num. of symbols in a text
	Verb mean	Mean number of verbs in a text
	Verb total	Num. of verbs in a text
	Verb 3rd person singular present	Num. of verbs in a text used in present time, with singular nouns, and with the pronouns he/she/it/one
	Verb base form	Num. of verbs in infinitive without the preposition "to" in a text
	Verb gerund or present participle	Num. of verbs that end in -ing
	Verb modal	Num. of modal verbs (e.g., can, could, shall, should, ought to, will, or would) in a text
	Verb past participle	Num. of verbs in past participle (usually ending in -ed) in a text
	Verb past tense	Num. of verbs in past tense in a text
Discursive characteristics	Adverbs	Frequency of common adverbs in a text (however, still, yet, nevertheless, but, although, despite, therefore, additionally, consequently, etc.)
	Articles	Frequency of common articles in a text
	Background	Frequency of words or phrases in a text to emphasize a research background
	Comparison	Frequency of words or phrases in a text used to contrast differences or similarities of certain information
	Conclusion	Frequency of words or phrases in a text to conclude ideas in research
	Connectives	Frequency of connectors in a text (words or phrases to link statements or linguistic units together)
	Connectors add information	Frequency of connectors in a text to adjoin new information
	Connectors comparison	Frequency of connectors to compare in terms of similarities
	Connectors contrast	Frequency of connectors to compare in terms of differences
	*Connectors emphasis	Frequency of connectors in a text to emphasize something
	Connectors explain	Frequency of connectors in a text to explain ideas
	Connectors expressing facts actually	Frequency of connectors to give facts as objective reality
	Connectors expressing opinion	Frequency of connectors to express judgments
	Connectors reason & cause	Frequency of connectors in a text to give a reason for an action or condition
	Connectors time & sequence	Frequency of connectors in a text to organize actions in time or following a sequence
	Demonstratives	Frequency of demonstratives in a text (this, those, etc.)
	Determiners	Frequency of common determiners in a text (a, the, every, etc.)
	Discourse markers	Frequency of discourse markers in a text (oh, well, now, then, you know, I mean)
	Future	Frequency of grammatical categories to express future tense

(*continued on next page*)

Table A.7 (*continued*).

	Importance	Frequency of words or phrases in a text to emphasize the importance of a claim or statement in research
	Key connectors	Frequency of connectors usually used in empirical research
	LIWC linguistic dimension	Frequency of words in a text included in the linguistic dimension of LIWC [41]
	LIWC psychological dimension	Frequency of words in a text included in psychological dimension of LIWC [41]
	Mechanism	Frequency of words or phrases in a text to describe mechanisms in a research process
	Negations	Frequency of words in a text expressing negative statements
	Negative words	Frequency of words in a text to express negative feelings
	Positive words	Frequency of words in a text to express positive feelings
	*Nouns for research	Frequency of nouns usually used in empirical research
	Novelty	Frequency of words or phrases in a text to emphasize originality of a research
	Objective	Frequency of words or phrases in a text to introduce the aim of a research
	Perspectives	Frequency of words or phrases in a text to introduce opportunities or consequences of findings
	Prepositions	Frequency of common prepositions in a text
	Problem	Frequency of words or phrases in a text to identify a research problem
	Number of pronouns	Frequency of common pronouns in a text (he, she, it, they, someone, etc.)
	Pronouns	Frequency of common pronouns in an academic text (this, these, we, our, etc.)
	First person pronouns	Frequency of 1st person pronouns (I, me, mine, myself, etc.)
	Second person pronouns	Frequency of 2nd person pronouns (you, your, yours, etc.)
	Third person pronouns	Frequency of 3rd person pronouns (he, she, it, they, etc.)
	Results	Frequency of words or phrases in a text to introduce findings
	State of the art	Frequency of words or phrases in a text to introduce antecedents of research about a topic
	Summary	Frequency of words or phrases in a text to summarize information
	Swear words	Frequency of profane or obscene oaths or words
Readability indices	ARI	Automated readability index. A value 10 means a high school student level (15-16 years old); a value 3 means students in 3rd grade (8-9 years old).
	CLI	Coleman–Liau index. It calculates samples of hard-copy text, instead of manually hard-coding the text. Unlike syllable-based readability indicators, it does not require to analyze the characters that create the words (such as syllable counts), but only their length in characters.
	DCRS	Dale-Chall readability formula. It is based on sentence length and the number of 'hard' words (i.e., words that do not appear on a specially designed list of common words familiar to most 4th-grade students). A value ≤ 4.9 means grade ≤ 4, while a value ≥ 10 means grades ≥ 16 (college graduate). The New Dale-Chall formula is based on familiar words, rather than syllable or letter counts.
	DW	Dale-Chall Word List. It contains approximately 3000 familiar words for at least 80% of the children in grade 5. It gives a significant correlation with reading difficulty.
	FKG	Flesch Kincaid Grade Level index. It is based on FRE. A level 8 means the reader needs a grade level ≥ 8 to understand (13-14 years old). Even for advanced readers, it means the content is less time-consuming to read.
	FRE	Flesch Reading Ease index. It gives a score between 1-100, with 100 being the highest readability score. Scoring between 70-80 is equivalent to school grade level 8. (13-14 years old). Sentences that contain a lot of words are more difficult to follow than shorter sentences.
	GFOG	Gunning's Fog Index. It considers that short sentences written in Plain English achieve a better score than long sentences written in complicated language. The ideal score is 7-8. A value ≥ 12 is too hard for most people to read.
	LWF	Linsear Write formula. It was developed for the United States Air Force to help them calculate the readability of their technical manuals.
	SMOG	Simple Measure of Gobbledygook. It is a popular method for health literacy materials.

provide additional evidence of the linguistic features' explanatory power, providing a structured way to select and analyze linguistic features in the context of fake news detection.

Focusing on interpretation, Section 5.3 illustrates an example of well-classified fake news, characterized as one with a greater tendency towards psychosocial-affective and appellative-evaluative communication, with a high presence of positive and varied words that qualify things and objects. In addition, high temporal-spatial reference and pronouns in the second and third person are observed.

As future work, it is proposed to define a model based on linguistic features to detect fake news in Spanish. In addition, progress should be made towards sensitivity tests, that is, testing the model in different contexts.

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.

Acknowledgments

E. Puraivan has been partially funded by the Escuela de Ingeniería Informática, Universidad de Valparaíso, Chile, through grant No. 101.016/2020. F. Riquelme has been partially supported by Fondecyt de Iniciación 11200113 from ANID, Chile. R. Venegas has been partially supported by Fondecyt 1190639 from ANID and the Applied Natural Language Processing Research Nucleus of the Pontificia Universidad Católica de Valparaíso, Chile.

Appendix. Linguistic features

See Table A.7.

References

[1] R. Vila de Prado, The post truth and the spyral of silence, Revista Aportes de la Comunicación y la Cultura (24) (2018) 9–19.
[2] B. Collins, D.T. Hoang, N.T. Nguyen, D. Hwang, Fake news types and detection models on social media a state-of-the-art survey, in: Communications in Computer and Information Science, Springer Singapore, 2020, pp. 562–573, http://dx.doi.org/10.1007/978-981-15-3380-8_49.
[3] G. Pennycook, T.D. Cannon, D.G. Rand, Prior exposure increases perceived accuracy of fake news., J. Exp. Psychol. [Gen.] 147 (12) (2018) 1865–1880, http://dx.doi.org/10.1037/xge0000465.
[4] C. Martel, G. Pennycook, D.G. Rand, Reliance on emotion promotes belief in fake news, Cogn. Res Princ. Implic. 5 (1) (2020) http://dx.doi.org/10.1186/s41235-020-00252-3.
[5] A. Campan, A. Cuzzocrea, T.M. Truta, Fighting fake news spread in online social networks: Actual trends and future research directions, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 4453–4457, http://dx.doi.org/10.1109/bigdata.2017.8258484.
[6] A. Gelfert, Fake news: A definition, Informal Log. 38 (1) (2018) 84–117, http://dx.doi.org/10.22329/il.v38i1.5068.
[7] D. Spohr, Fake news and ideological polarization, Bus. In. Rev. 34 (3) (2017) 150–160, http://dx.doi.org/10.1177/0266382117722446.
[8] A. Giachanou, P. Rosso, The battle against online harmful information, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, 2020, pp. 3503–3504, http://dx.doi.org/10.1145/3340531.3412169.
[9] C. Chen, K. Wu, V. Srinivasan, X. Zhang, Battling the internet water army, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, 2013, pp. 116–120, http://dx.doi.org/10.1145/2492517.2492637.
[10] J. Corner, Fake news, post-truth and media–political change, Media Cult. Soc. 39 (7) (2017) 1100–1107, http://dx.doi.org/10.1177/0163443717726743.
[11] S. Lewandowsky, U.K. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the ‘‘post-truth’’ era, J. Appl. Res. Memory Cogn. 6 (4) (2017) 353–369, http://dx.doi.org/10.1016/j.jarmac.2017.07.008.
[12] N. Corbu, E. Negrea-Busuioc, Populism meets fake news: Social media, stereotypes and emotions, in: Perspectives on Populism and the Media, Nomos Verlagsgesellschaft mbH & Co. KG, 2020, pp. 181–200, http://dx.doi.org/10.5771/9783845297392-181.
[13] S. van der Linden, J. Roozenbeek, J. Compton, Inoculating against fake news about covid-19, Front. Psy. 11 (2020) http://dx.doi.org/10.3389/fpsyg.2020.566790.
[14] S. Kogan, T.J. Moskowitz, M. Niessner, Fake news: Evidence from financial markets, SSRN Electr. J. (2018) http://dx.doi.org/10.2139/ssrn.3237763.
[15] O.D. Apuke, B. Omar, Fake news and COVID-19: modelling the predictors of fake news sharing among social media users, Telemat. Inform. 56 (2021) 101475, http://dx.doi.org/10.1016/j.tele.2020.101475.
[16] G. Pennycook, J. McPhetres, Y. Zhang, J.G. Lu, D.G. Rand, Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention, Psychol. Sci. 31 (7) (2020) 770–780, http://dx.doi.org/10.1177/0956797620939054.
[17] T. Adhanom, Munich security conference. World health organization, 2020, https://www.who.int/director-general/speeches/detail/munich-security-conference. (Online Accessed June 2021).
[18] S.M. Ghafari, A. Beheshti, A. Joshi, C. Paris, A. Mahmood, S. Yakhchi, M.A. Orgun, A survey on trust prediction in online social networks, IEEE Access 8 (2020) 144292–144309, http://dx.doi.org/10.1109/access.2020.3009445.
[19] X. Zhang, A.A. Ghorbani, An overview of online fake news: Characterization, detection, and discussion, Inf. Process. Manage. 57 (2) (2020) 102025, http://dx.doi.org/10.1016/j.ipm.2019.03.004.
[20] S.R.K. Indarapu, J. Komalla, D.R. Inugala, G.R. Kota, A. Sanam, Comparative analysis of machine learning algorithms to detect fake news, in: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), IEEE, 2021, pp. 591–594, http://dx.doi.org/10.1109/icspc51351.2021.9451690.
[21] S. Kumar, B. Arora, A review of fake news detection using machine learning techniques, in: 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 2021, pp. 1–8, http://dx.doi.org/10.1109/ICESC51422.2021.9532796.
[22] R. Varma, Y. Verma, P. Vijayvargiya, P.P. Churi, A systematic survey on deep learning and machine learning approaches of fake news detection in the pre- and post-COVID-19 pandemic, Int. J. Intell. Comput. Cybern. 14 (4) (2021) 617–646, http://dx.doi.org/10.1108/IJICC-04-2021-0069.
[23] I.H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Computer Science 2 (3) (2021) http://dx.doi.org/10.1007/s42979-021-00592-x.
[24] A. Arango, J. Pérez, B. Poblete, Hate speech detection is not as easy as you may think: A closer look at model validation (extended version), Inf. Syst. (2020) 101584, http://dx.doi.org/10.1016/j.is.2020.101584.
[25] M. Mahyoob, J. Algaraady, M. Alrahaili, Linguistic-based detection of fake news in social media, Int. J. English Linguist. 11 (1) (2020) 99, http://dx.doi.org/10.5539/ijel.v11n1p99.
[26] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 3391–3401, URL https://aclanthology.org/C18-1287.
[27] N. Hassan, W. Gomaa, G. Khoriba, M. Haggag, Credibility detection in Twitter using word N-gram analysis and supervised machine learning techniques, Int. J. Intell. Eng. Syst. 13 (1) (2020) 291–300, http://dx.doi.org/10.22266/ijies2020.0229.27.
[28] A. Giachanou, B. Ghanem, E.A. Ríssola, P. Rosso, F. Crestani, D. Oberski, The impact of psycholinguistic patterns in discriminating between fake news spreaders and fact checkers, Data Knowl. Eng. 138 (2022) 101960, http://dx.doi.org/10.1016/j.datak.2021.101960.
[29] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, Y. Kompatsiaris, Detection and visualization of misleading content on Twitter, Int. J. Multimedia Inf. Retriev. 7 (1) (2017) 71–86, http://dx.doi.org/10.1007/s13735-017-0143-x.
[30] O. Ajao, D. Bhowmik, S. Zargari, Fake news identification on Twitter with hybrid CNN and RNN models, in: Proceedings of the 9th International Conference on Social Media and Society, ACM, 2018, pp. 226–230, http://dx.doi.org/10.1145/3217804.3217917.
[31] T. Hamdi, H. Slimi, I. Bounhas, Y. Slimani, A hybrid approach for fake news detection in Twitter based on user features and graph embedding, in: Distributed Computing and Internet Technology, Springer International Publishing, 2019, pp. 266–280, http://dx.doi.org/10.1007/978-3-030-36987-3_17.

[32] S. Kumar, B. Huang, R.A.V. Cox, K.M. Carley, An anatomical comparison of fake-news and trusted-news sharing pattern on Twitter, Comput. Math. Organ. Theory (2020) http://dx.doi.org/10.1007/s10588-019-09305-5.
[33] S. Kumar, R. Asthana, S. Upadhyay, N. Upreti, M. Akbar, Fake news detection using deep learning models: A novel approach, Trans. Emerg. Telecommun. Technol. 31 (2) (2019) http://dx.doi.org/10.1002/ett.3767.
[34] N. Seddari, A. Derhab, M. Belaoued, W. Halboob, J. Al-Muhtadi, A. Bouras, A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media, IEEE Access 10 (2022) 62097–62109, http://dx.doi.org/10.1109/access.2022.3181184.
[35] D.K. Sharma, P. Shrivastava, S. Garg, Utilizing word embedding and linguistic features for fake news detection, in: 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), 2022, pp. 844–848, http://dx.doi.org/10.23919/INDIACom54597.2022.9763294.
[36] A. Choudhary, A. Arora, Linguistic feature based learning model for fake news detection and classification, Expert Syst. Appl. 169 (2021) 114171, http://dx.doi.org/10.1016/j.eswa.2020.114171.
[37] M.N. Nikiforos, S. Vergis, A. Stylidou, N. Augoustis, K.L. Kermanidis, M. Maragoudakis, Fake news detection regarding the Hong Kong events from tweets, in: Artificial Intelligence Applications and Innovations. AIAI 2020 IFIP WG 12.5 International Workshops, Springer International Publishing, 2020, pp. 177–186, http://dx.doi.org/10.1007/978-3-030-49190-1_16.
[38] R. Agarwal, S. Gupta, N. Chatterjee, Profiling fake news spreaders on Twitter: A clickbait and linguistic feature based scheme, in: Natural Language Processing and Information Systems, Springer International Publishing, 2022, pp. 345–357, http://dx.doi.org/10.1007/978-3-031-08473-7_32.
[39] R. Sousa-Silva, Fighting the fake: A forensic linguistic analysis to fake news detection, Int. J. Semiotics of Law - Revue internationale de Sémiotique juridique (2022) http://dx.doi.org/10.1007/s11196-022-09901-w.
[40] S. Crossley, K. Kyle, L. Allen, L. Guo, D. McNamara, Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation, J. Writing Assess. 7 (1) (2014) 10–16.
[41] J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn, The Development and Psychometric Properties of LIWC2015, Technical Report, University of Texas at Austin, 2015, pp. 1–25, http://dx.doi.org/10.15781/T29G6Z.
[42] J.N. Farr, J.J. Jenkins, D.G. Paterson, Simplification of flesch reading ease formula., J. Appl. Psychol. 35 (5) (1951) 333–337, http://dx.doi.org/10.1037/h0062427.
[43] J.P. Kincaid, R.P. Fishburne, R.L. Rogers, B.S. Chissom, Derivation of new readability formulas (automated readability index, FOG count, and Flesch Reading Ease formula) for Navy enlisted personnel, Technical Report, (8–75) Chief of Naval Technical Training: Naval Air Station Memphis, 1975.
[44] G.R. Naik (Ed.), Advances in principal component analysis, Springer Singapore, 2018, http://dx.doi.org/10.1007/978-981-10-6704-4.
[45] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 6, Pearson, Upper Saddle River, NJ, 2007.
[46] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdiscip. Rev. Comput. Stat. 2 (4) (2010) 433–459, http://dx.doi.org/10.1002/wics.101.
[47] F. Nwanganga, M. Chapple, Practical Machine Learning in R, Wiley, 2020, http://dx.doi.org/10.1002/9781119591542.
[48] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA, IEEE, 2018, pp. 80–89, http://dx.doi.org/10.1109/dsaa.2018.00018.
[49] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, IEEE Trans. Radiat. Plasma Med. Sci. (2021) 1, http://dx.doi.org/10.1109/trpms.2021.3066428.
[50] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable, Lulu, Morisville, North Carolina, 2019.
[51] R.J. Wieringa, Design Science Methodology for Information Systems and Software Engineering, Springer Berlin Heidelberg, 2014, http://dx.doi.org/10.1007/978-3-662-43839-8.
[52] E. Puraivan, E. Godoy, F. Riquelme, R. Salas, Fake news detection on Twitter using a data mining framework based on explainable machine learning techniques, in: 11th International Conference of Pattern Recognition Systems, ICPRS 2021, Institution of Engineering and Technology, 2021, pp. 157–162, http://dx.doi.org/10.1049/icp.2021.1450.
[53] R. Core Team, R: A language and environment for statistical computing, 2020, URL https://www.R-project.org/.
[54] A. Kassambara, F. Mundt, Factoextra: Extract and visualize the results of multivariate data analyses, 2020, URL https://CRAN.R-project.org/package=factoextra, R package version 1.0.7.
[55] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Secur. Privacy 1 (1) (2017) e9, http://dx.doi.org/10.1002/spy2.9.
[56] H. Ahmed, I. Traore, S. Saad, Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques, in: Lecture Notes in Computer Science, Springer International Publishing, 2017, pp. 127–138, http://dx.doi.org/10.1007/978-3-319-69155-8_9.
[57] J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, et al., Análisis multivariante, vol. 491, Prentice Hall Madrid, 1999.
[58] J. Aldas Manzano, E. Uriel Jimenez, AnáLisis Multivariante Aplicado Con R, Ediciones Paraninfo, SA, 2017.
[59] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1) (1960) 141–151.
[60] S. Umadevi, K.S.J. Marseline, A survey on data mining classification algorithms, in: 2017 International Conference on Signal Processing and Communication, ICSPC, IEEE, 2017, pp. 264–268, http://dx.doi.org/10.1109/cspc.2017.8305851.
[61] E. Providel, D. Toro, F. Riquelme, M. Mendoza, E. Puraivan, CLNews: The first dataset of the Chilean social outbreak for disinformation analysis, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 4394–4398, http://dx.doi.org/10.1145/3511808.3557560.
[62] E. Amer, K.-S. Kwak, S. El-Sappagh, Context-based fake news detection model relying on deep learning models, Electronics 11 (8) (2022) http://dx.doi.org/10.3390/electronics11081255.
[63] M. Samadi, M. Mousavian, S. Momtazi, Deep contextualized text representation and learning for fake news detection, Inf. Process. Manage. 58 (6) (2021) 102723, http://dx.doi.org/10.1016/j.ipm.2021.102723.

Eduardo Puraivan received a bachelor's degree in engineering sciences and a master's degree in statistics. He is currently a student in the doctoral program in Ingeniería Informática Aplicada at Universidad de Valparaíso, Chile. He is lecturer and research at Universidad Viña del Mar, Chile.

René Venegas is Spanish teacher and graduated in Hispanic Language and Literature at the Universidad de Playa Ancha, Chile(1994). His master studies are in National Policies Management: Education and Culture at the before mentioned university. In year 2005, he obtained his doctor degree at the Pontificia Universidad Católica de Valparaíso, Chile. His research interests are academic discourse, the study of meaning with computer tools, and natural language processing. He has participated in several research projects as principal and co-principal, with special interest in the description of academic, professional and political discourse using a corpus linguistics and natural language processing approach. He is currently professor at Potificia Universidad Católica de Valparaíso, Chile.

Fabián Riquelme received the M.Sc. degree at Universidad de Concepción, Chile, and the PhD. in Computing from the Universitat Politécnica de Catalunya, Spain. Interested in social network analysis, social computing, decision systems and data analysis. He is currently the Research and Innovation Coordinator of the Faculty of Engineering of Universidad de Valparaíso, Chile.