Bribones_tras_la_esmeralda_perdida@FakeDeS 2021: Fake news detection based on random forests, k-nearest neighbors, and n-grams for a Spanish corpora

Victor Lomas Barrie[1], Nora Perez[1], Victor Manuel Lara[2], and Antonio Neme[3]

[1] IIMAS, Universidad Nacional Autónoma de México, México
[2] Facultad de Matemáticas, Universidad Autónoma de Yucatán, Mérida, Yucatán, México
[3] Unidad Académica Mérida, IIMAS, Universidad Nacional Autónoma de México, Mérida, Mexico
antonio.neme@iimas.unam.mx

Abstract. Fake news constitutes a social problem that affects democratic societies by influencing and manipulating public opinion. The concept of fake news covers a wide range of news that somehow resembles the truth. For example, they might contain explicit false assertions, contain some aspects of the truth, or even present the truth in a deformed fashion to disqualify a political figure or a specific group. The obnoxious impact of fake news makes it imperative to develop tools to help professional journalists, on the one hand, and the public, on the other, to detect news that is fake. In this contribution, we describe a methodology with the aim to tell apart fake from true news in a data set of several hundreds of manually curated journalistic pieces. The proposal is based on a bag-of-words approach and relies on shallow classifiers. We intended to validate the hypothesis that true and fake news can be told apart with simple assumptions. However, as discussed, the hypothesis was rejected, as it was not possible to classify the texts in a subsequent test stage successfully. The corpora used for the classification task is in Spanish, and it was presented as an open challenge in CodaLab.

Keywords: Fake news · Machine learning · bag-of-words

3

1 Introduction

The University of Michigan Library defines fake news *as those news stories that are false: the story itself is fabricated, with no verifiable facts, sources or quotes*

[3] *IberLEF 2021, September 2021, Málaga, Spain.*

[4]. Fake news are false, partially true, or modified versions of the true, with the aim of give the public a deformed view of a certain process, public actor, or social phenomena. Fake news are a major source of confusion in every kind of society, and thus, enormous efforts, mainly from public organizations, are in motion to try to expose not only fake news, but also the players behind them. The latter tends to be a harder problem, and is not further discussed.

The spread of fake news tends to be faster than the corresponding to true news [5]. This difference in diffusion rate amplifies the damages caused by misleading information, since several actions have to be taken in order to counteract false information. The Collins English dictionary declared *Fake news* as the word of the year in 2017. It is an indication of the growing exposure of the public to such kind of news.

In Mexico, the problem of fake news is an increasing malady. It has been documented in several studies [7, 1]. The vast majority of fake news tend to be of political content, towards some of the major public personalities. However, since 2020 and the arrival of the COVID19 pandemics, the presence of fake news in the Mexican media, both online and in traditional TV and radio shows has been a major problem. An obvious problem is the diffusion of false news that has a negative impact in public health.

The human factor in the detection of fake news is so far, the best alternative [8]. The identification of suspicious news is a complex task. In order to detect a fake news, the reader should follow a series of steps. First, she/he should determine the subject of the piece, then, the second step, try to put in context of what it is already known, and, in a third step, corroborate that information in other trustworthy sites or media. Alternatively, he/she could confirm the information with an expert.

Big mass media is not exempt of fake news diffusion [6]. In fact, several major newspapers and broadcasting companies tend to be embedded, knowingly or unknowingly, in the diffusion of fake news.

An automatic tool that could be of help to the public in the identification of a suspicious news can be obtained via a classifier. The problem of automatic classification involves two main components. The first one is a dataset, usually consisting of a large enough sample containing a relevant number of instances belonging to one of the relevant classes or labels. The second component is an algorithm that, based on certain characterizations of instances in the dataset, is able to properly tell apart the relevant classes. In the problem at hand, the recognition of a journalistic piece as true or false (fake), the classification is a dichotomy, that is, it belongs to either one class or the other.

In this contribution, we describe our efforts to classify journalist pieces as either true or false. The corpora (dataset) consists of almost 700 of pieces manually curated as true or fake. The dataset was generated by researchers working in Mexican Public institutions, and a public challenge was started to obtain an algorithm able to cope with the task at hand. The rest of the contribution goes as follows.

In this job, we departed from the hypothesis that true and fake news can be adequately tell apart from each other based on the frequency of use of certain words. We describe our efforts in those line as follows. In section 2 we briefly describe related work, then, in section 3 we describe our proposal. Finally, in sections 4 and 5 we describe the results of the proposed methodology in the above mentioned dataset, and offer some conclusions and discussion.

2 Related work

In [9], it is described a methodology to characterize a given text in terms of n-grams. The main idea is to train an algorithm to find differences in use of n-grams among true and fake pieces. Also, authors provide a tool helps the reader to characterize the source of the site as reliable or not, by checking the reliability of the publishing site. Authors use as classifiers random forests and naive Bayes classifiers.

In [10], a deep learning architecture was trained from a corpora of several hundred papers made available from a public challenge (Kaggle). The relevant idea was to correlate the title of the journalistic piece and the main body. When an inconsistency between the title and the body, processed by word2vec (more on this in the next section), was found, the text was a candidate fake news.

In [8], several classifiers were trained to tell apart true and fake news, based on several features extracted from the texts. Some linguistic features were used as input to the classifiers. In particular,the percentage of words linked to positive or negative emotions, the percentage of stop words, and the use of informal language, were of particular relevance.

In [11], authors apply random forests, boosting and linear regression algorithms to a corpora of news in Spanish. Each text is characterized following a bag-of-words, an n-gram approach and a POS tags n-gram representation. The results are relevant since it is one of the fist experiences in texts in Spanish, and in Mexican media in particular.

3 The algorithm

Our approach is based on describing each text as a point in a high-dimensional space. That space is the relative frequency of use of each of the words in the whole vocabulary. The vocabulary was obtained from all texts. The Bag-of-Words (BoW) approach treats language as structure-less objects, in which the relevant features are the relative frequency of use or appearance of words (tokens). This approach leaves grammar aside. The term BoW was coined more than fifty years ago, in a diminishing way to bold the relevance of structure in language [12]. Despite heavy criticism, BoW has offered interesting results in several contexts, such as in spam detection [13] and in authorship attribution [14].

Aligned with the BoW approach, the perspective offered by word2vec is a rather powerful one [15], based on coding of words as weights in neural networks. The tools under this umbrella allow a link from syntaxis to semantics. However, this technique requires large datasets, and more importantly, might hide the changes in writing styles. Since we are interested in detecting changes in the general way an author writes after a given date, we postulate that relative frequency is enough. We will give evidence of this postulate while the paper unveils.

Different alternatives exist to BoW, such as that presented in [16]. There, a deep structure between words is obtained, and from it, inferences about the possible authorship are made. Neural networks with deep architectures offer also a possibility, at the expense of making inferences about the relevant attributes a task on itself [17].

First, all texts from the same columnist j are scanned to create his/her vocabulary, V_j. Then, in a second scan, each text t is mapped into a new space. This space is generated from the vocabulary, and it has $|V_j|$ dimensions. The coordinates of text t are linked to the relative frequency (probability) of use of each of the word in the vocabulary. Thus, the text t is a point in the space of relative frequencies. The $N-$gram approach, which is what we just described, although simple, allows the identification of relevant patterns. We define γ_t as the location of text j in the vocabulary space V_j.

Each text t has associated a label, indicating whether it is true or fake. This label is referred to as L_j. The values for L_j can be either 0, if the text is true, or 1, if it is fake, all accordingly to the manual classification. Now we have defined all the required elements to frame the problem we are attacking within classification theory. The attributes are the relative frequency of use of each word within the vocabulary, and the label is whether the text was true or fake. What we are interested in solving is thus the classification task $L_j = \gamma_j$.

The challenge consisted of two stages: 1. Training / calibration; and 2. Evaluation. In the first stage, 676 news from the Mexican media were made public, 338 of which were true, and 338 were false. The sample was balanced, which facilitates, theoretically, the classification task. In the second stage, the label was not unveiled, and the system was evaluated online, at maximum two times in total. We give a panoramic view of the algorithm in fig. 1

A classifier is an algorithm that, given a set of vectors or observations, described by several attributes, tries to link them to elements in a second set, namely the labels or classes. In other words, a classifier provides a function between elements in a set, with elements in a second one. This function can be either implicit, such as in the case of neural networks, or explicit, as in linear regression. The former tend to show very low errors (when provided by enough data both in quantity and quality), but tend to be only poorly interpretable. The latter are very interpretable, but tend to have very high errors, since almost no relevant phenomena is well described by a linear association.

Two classifiers were applied in our proposal. The first is random forests (RF), and the second one is k nearest neighbors (kNN). RF offers results that are

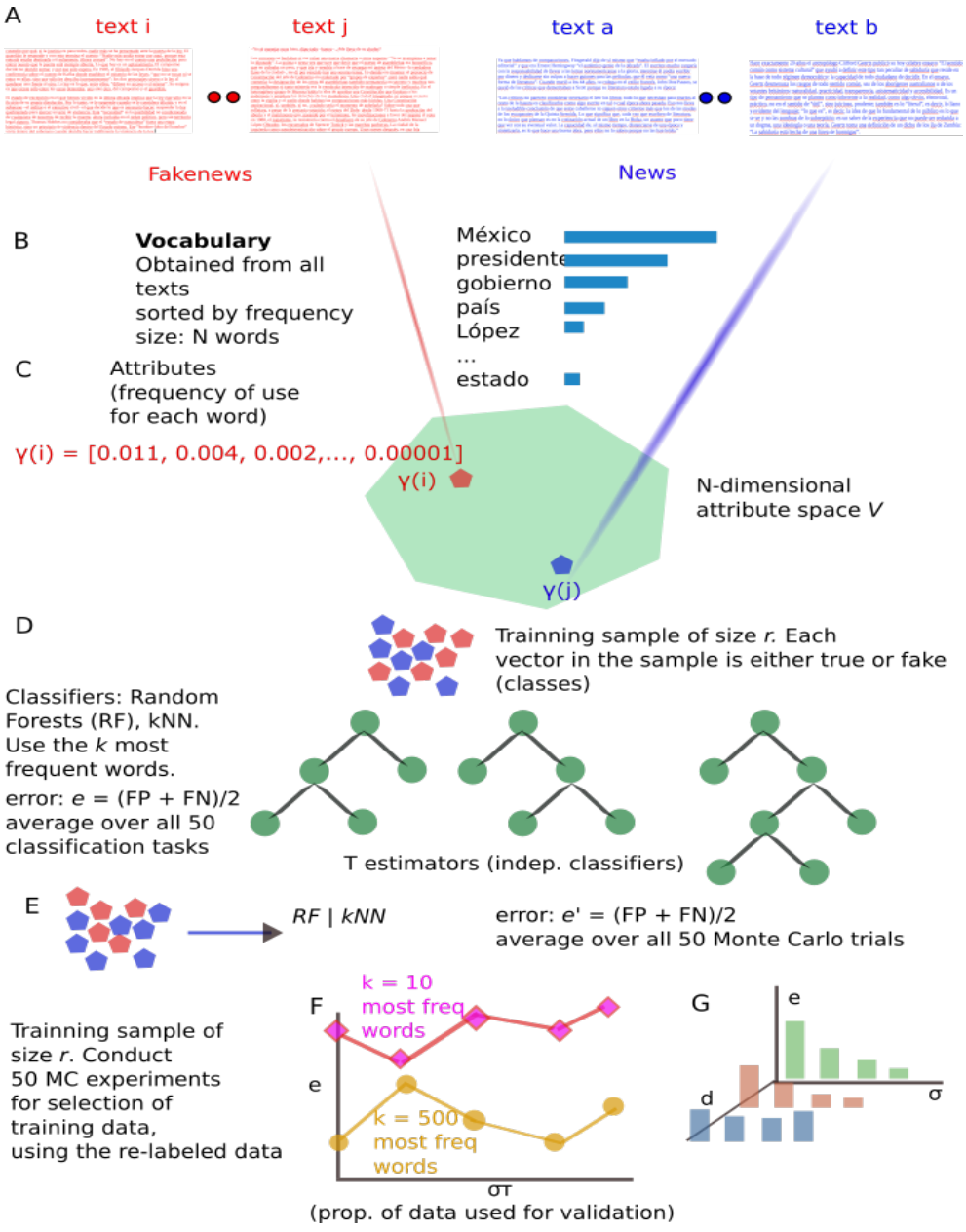

Fig. 1. The algorithm. The methodology is based on obtaining the vocabulary from all texts, both true and fake. Then, each text is characterized as a point in a $d-$dimensional space. Each dimension is the relative frequency of use of the k most common words among all texts.

both, interpretable, as well tend to present low classification errors [18]. It is an ensemble method that generates a group of decision trees [19], and for each one of them, randomly selects a subsample of the whole training set. The overall decision about the class a given vector belongs is computed taking into account the decision made by the majority [20] of the decision trees. In several implementations, the subsamples are randomly modified as to give more robustness to the ensemble.

The second classifier we tested is kNN. It is an algorithm of the family commonly known as *guilty by association*. Here, the label of the vector to be classified is determined by the most common label among its k nearest neighbors. The k neighbors of a vector v are its closest k vectors, based on Euclidean distance.

4 Results

The dataset T, consisting of 678 texts [2], in which half were true and half fake, was used to tune the algorithm. We followed a permutation approach to create several subsambles from T, and train. The first stage, as described in the previous section, was to obtain the vocabulary from the available corpora. Fig. 2 shows the 25 most common words, leaving aside the majority of prepositions, conjunctions, connectors, and so on.

The number of attributes that are relevant in the classification task in hand, that is, the dimension of the feature space, is an unknown quantity. We estimated it by increasing d from 10 to 1000, and evaluating the system classification capabilities when the d most frequent words were considered. We were interested in the validation error, and for that, we split the dataset T in two sets, the training and the validation sets. Fig. 3 shows the classification error for the validation dataset σ derived from T. σ is the percentage of T used for validation, whereas $1-\sigma$ is the percentage of T used for training. We varied σ from 0.02 to 0.5. For each pair of (d, σ), 50 Monte Carlo trials were conducted, selecting each time a possible different training and validation datasets. The average error is reported.

For RF, the maximum depth of the trees was $min(10, 3 \times log_{10}(d))$ and the number of estimators was $2 \times d$. For kNN, we varied k as a function of d by $k = max(5, 2 \times log_{100}(d))$. As observed for kNN, we linked k and the dimensionality d.

As observed in the figure, the classification error (number of misclassified texts) did not reach good error measures. Thus, the hypothesis that a n-gram approach, with n = 1, using the $d-$ most frequent words in the vocabulary is enough to classify a text as either true or false (fake) has to be rejected.

Once we trained the RF and kNN classifiers, we tested it in the second stage of the competition. The results were unsatisfactory. The performance for RF was 0.56 and for kNN was 0.54. Both results were below the average for the rest of the competitor.

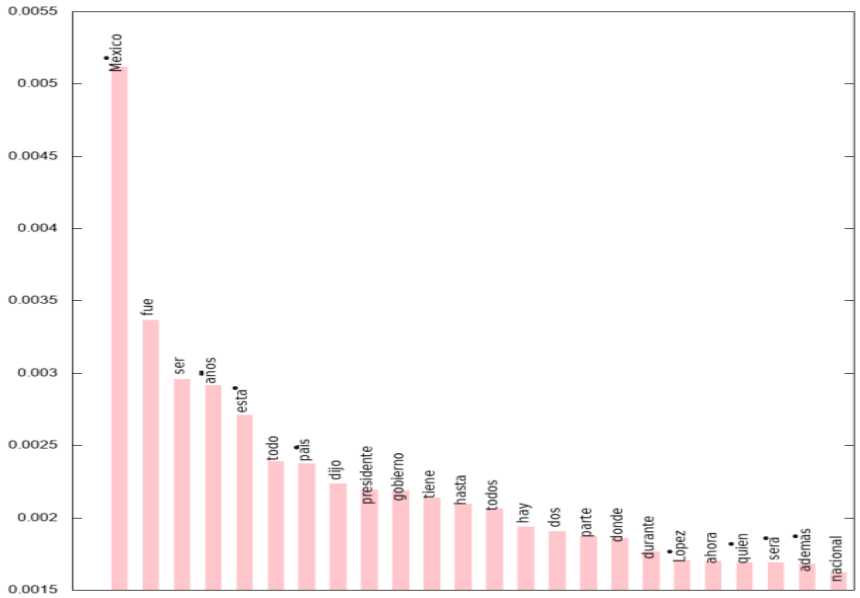

Fig. 2. The 25 most words without prepositions, conjunctions and connectors. At least 1/3 of the true and 1/3 of the fake news had to make use of the word in order to consider that specific word, otherwise, it was dropped from the vocabulary.

5 Conclusions

The hypothesis that true and fake news can be tell apart based on the relative frequency of words has to be rejected, as shown in the previous section. We have shown that the n-gram approach, with n = 1, linked to random forests and k-nearest neighbors, did not offer good results.

The reason for the low performance over the test dataset of the second stage of the challenge, was that the vocabulary varied to an extent in which several of the most frequent words were not present in one of the two corpora. This limitation cannot easily be solved.

We plan to use a more complex feature selection algorithm, at the time that we will make use of deep learning architectures over the attributes described in this contribution. We intend to verify whether a deeper algorithm could offer better results over the same attributes, or if there is something inherently incomplete in them.

References

1. Gómez-Adorno H, Posadas-Durán JP, Bel-Enguix G, Porto C. Overview of FakeDeS Task at Iberlef 2020: Fake News Detection in Spanish. Procesamiento del Lenguaje Natural. V. 67 No. 0 (2021).

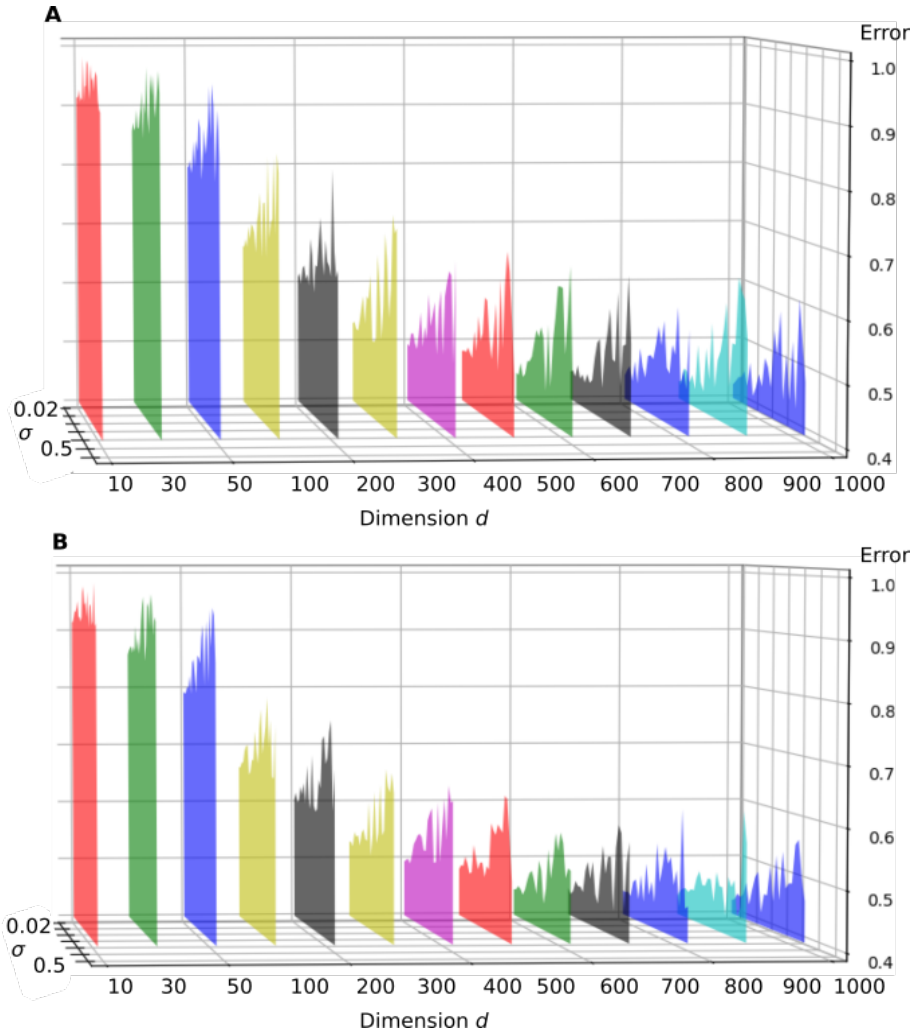

Fig. 3. Test error as a function of the number of words (dimensionality) d, and the percentage of the sample used to test, σ. The percentage of the dataset T used to train the classifiers was thus 1 - σ. A. Error from random forests. B. Error from kNN.

2. Posadas-Durán JP, Gómez-Adorno Helena, Grigori S, Escobar J, Moreno J. Detection of fake news in a new corpus for the Spanish language. Journal of Intelligent & Fuzzy Systems. Vol 36, No. 5 4869-4876. (2019)
3. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Carmona, M., Mellado, E., Albornoz, J., Chiruzzo, L., Freitas, L., Adorno, H., Gutiérrez, Y., Zafra, S., Lima, S., Arco, F. & Taulé, M. The overview of the Iberlef 2021. *Proceedings Of The Iberian Languages Evaluation Forum (IberLEF 2021)*. (2021)
4. Desai, S., Mooney, H. & Oehrli, J. "Fake News," Lies and Propaganda: How to Sort Fact from Fiction. *University Of Michigan Library*. (2021), https://guides.lib.umich.edu/fakenews
5. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 09 Mar 2018: Vol. 359, Issue 6380, pp. 1146-1151 DOI: 10.1126/science.aap9559
6. Hailey M. Fake News and the Sociological Imagination: Theory Informs Practice. Forty-sixth National LOEX Library Instruction Conference Proceedings (Library Orientation Series No. 51). http://hdl.handle.net/2027.42/143532
7. ADEMÁS DE PANDEMIA POR COVID-19, MÉXICO ENFRENTA PROPAGACIÓN DE NOTICIAS FALSA, May, 2020. howpublished=`https://www.dgcs.unam.mx/boletin/bdboletin/2020_318.html`
8. Ahmad I, Yousaf M, Ahmad M. Fake News Detection Using Machine Learning Ensemble Methods. Volume 2020. Complexity. doi.org/10.1155/2020/8885861
9. Sharma U, Saran S, Patil S. Fake News Detection using Machine Learning Algorithms. International Journal of Engineering Research and Technology (IJERT) NTASU – 2020 (Volume 09 – Issue 03).
10. Maslej V, Butka. (2019). Deep learning methods for Fake News detection. 10.1109/CINTI-MACRo49179.2019.9105317.
11. Posadas Duran J, Sidorov G, Gomez Adorno H, Moreno J. Detection of Fake News in a New Corpus for the Spanish Language. Journal of Intelligent and Fuzzy Systems · May 2019 DOI: 10.3233/JIFS-179034
12. Harriz Z. Distributional Structure. Word. 10 (2/3): 146–62. doi:10.1080/00437956.1954.11659520 (1954).
13. Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail AAAI'98 Workshop on Learning for Text Categorization. (1988).
14. Boughaci D, Benmesbah M, Zebiri A. An improved N-grams based Model for Authorship Attribution 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia. 1-6, doi: 10.1109/ICCISci.2019.8716391. (2019).
15. Mikolov T, Sutskever I, Chen K; Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. arXiv:1310.4546. Bibcode:2013arXiv1310.4546M. (2013).
16. Gómez-Adorno H, Sidorov G, Pinto D, Vilarino D, Gelbukh A. Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. Sensors. Vol. 16. No. 1374. doi:10.3390/s16091374. (2016).
17. Shrestha P, Sierra S, Gonzalez F, Montes M, Rosso P, Solorio T. Convolutional Neural Networks for Authorship Attribution of Short Texts. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2 (2017).
18. Ho TK (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal. 278–282. (1995).
19. Quinlan JR Induction of decision trees. Machine Learning. 1: 81–106. doi:10.1007/BF00116251 (1996).

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