Automatic Detection of Fake News in Spanish: Ecuadorian Political Satire

Mafla, Nicolás [1,*]; **Flores, Miguel** [2]; **Castillo-Páez, Sergio** [3]; **Andrade, Roberto** [4]

[1]*Escuela Politécnica Nacional, Facultad de Ciencias, Quito, Ecuador*
[2]*Escuela Politécnica Nacional, Facultad de Ciencias, Departamento de Matemática, Grupo MODES, SIGTI, Quito, Ecuador*
[3]*Universidad de las Fuerzas Armadas ESPE, Departamento de Ciencias Exactas, Ecuador*
[4]*Escuela Politécnica Nacional, Facultad de Ingeniería en Sistemas, Quito, Ecuador*

Abstract: The circulation of fake news on internet, especially those of political satire through social media, has affected the majority of the Ecuadorian population. This work presents a methodology based on statistical learning that accurately and automatically detects fake news in Spanish using machine learning and natural language processing techniques. The document begins by presenting basic concepts related to fake news and works related to their automatic detection. The second section explains the news corpus creation process, text processing, numerical representation with TF-IDF and training of supervised classification algorithms with two different data sets. Results obtained from the training are analyzed in the third section, being the models with support vector machines the ones that offer the best predictions, improving approximately 15%, 6% and 3% to the performance of the models with naive bayes, random forests and boosting trees respectively. Finally, conclusions of the research and future work is presented in the fourth section.

Keywords: Fact-checking, machine learning, natural language processing, supervised classification

Detección Automática de Noticias Falsas en Español: Sátira Política Ecuatoriana

Resumen: La circulación de noticias falsas en internet, especialmente las de sátira política a través de redes sociales, ha afectado a la mayoría de la población ecuatoriana. Este trabajo presenta una metodología basada en el aprendizaje estadístico que detecta de forma precisa y automática noticias falsas en español utilizando técnicas de aprendizaje automático y procesamiento del lenguaje natural. El documento comienza presentando conceptos básicos relacionados con las noticias falsas y trabajos relacionados con su detección automática. La segunda sección explica el proceso de creación del corpus de noticias, procesamiento de los textos, representación numérica con TF-IDF y entrenamiento de algoritmos de clasificación supervisados con dos conjuntos de datos diferentes. Los resultados obtenidos del entrenamiento se analizan en la tercera sección, siendo los modelos con máquinas de soporte vectorial los que ofrecen mejores predicciones, mejorando aproximadamente un 15%, 6% y 3% al rendimiento de los modelos con naive bayes, random forests y árboles boosting respectivamente. Finalmente, las conclusiones de la investigación y el trabajo futuro se presentan en la cuarta sección.

Palabras claves: Fact-checking, machine learning, procesamiento del lenguaje natural, clasificación supervisada

1. INTRODUCTION

The time we currently spend browsing the internet and consuming content on social media occupies a large part of our day (Bergström and Jervelycke Belfrage, 2018), without a doubt they have become one of the most powerful communication tools used today since it allows access to consumption and disclosure information instantly available to anyone. This has led to the preference of digital media over traditional media, consequence of this free access and the easy generation of content, the information that circulates online is not always reliable (López-Buroll et al., 2018). The growth of social networks increases the spread of fake news on the internet, information distributed by these media is massive, fast and heterogeneous, which can cause serious impact on the entire society (Zhang and Ghorbani, 2020).

Fake news has had a negative impact on several sectors, among these, the most important communication, economy, health and politics; mentioning some cases we have: the 2016 US presidential elections (Allcott y Gentzkow, 2017), the threat to global public health caused by the massive infodemic originated around the pandemic produced by the covid-19 virus (Pulido et al., 2020) and a curious case is of a scientific publication in the American Journal of Biomedical Science & Research

*nicolas.mafla.checa@gmail.com
Recibido: 11/10/2021
Aceptado: 01/05/2022
Publicado: 23/12/2022
10.33333/rp.vol50n3.01

(Shelomi, 2020), noting that even the scientific community is not exempt from sharing false information that pretends to be true.

The problem of fake news that circulates on the internet and also on social networks affects us all directly or indirectly, it is for this reason that this research provides the theoretical bases for the creation of applications that help to fight against media misinformation, as mentioned above, many sectors are affected. Additionally, the development that is achieved will potentially serve for future research on this topic and will be an advance with respect to supervised classification models that involve natural language processing techniques in the Spanish language.

The objective of this research is to create a model that allows to accurately and automatically identify fake news in Spanish from Ecuadorian news and satire pages. For this, a text corpus was created extracting the news manually or using web-scraping techniques, which were later processed with NLP techniques to create the database that served to train the algorithms and compare their results.

1.1 Fake news

The term fake news is recent and has gained popularity in last years, but false information has been part of humanity for a long time, fake news is as old as the printed news that circulated since the invention of the printing press, or even older (Soll, 2016). There is no universal definition for fake news and it is not easy to formulate a generally accepted one for the term, since these tend to be diverse in terms of topics, styles and even platforms, which is why several authors have proposed their own definitions. Fake news is defined as manufactured information that imitates the content of the media in form but not in the organizational process or intention, in addition they lack the standards and editorial processes of the media that ensure the accuracy and credibility of the information (Lazer et al., 2018). They refer to all kinds of false stories that are published and distributed mainly on the internet, in order to deceive or deliberately entice readers for financial, political or other benefit (Zhang and Ghorbani, 2020). And lastly, fake news is news articles that are intentionally created and verifiable false that could mislead readers (Allcott y Gentzkow, 2017). Based on how the concepts presented were created, they share three characteristics in common, the first is related to the authenticity of the news information (containing some statement based on facts or not), the second refers to the intention to create fake news (with the aim of deceiving or entertaining the public), and finally, if they are really news (Zhou and Zafarani, 2020).

1.2 Knowledge-based detection

According to Zhang and Ghorbani (2020), and Zhou and Zafarani (2020), a process known as fact-checking is generally used to detect fake news from a knowledge approach, this method was initially developed by journalists, and is currently used by a large part of the media. Fact-checking process aims to verify the authenticity of the information, comparing the knowledge extracted from the content of the news to be verified with known facts. Both evaluation criteria and visual metrics are used to determine the level of veracity of the news in the fact-checking process.

Manual fact-checking is usually carried out by a select group of professionals called fact checkers of great credibility since this leads to highly accurate results. This process can require a lot of execution time with a high maintenance cost, it also presents difficulties when the information content to be verified is massive. Another way of conducting fact checking is data verification through collective sources, which is based on a large population of regular individuals acting as fact-checkers.

Compared to expert-based fact checking, it is relatively difficult to administer, less credible and accurate due to the political bias of its verifiers, but having better scalability. Manual fact-checking does not adapt to the volume of information that is created every day online, especially on social media. For a better scalability of fake news detection, automatic fact-checking techniques have been developed, which are largely based on extraction of information through natural language processing and classification using machine learning techniques (Bondielli and Marcelloni, 2019).

Given the diversity of ways of speaking Spanish, which largely depends on the geographical area where the speaker comes from, this research focused on creating a model capable of identifying fake news in Spanish from Ecuador. The model was trained from a set of publications identified as real and fake news, extracted from the main national newspapers and pages of political satire on Facebook, in addition to comparing different types of classification algorithms used in the detection of fake news in the English language (Bondielli and Marcelloni, 2019).

1.3 Related work

Singharia et al. (2017), present a three-level hierarchical attention artificial neural network (3HAN): words, sentences and headings for accurate detection of fake news. The model is based on the representation of a news article as a vector, which is used to classify an article by assigning a probability of being false. The data used for the training of the model belongs to the period of the USA presidential elections of 2016 and was extracted from the PolitiFact platform to create a set of fake news and from the list of popular verified sites in the USA provided by Forbes to create the set of real news. Ciampaglia et al. (2015), present a model based on networks or graphs, in which the verification of facts can be approximated quite well by finding the shortest path between nodes denoting statements under properly defined semantic proximity metrics on knowledge graphs. The data used was extracted from Wikipedia, which includes all the factual statements extracted from the Wikipedia information boxes, thus creating a knowledge graph with 3 million entities linked by approximately 23 million edges. Posadas-Durán et al. (2019), present a model to analyze and detect misleading information present in a large number of Spanish-language websites. For the training of the model, a set of news collected manually from different websites was used to create a corpus of news labeled as fake and real news. The training was carried out using supervised classification algorithms: vector support machines, logistic regression, random forests and gradient boosting, evaluating performance by removing stop words and

considering stop words. Extraction of information and its representation was carried out through linguistic characteristics obtained from three techniques: bag-of-words, n-grams and POS tags n-grams.

2. METHODOLOGY

In the absence of a set of news data classified as true and false, the first step was to create it from Ecuadorian information sources available on the internet and social media. News texts were cleaned and represented numerically with the TF-IDF technique to create the detection models. In order to compare the ability to detect fake news, four algorithms were compared: vector support machines, random forest, boosting trees and the naive bayes classifier. The entire process from extraction to modeling was carried out in Python, specifically for the training of the classification algorithms scikit-learn was used (Pedregosa et al., 2011). Methodology for creating the models from text processing to algorithm training was based on followed by Posadas-Durán et al. (2019).

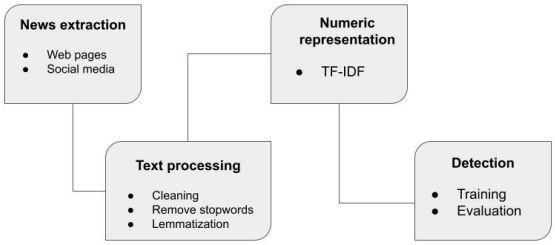

Figure 1. Graphic diagram of the problem modeling process

2.1 News extraction

The corpus consists of a collection of news from the Ecuadorian sphere extracted from different pages within the internet: websites of the main real national newspapers and pages of political satire on social networks, collected from September to December 2020 what is the period in which the information gathering process was carried out.

The publications made by the political satire pages on Facebook used in this work are characterized by being short messages, so the information obtained from the newspaper pages was not the complete story, it was the summary of the news provided by the same newspapers in order to maintain the pattern that the pages of satire follow. To label the collected news, the following was considered:

- The news was labeled as real if there was evidence that the publication came from a trusted page, such as it is for newspaper web pages.
- The news was labeled as false if there was no evidence about the veracity of the information presented and the credibility of the page from which this news came, in this case the pages of political satire.

2.2 Text processing

The components of interest in this work are the words of each news item, which is why elements that cannot be considered words were extracted from the corpus. The elements removed from the texts were:

- Links referring to external pages
- Tags to users´ accounts
- Hashtags
- Punctuation marks
- Numeric and special characters

Finally, all the words frequently used in languages that are not helpful by themselves in understanding the context of the news and in general of a text called stopwords were removed and all document in the corpus were transformed to lowercase.

Lemmatization is the process in which given all the different inflected forms of a word, its base form is found, this helps us to considerably reduce the number of words with similar meanings in a text. This task was carried out with the help of the text processing library spaCy, in it there is a great variety of pre-trained models based on neural networks in different languages including Spanish.

2.3 Numerical representation

To numerically represent the news texts, the TF-IDF technique was used, which is based on representing each of the texts as a vector. The term frequency (TF) measures the frequency with which a term appears in a given document, while the inverse document frequency (IDF) measures the importance of a term within the corpus, weighing less weight to the terms that are very common within the corpus, while it weighs more unusual. Multiplying both metrics gives the TF-IDF representation (Vajjala et al., 2020).

$$TF\text{-}IDF(p_i,d_j) = \frac{f(p_i,d_j)}{|d_j|} \cdot log\left(\frac{M}{|\{d \in D : p_i \in d\}|}\right) \quad (1)$$

For a corpus $D = \{d_1,...,d_M\}$ and a vocabulary $V = \{p_1,...,p_N\}$, where $f(p_i,d_j)$ is the relative frequency of the word p_i of the document d_j in the corpus D.

2.4 Fake news detection

The problem consists of predicting if a news item is fake based on the information provided by it, captured in the form of vectors with the TF-IDF technique. The response variable that we want to predict is defined as a binary variable, this will take the value of 1 if it is a fake news and 0 if it is a real news. For the creation of the models, the usual process of training of supervised classification algorithms was followed, partitioning the data set in training,

validation and tests to obtain the optimal hyperparameters and the algorithms that presented the best performance to new data sets.

In several investigations, it has been shown that classical supervised classification methods solve the problem of detecting fake news in the English language with excellent results, the ones that stand out the most are: SVM, random forests and boosting algorithms (Zhou and Zafarani, 2020). On the other hand, in the research carried out by Posadas-Durán et al. (2019) they solve the detection problem in the Spanish language with the same classifiers and additionally logistic regression obtaining good results. For these reasons, the classifiers mentioned above were chosen, except for regression, which was replaced by the naive bayes classifier since it has less demanding assumptions.

SVM

The support vector machine is a generalization of a linear classifier called the *maximal margin classifier*, that has the objective of solving binary classification problems in which the data set has two classes but these are not separable by a linear boundary. To achieve this, the space of the training data is transformed into one of higher dimension in which a hyperplane can be fitted that maximizes the separation of the two classes. The problem solved is the following:

$$\begin{aligned} &\max_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.a} \quad &y_i(\beta_0 + \beta\cdot\phi(x_i)) \geq 1-\xi_i, \forall i \in \{1,...,n\} \\ &\xi_i \geq 0, \quad C>0 \end{aligned} \tag{2}$$

The hyperplane that solves the classification problem is represented by $\beta_0 + \beta\cdot\phi(x_i)$ where ϕ is the function with which the space of the training data is transformed. $\xi_1,...,\xi_n$ are variables that allow the problem to be more flexible and allow a given number of observations to be on wrong sides of the hyperplane controlled by the parameter C.

The classification of an observation x is given by the expression $sign\{\beta_0 + \sum_{i=1}^{n}\alpha_i y_i \phi(x_i)\cdot\phi(x)\}$ and here is the importance of the kernels, since it is not necessary to have an explicit representation of the function ϕ, it is enough to define the inner product of the transformed data, so the kernel functions are represented as follows:

$$K(x,x') = \phi(x)\cdot\phi(x') \tag{3}$$

There are several kernel functions that are used to solve classification problems with SVM, the most used and the ones we chose for this investigation are: linear, sigmoid and gaussian.

Random Forest

The Random Forest algorithm involves a sequence of models that segment the predictor space into simple regions and its classification rules can be modeled by a series of related decisions known as *decision trees*. The most common way to measure the power of division in trained classification trees is the classification error rate, this measures the proportion of observations that do not belong to the most common class, in practice more sensitive measures are used, such as gini index or *cross-entropy* coefficient, since these are a way of measuring *impurity* of the segmentation. The algorithm is shown below:

Algorithm 1: Random Forest for classification

For $m = 1$ **to** M **:**

(a) Take a bootstrap sample Z of size N from the training space

(b) Fit a tree T_m from the sample Z in which each division of the tree is carried out with the following steps

1. Select m variables from the total of variables
2. Take the best variable of the selected m to split the tree based on the metric used (gini or cross-entropy)
3. Segment the space

Output: Collection of trees $\{T_m\}_1^M$

The classification for an observation x will be the class to which x belongs that is repeated the most in the predictions of the trees $\{T_m\}_1^M$

Boosting Trees

Similar to random forests, boosting models are based on the use of a series of weak that together provide a great predictive power and are widely used for regression and classification problems. Unlike random forests, decision trees are not trained on bootstrap samples, instead they are trained on sequential modifications of the training space, each decision that is fitted uses the information from its predecessors.

The objective is to repeatedly apply a weak classifier to modifications of the training set, generating M new weak classifiers and these will be used for the final prediction. The classification of an observation x will be the weighted combination of the predictions of these weak classifiers.

$$G(x) = sign\left\{\sum_{m=1}^{M}\alpha_m G_m(x)\right\} \tag{4}$$

where α_m controls the contribution that the decision tree G_m makes to the final model, giving more relevance to those that are more accurate. The modifications made to the training set consist of applying weights $w_1,...,w_N$ to the N observations $x_1,...,x_N$, the weights are initialized with $w_i = \frac{1}{N} \quad \forall i = 1,..,N$ and as the iterations go by, the weights are modified one by one, making the algorithm classification uses the new weighed data.

In each iteration, the observations that were erroneously classified increase their weight, while those correctly classified have their weight reduced. Thus, as the process progresses, the observations with problems being classified acquire more relevance for the following iterations, forcing to the following classifiers to concentrate on those that did not obtain an accurate prediction by the previous classifiers.

Naive Bayes Classifier

The naive bayes classifier is based on Bayes theorem, which is used to estimate the conditional probability of the occurrence of an event given a certain amount of information about the event. For a training space with p predictors and response variable Y can take K two classes, using Bayes the prediction that the observation x belongs to the class K is the following:

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)} = \frac{\pi_k \prod_{j=1}^{p} f_{jk}(x_j)}{\sum_{i=1}^{K} \pi_i f_i(x)} \quad (5)$$

where $\pi_k = P(Y = k)$ is the prior probability that an observation taken at random belongs to class k, $f_k(X) = P(X|Y = k)$ denotes the density function of X for observations from classes k. This model assumes that given a class k the marginal distributions of the predictors $X_1, ..., X_p$ are independent and normal distributed.

2.5 Model evaluation

The evaluation of the models was measured by the F-score, area under the receiver operating characteristic curve (ROC) and the estimation of the general prediction error. By having an imbalance in the classes of the data, that is, there is much more true news than fake news, just measuring the accuracy of the model could lead to biased results (Zaki and Meira, 2014). For this reason, the F-score measure was chosen since it combines both the precision and the recall of the model, thus having a more real measure of the prediction power.

$$F_P = \frac{2TP}{2TP + FP + FN} \quad F_N = \frac{2TN}{2TN + FN + FP}$$
$$F = \frac{F_P + F_N}{2} \quad (6)$$

Where TP, TN, FP and FN are the true positives, true negatives, false positives and false negatives respectively of the model in the test set.

Algorithm 2: General test error estimation (K-fold cross validation)

Partition Corpus D in $[D_1, ..., D_K]$ equal parts
for $i = 1$ **to** K **do**
  $G_i \leftarrow$ train the classifier on $D \setminus D_i$
  $\hat{Err}_i \leftarrow$ calculate prediction error of the classifier G_i on $D \setminus D_i$
end
$\hat{Err} = \frac{1}{K} \sum_{i=1}^{K} \hat{Err}_i$
Output: $\hat{Err}$

3. RESULTS

Collecting the text of the pages on Facebook was done manually by copying the content, on the other hand, the texts collected from the pages of the newspapers was done with the help of web scrapping. Pages of national newspapers chosen were: "El Comercio" and "El Universo" for the real news group and political satire pages were "El Uninverso", "El Culimercio" and "El Mercioco" for the fake news group.

Table 1. National newspaper pages and political satire pages

Page	Website	Type of news
El Universo	www.eluniverso.com	National newspaper
El Comercio	www.elcomercio.com	National newspaper
El Uninverso	www.facebook.com/DiarioUninverso	Political satire
El Culimercio	www.facebook.com/elculimercioec	Political satire
El Mercioco	www.facebook.com/mercioco	Political satire

The news corpus consists of a total of 1629 news items, approximately 59% of these belong to "El Comercio", since due to the structure of its website, it facilitated the process of collecting the news through webscrapping, so it was possible to extract large amount of news with great speed. In general, the pages that create false information tend to have a short lifetime (Allcott y Gentzkow, 2017), so the information obtained from two of the satire pages is much less than that obtained from the pages of national newspapers, except from the "El Uninverso" page, which is a fairly active page that has published a lot of content since its creation; the advantage of this was that the time invested for the extraction of the false texts was not excessive and it was feasible to execute it manually.

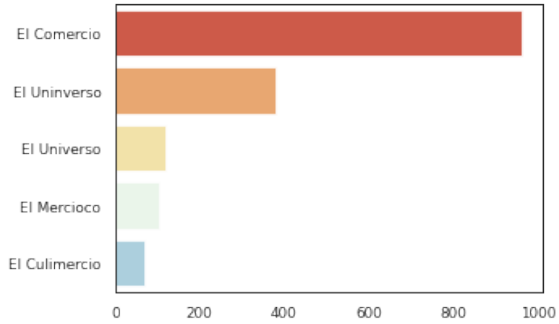

Figure 2. Distribution of the news collected according to the page where they come from

Of the total news we have that 34% corresponds to fake news and the remaining 66% to real news. Analyzing the real news and the fake news from a point of view at the level of terms, we can see that the distribution of the number of terms is quite similar for both groups, the average of terms for the real news is approximately 41 terms while for the fake news is about 42. On the other hand, the number of terms present in the fake news are a little more variable than the real ones, having 15.7 and 19.6 the approximate deviations respectively. The resulting news set was stored in a comma separated values (csv) file with utf-8 encoding to be able to store words in Spanish without losing the accents and the letter ñ.

Vocabulary consists of 9953 words, as this technique depends on the total number of words in a corpus, if the resulting vocabulary consists of an excessive number of words, this quantity will be the number of variables when representing the texts numerically, so remove stop words and lemmatization helped us reduce the number of variables without losing the structure of the text. The original database created has 1629 records with 9953 variables, but applying the mentioned reduction techniques, the number of

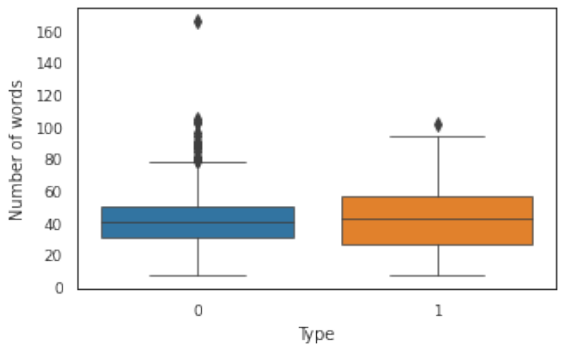

Figure 3. Distribution of the number of terms by news groups

variables decreases to 6336. It was analyzed whether reducing the number of variables with these techniques affects the performance of the classification models, so these two data sets were used to train the algorithms.

Since there is only 34% fake news, the train database was balanced in order not to obtain biased results. To balance the data sets, a technique called SMOTE (Bowyer et al., 2011) was used, which unlike the traditional ones based on resampling with replacement, this creates synthetic observations of the minority class using its K closest neighbors. The minority class is oversampled by taking each of the observations and introducing synthetic observations along the segments that join them with one or all of the K closest neighbors, that is, the difference between an observation and its closest neighbor is taken and this difference is multiplied by a random number between 0 and 1, thereby selecting a random point along the length of the two observations.

F-score values for all the models in general are above 80% which is very good since this indicates that the models make correct predictions most of the time. In the same way, the area under the ROC curve for all the models in general is excellent, having values very close to the best of the cases, this indicates that the models can easily differentiate fake news from real news. The results are shown in Table 2 and models trained from unprocessed texts are represented by $(*)$.

Reducing the number of variables in the data set when processing the news texts could reduce the predictive power of the models, which is true, but it is clear that in this case the reduction is minimal, since the values F-score and area under ROC curve for models without stop words and lemmatization are practically equal to models with stop words and non-lemmatization.

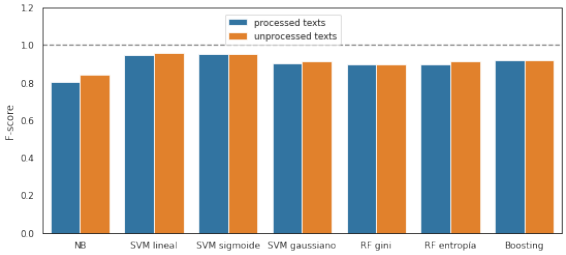

Figure 4. Results of the F-score measure for processed texts and unprocessed texts

The estimate of the expected prediction error was obtained through the cross-validation process with $K = 5$ and $K = 10$ partition groups of the data set. With $K = 5$ of the 1629 news of the corpus, 5 groups of approximately 326 news each were randomly taken, with which the 5 models were trained with a set of approximately 1304 and the predictions were made on approximately 326 news. In a similar way, the estimates were made with $K = 10$, thus, from the 1629 news of the corpus, 10 groups of approximately 163 news each were randomly taken, with which the 10 models were trained with a set of approximately 1630 and made the predictions on approximately 163 news items.

Table 2. National newspaper pages and political satire pages

model	Error	AUC	F-score
SVM linear	4.91%	99,24%	94.41%
SVM sigmoid	4.29%	99,2%	95.13%
SVM gaussian	7.98%	99,31%	90.35%
SVM lineal*	3.68%	99.40%	95.81%
SVM sigmoid*	4.29%	99,30%	95.08%
SVM gaussian*	7.36%	99.20%	91.35%
NB gaussian	17.79%	81.09%	80.36%
NB gaussian*	13.5%	83.37%	84.31%
RF gini	8.59%	97.47%	89.67%
RF entropy	8.59%	97.63%	89.91%
RF gini*	8.59%	98.30%	89.91%
RF entropy*	7.36%	97.93%	91.44%
Boosting trees	6.75%	97.27%	92.11%
Boosting trees*	7.36%	97.60%	91.61%

Based on the results obtained, the models are able to accurately predict whether a news item is real or fake, achieving high values for the F-score measure, being the models with vector support machines with linear and sigmoid kernels that offer the best predictions, reaching accuracy above 95%. Apparently the preprocessing of the texts does not have much relevance on the predictive power of the models, but if we analyze Figure 5, this process does have a direct impact on the estimates of the prediction errors, these are much less variable and in lower value for the models with stop words and non-lemmatized texts than for the models without stop words and lemmatized texts, which is related to the reduction of information as a consequence of reducing the number of variables in the data set; This is true for the models with vector support machines and the two models with a naive bay classifier, since for the models trained with processed texts that use decision trees (boosting and random forests), they present estimates of prediction errors even better than models with decision trees with raw texts.

4. CONCLUSIONS

The study carried out demonstrates the capacity of the supervised classification algorithms to identify with great precision fake news, with F-score scores above 90%, on political satire pages focused on Ecuadorian problems that circulate on social networks, whose objective is not malicious, it is humorous, but the information that is created and shared can be misinterpreted out of context and can mislead people. Social networks have proven to be a powerful mass communication tool, but they are also an excellent source to obtain data from both users and the pages created on these platforms. Information extracted from Facebook

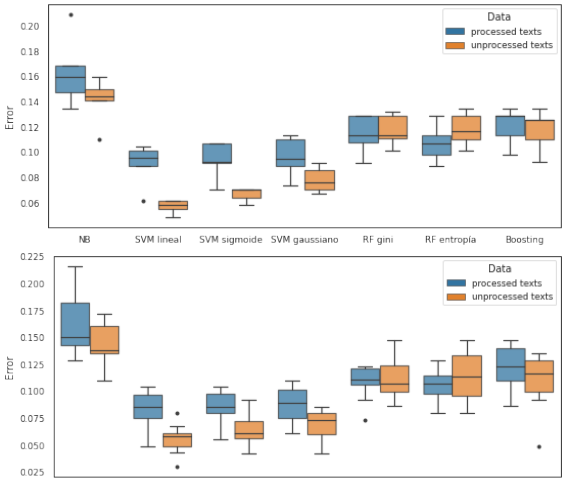

Figure 5. Estimation of the expected prediction error using cross validation with 5 (first) and 10 (second) groups for each of the proposed models with their variants

Table 3. Comparison of mean prediction error estimate

comparation		statistic	p-value
SVM linear	NB gaussian	7.522058	2.138598e-06
	RF gini	3.495373	0.002619
	RF entropy	3.092006	0.006365
	Boosting trees	4.199826	0.000630
SVM sigmoid	NB gaussian	7.507324	3.397924e-06
	RF gini	3.370072	0.003411
	RF entropy	2.940428	0.008746
	Boosting trees	4.093132	0.000895
SVM gaussian	NB gaussian	7.227926	4.703457e-06
	RF gini	2.973300	0.008146
	RF entropy	2.547322	0.020219
	Boosting trees	3.782543	0.001662
SVM linear*	NB gaussian*	12.078495	1.391164e-09
	RF gini*	7.655993	8.332308e-07
	RF entropy*	6.752743	7.627163e-06
	Boosting trees*	6.125672	1.846512e-05
SVM sigmoid*	NB gaussian*	10.544055	6.033120e-09
	RF gini*	6.304513	7.377372e-06
	RF entropy*	5.638357	4.115911e-05
	Boosting trees*	4.966191	1.356008e-04
SVM gaussian*	NB gaussian*	9.976353	1.608514e-08
	RF gini*	5.671745	2.765814e-05
	RF entropy*	5.069808	1.329329e-04
	Boosting trees*	4.363540	4.935202e-04

was vitally important to create the news corpus used to train the models, since without an expert verified fake news database, these pages were the only easily verifiable source for news with fake content and accessible to anyone.

Numerical representation of texts in the form of vectors was a very important aspect in this work, since it was the main piece on which it was based to obtain the necessary data for the training process. Inverse term frequency technique captured the form of writing to a fake news since as we saw this is based on calculating the importance of each of the words both at the level of the news and at a more general level within the corpus. The writing style of fake news is similar to that of a real news story since it pretends to simulate being one but the type of words that are used to elaborate them are different, especially because they make use of words typical of the Ecuadorian slang and it is precisely this what the

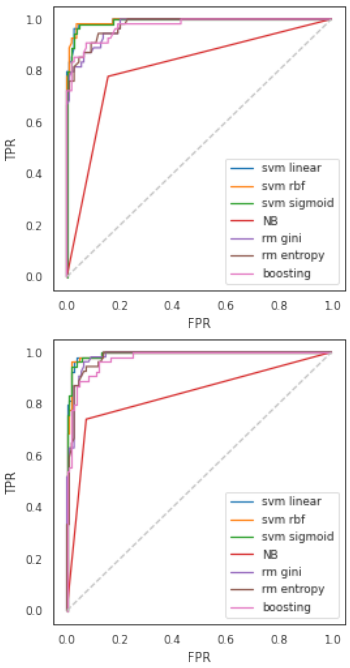

Figure 6. Roc curves of the models. Models trained with lemmatized texts and without stop words (left) models trained with non-lemmatized texts and with stop words (right)

inverse term frequency manages to represent numerically. The number of variables resulting from the numerical representation of the news of the corpus for this case was close to ten thousand, which translates to a large computational expense, so that previously processing the texts was of great help to reduce the time of execution of the algorithms without losing almost any information and maintaining excellent performance of the models when predictions are made.

This study has been carried out with public data from social networks and from national newspaper pages, the investigation could be deepened by creating a more sophisticated model with the help of fact-checking experts to create a larger news corpus addressing more topics in order to generalize the detection of fake news. The model created has a great application within the technological field, thus being able to be used as a tool to help both individuals and companies to directly combat misinformation, especially in crisis situations in which social networks play a fundamental role within communication, which would achieve a better understanding of the behavior of users on social networks during critical events. This work is an advance with respect to the detection of fake news and natural language processing in Spanish, so it is recommended to deepen with different methodologies and adding elements such as images, videos or audios and new text processing techniques in the Spanish language.

REFERENCES

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* (31), 211-236. https://doi.org/10.1257/jep.31.2.211

Bergström, A. and Jervelycke Belfrage, M. (2018). News in Social Media: Incidental consumption and the

role of opinion leaders. *Digital Journalism* (6), 1-16. http://dx.doi.org/10.1080/21670811.2018.1423625

Bondielli, A. and Marcelloni, F. (2019). A Survey on Fake News and Rumour Detection Techniques. *Information Sciences* (497), 38-55. https://doi.org/10.1016/j.ins.2019.05.035

Bowyer, K., Chawla, N., Hall, L., Kegelmeyer, W. (2011). SMOTE: Synthetic Minority Over-sampling Technique *J. Artif. Intell. Res. (JAIR)* (16), 321-357. https://doi.org/10.1613/jair.953

Ciampaglia, G., Shiralkar, P., Rocha, L., Bollen, J., Menczer, F., Flammini, A. (2015). Computational Fact Checking from Knowledge Networks. *PloS one* (10). http://dx.doi.org/10.1371/journal.pone.0128193

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sunstein, C., Thorson, E., Watts, D., Zittrain, J. (2018). The science of fake news. *Science* (359), 1094-1096. http://dx.doi.org/10.1126/science.aao2998

López-Borrull, A., Vives-Gràcia, J. and Badell, J. (2018). Fake news, ¿amenaza u oportunidad para los profesionales de la información y la documentación? *El Profesional de la Información* (27), 1346. http://dx.doi.org/10.3145/epi.2018.nov.17

Posadas Durán, J., Gomez Adorno, H., Sidorov, G., Moreno, J. (2019). Detection of fake news in a new corpus for the Spanish language *Journal of Intelligent & Fuzzy Systems* (36), 4869-4876. http://dx.doi.org/10.3233/JIFS-179034

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

Pulido CM, Ruiz-Eugenio L, Redondo-Sama G, Villarejo-Carballido B. (2020). A New Application of Social Impact in Social Media for Overcoming Fake News in Health. *International journal of environmental research and public health* (17), 2430. http://dx.doi.org/10.3390/ijerph17072430

Shelomi, M. Opinion: Using Pokémon to Detect Scientific Misinformation. Obtained from: https://www.the-scientist.com/. (November, 2020).

Singhania S., Fernandez N., Rao S. (2017). 3HAN: A Deep Neural Network for Fake News Detection. *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science* (10635).

Soll, J. The Long and Brutal History of Fake News: Bogus news has been around a lot longer than real news. And it's left a lot of destruction behind. Obtained from: https://www.politico.com/magazine/. (November, 2016).

Vajjala, S., Majumder, B., Gupta, A., Surana, H. (2020). *Practical Natural Language Processing*.

Zaki, M. and Meira, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. *Cambridge University Press*, Cambridge University Press, USA.

Zhang, X. and Ghorbani, A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* (57), 102025. https://doi.org/10.1016/j.ipm.2019.03.0045

Zhou, X. and Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* (53), 40. https://doi.org/10.1145/3395046

BIOGRAPHIES

Nicolás Mafla. Born in Quito, in 2015 he entered to Mathematical Engineering with a mention in Statistics and Operations Research at EPN. During this period he was part of the CLAVEMAT project as a volunteer tutor, providing academic support in fundamental subjects for students of initial engineering levels. He obtained his degree as a Mathematical Engineer in 2021 and currently works as a data scientist in an Ecuadorian fin-tech.

Miguel Flores. Ph.D. in Statistics and Operations Research, Master in Statistical Techniques (University of La Coruña). Has experience in higher education and professional training, university and business in the field of Statistics & Machine Learning. Full Professor of the Probability and Statistics chair at EPN. Member of the Multidisciplinary Research Group on Information Systems, Technology Management and Innovation (SIGTI) of the National Polytechnic School and of the Modeling, Optimization and Statistical Inference Group (MODES) of the University of La Coruña.

Sergio Castillo. Basic academic formation is in Mathematical Engineering with a major in Statistics, Finance and Business Management, from the EPN; with a Higher Specialization in Finance from the Universidad Andina Simón Bolívar; a Master's Degree in University Teaching from ESPE, and then thanks to a

SENESCYT scholarship, He completed his PhD studies in Statistics and Operations Research, at the University of Vigo in Spain, specializing in research fields related to Geostatistics and Non-parametric Statistics.

Roberto Andrade. PhD student in Security Systems at the Faculty of Systems Engineering at EPN, his master's degree is in Network and Telecommunications Management at the Army Polytechnic School in 2013 and his engineering degree is in Electronics and Telecommunications at the National Polytechnic School (EPN) in 2007. Security Officer of the Ecuadorian Ministry of Education (MINEDUC) in 2015, Technological Infrastructure Coordinator at the National Planning Secretariat SENPLADES 2013-2014, Data Center, security and network administration in SENPLADES and Tecnología Sucre 2009-2013 and Technical Engineering for VoIP systems in SERATVoIP 2007-2011. He is a Certified CCNA, CCNP and CCNA Security Technical Instructor at EPN from 2010 to date.