# Detection of fake news in a new corpus for the Spanish language

**Article** *in* Journal of Intelligent and Fuzzy Systems · May 2019

DOI: 10.3233/JIFS-179034

CITATIONS
4

READS
710

**4 authors:**

Juan Pablo Francisco Posadas Durán
Instituto Politécnico Nacional
**17** PUBLICATIONS **169** CITATIONS
SEE PROFILE

Helena Gomez Adorno
Universidad Nacional Autónoma de México
**47** PUBLICATIONS **531** CITATIONS
SEE PROFILE

Grigori Sidorov
Instituto Politécnico Nacional
**226** PUBLICATIONS **1,745** CITATIONS
SEE PROFILE

Jaime Moreno
School of Mechanical and Electrical Engineering
**44** PUBLICATIONS **31** CITATIONS
SEE PROFILE

# Detection of Fake News in a New Corpus for the Spanish Language

*Pre-print version*

Juan-Pablo Posadas-Durán [a], Helena Gómez-Adorno [b,*], Grigori Sidorov [c], Jesús Jaime Moreno Escobar [a]

[a] *Escuela Superior de Ingeniería Mecánica y Electrica Unidad Zacatenco (ESIME Zacatenco), Instituto Politécnico Nacional, Mexico*

[b] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, Mexico*

[c] *Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional, Mexico*

**Abstract.** We present a new resource to analyze and detect deceptive information that is present in a huge amount of news websites. Specifically, we compiled a corpus of news in the Spanish language extracted from several websites. The corpus is annotated with two labels (real and fake) for automatic fake news detection. Furthermore, the corpus also provides the category of the news, presenting a detailed analysis on vocabulary overlap among categories. Finally, we present a style-based fake news detection method. The obtained results show that the introduced corpus is an interesting resource for future research in this area.

Keywords: Fake news, Corpus, Spanish, Resource, Machine learning

## 1. Introduction

The dissemination of information on social networks can be defined as a process in which news, events, and opinions are published, received and resent through users. The information disseminated in social networks follows a route, from one user to another and from one site to another, which allows the information to be traceable, however, the verification of this can be difficult. This is due to the fact that currently the volume of existing information exceeds the capacity of the human being to process and understand it.

Fake news provides information that aims to manipulate people for different purposes [1]. In social networks, misinformation extends in seconds among thousands of people, so it is necessary to develop tools that help control the amount of false information in the web. Similar tasks are detection of popularity in social networks [2] and also detection of subjectivity of messages in this media [3].

A fake news detection system aims to help users detect and filter out potentially deceptive news. The prediction of intentionally misleading news is based on the analysis of truthful and fraudulent previously reviewed news, i.e., annotated corpora. In [4] the authors divide the approaches to fake news detection divide into three categories: knowledge-based (by relating to known facts), context-based (by analyzing news spread in social media), and style-based (by analyzing writing style). In all cases, the approaches require annotated corpora, which implies an additional challenge given that the number of labeled fake news corpora is scarce and the few available resources are in the English language.

The main contributions of this research work can be summarized as follows:

*Corresponding author: Helena Gómez-Adorno Universidad Nacional Autónoma de México, Mexico City, Mexico. E-mail: helena.gomez@iimas.unam.mx

– the development of the first Spanish corpus consisting of fake and real news extracted from news websites. It is a new resource to investigate and analyze different aspects of the style-based fake news detection. This is the first resource for Spanish language in this research field.
– the annotation procedure for classifying real and fake news.
– statistics of the corpus, including vocabulary overlap of the different news topics and classes (real vs. fake). The same statistics is provided for other state-of-the-art fake news labeled corpus in English.
– experiments for automatic fake news detection using supervised learning on linguistically motivated features.

The rest of the paper is organized as follows. Section 2 provides a review of related work. Section 3 presents the methodology of the corpus construction, the annotation steps, the statistics of the corpus, and a comparison with other fake news resource. Section 4 describes a baseline approach for fake news detection and the obtained results. Finally, Section 5 draws the conclusions and points to the possible directions of future work.

## 2. Related Work

Although the publication and dissemination of fake news is not a new issue, nowadays their propagation is being potentiated by social media platforms. It is easy and fast to spread fake news in social media so its identification turns out to be a process with a certain degree of complexity. Therefore, the detection of fake news is attracting a lot of attention in recent years.

*Shu et al.* [1], presents an overview of research on the detection of fake news in social networks, focusing on psychology, social theories, and algorithms. Two aspects of the problem are reviewed: characterization and detection. Their fake news detection approach is divided into two stages: (i) extraction of characteristics and (ii) construction of the model. The feature extraction stage aims to represent news content and related information in a formal mathematical structure and the model construction stage builds an automatic learning model to differentiate between fake and real news.

Other works tackle the problem by analyzing the source of origin, for example, *Nazer et al.* [5] introduced a methodology for the detection of fake news related to natural disasters by means of social networks, specifically Twitter, analyzing the characteristics of the language of the tweets, as well as their variations throughout the period of time in which the disaster. This work analyzes the problem of news streaming because at the beginning of disaster situations the bombardment of information is such that it is difficult to keep track of all publications to determine their veracity. It also considers the dissemination of unwanted content such as spam, rumors, generic opinions and the use of bots, which has become a daily matter. The bots are capable of spreading large amounts of information in a short period of time.

There are three generally used characteristics of fake news: the text, the responses of the users to such news and the users that disperse the news. In [6], the authors introduced a model that integrates such characteristics for a more accurate prediction model. The model uses the text and users responses to train a Recurrent Neural Network to capture the temporal pattern of user activity on a given article. Then, the behavior of the users that disperse the news is learned and combined with the other characteristics in order to decide if the news is real or fake.

The Fake News Challenge project (FNC-1) [1] propose to break down the problem into stages. A first step would be to understand what several news organizations are writing about the topic, through a Stance Detection task. This task seeks to estimate the stance of the text in relation to the title (of the news), the text can agree, disagree, discuss or not be related to the title. So the goal of this stage is to develop tools to organize the news so that later people can analyze and quickly identify fake news [7,8].

Another interesting dataset [9] considers six class labels: pants-fire, false, barely true, half-true, mostly-true, and true. The data for this corpus are taken from the manually labeled news by Politifact [2]. The author evaluated four classification methods: regularized logistic regression (LR), support vector machine classifier (SVM), a bi-directional short-term memory network model (Bi-LSTMs), and a convolutional neural network model (CNN) (to integrate text and metadata). The hybrid CNN model outperformed all models, resulting in a precision of 0.270 in the test set.

[1] http://www.fakenewschallenge.org/

[2] https://www.politifact.com/truth-o-meter/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

As we can see, a considerable large amount of annotated corpora can be found in English for fake news detection. However, to the best of our knowledge, there is still no resource for building a machine-learning-based approach, which requires annotated corpora [10]. There have been huge advances on research work for the Spanish language, for example, the Spanish corpus manager [11] consists in a huge database of general purpose corpora in Spanish. Another initiative is the *Sociolinguistic Corpus of WhatsApp* [12] and the *Spanish language proverbs* [13]. Although there are already tools that focus on the detection of fake news in the Spanish language, most of them carry out the verification process manually, that is, through exhaustive investigations by the work team and users. This implies that the identification of deceptive content is subjective and delayed. An example of this type of platform is the *VerificadoMX*[3] site, which focuses on publications related to the political sphere.

## 3. The Spanish Fake News Corpus

The Spanish Fake News Corpus contains a collection of news compiled from several resources on the Web: established newspapers websites, media companies websites, special websites dedicated to validating fake news and websites designated by different journalists as sites that regularly publish fake news.

The news were collected from January to July of 2018 and all of them were written in Spanish. The resource is freely available at `https://github.com/jpposadas/FakeNewsCorpusSpanish`.

In contrast to other works [4,9], where a more detailed classification of the news is used, the presented corpus was tagged considering only two classes (true or fake). Although in some cases the news is not completely true or fake, there is not a convention on how to classify the news and some of the proposed categories in previous works are not clear enough.

The process was manually performed and the following aspects were considered: 1) news were tagged as true if there was evidence that it has been published in reliable sites, i.e., established newspaper websites or renowned journalists websites; 2) news were tagged as fake if there were news from reliable sites or specialized website in detection of deceptive content (for example VerificadoMX) that contradicts it or no other evidence was found about the news besides the source; 3) the correlation between the news was kept by collecting the true-fake news pair of an event; 4) we tried to find the source of the news.

Table 1 presents a list of websites considered as reliables. *Animal Político* and *Aritegui Noticias* are websites of news managed by a prestigious journalist that frequently appear in the media like newspapers, TV or radio. *Proceso* is a magazine focused on political and social themes that offers news on his site and *MVS Noticias* correspond to a company that broadcasts news on TV and radio. The rest of the sites mentioned in the Table 1 correspond to the digital version of established newspapers.

Table 1
Reliable sites

| Name | URL | Origin |
|---|---|---|
| **ABC** | `www.abc.es` | Spain |
| **Animal Político** | `www.animalpolitico.com` | Mexico |
| **Aristegui Noticias** | `aristeguinoticias.com` | Mexico |
| **BBC News** | `www.bbc.com/mundo` | England |
| **CNN Spanish** | `cnnespanol.cnn.com` | USA |
| **El Clarín** | `www.clarin.com` | Argentina |
| **El Espectador** | `www.elespectador.com` | Colombia |
| **El Financiero** | `www.elfinanciero.com.mx` | Mexico |
| **El Mundo** | `www.elmundo.es` | Spain |
| **El País** | `elpais.com` | Spain |
| **El Universal** | `www.eluniversal.com.mx` | Mexico |
| **Excelsior** | `www.excelsior.com.mx` | Mexico |
| **Forbes** | `www.forbes.com.mx` | USA |
| **Huffpost** | `www.huffingtonpost.com.mx` | USA |
| **La Jornada** | `www.jornada.com.mx` | Mexico |
| **La Vanguardia** | `www.lavanguardia.com` | Spain |
| **Marca** | `www.marca.com` | Spain |
| **Milenio** | `www.milenio.com` | Mexico |
| **MVS Noticias** | `www.mvsnoticias.com` | Mexico |
| **Proceso** | `www.proceso.com.mx` | Mexico |
| **Tiempo** | `www.tiempo.com.mx` | Mexico |

According to specialized websites in unmasking fake news, certain websites that systematically publish fake news have been detected. This type of deceptive websites contains news that are not true at all (completely fake, a mixture of true and fake, satire, humorist, among others). Deceptive websites usually combine true news along with fake news to confuse to the users. The site *VerificadoMX* mention some of this kind of websites.

[3] `https://verificado.mx/`

Websites that offer news validation service have appeared due to the rapid propagation of fake news. These websites detect and unmask some of the fake news that surf on the Web or social media to prevent the misinformation between users. Most of the validation websites perform the validation manually by a journalist. Some validation websites were used in the compilation of the proposed corpus, their names and descriptions are below.

- **VerificadoMX** (`verificado.mx`): it seeks to confront fake news and unreliable promises or unfounded criticism with a journalist approach.
- **Maldito Bulo** (`maldita.es/malditobulo`): it is an independent journalistic project dedicated to unmasking all kinds of rumors and fake news on social media like Facebook, WhatsApp or Twitter.
- **Caza Hoax** (`cazahoax.com`): it is a community of Hispanic origin specialized in unmasking false information on the internet.

### *3.1. Corpus Compilation Procedure*

The fake news was recollected from the validation sites and the deceiving sites and tagged manually. In the case of the validation sites the following steps were performed to gather a real-fake news pair:

**Step 1** Select a news from the validation site.
**Step 2** Look for the link to the source. If the link is missing, the news is discarded and go back to step 1.
**Step 3** Verify the link to the source. If the link is broken then the news is discarded an go back to step 1 else download and save the news as fake.
**Step 4** Identify the keywords of the news by answering the questions What?, Who?, How?, When? and Where?.
**Step 5** Use the Google search service to look for the real news counterpart of the fake news by typing its headline or keywords.
**Step 6** From the results of the search performed in the previous step, select the news whose origin belong to a reliable site (see table 1) and best match with the headline and keywords.
**Step 7** Download and save the selected news from the previous step as a true news.

Note that in the **Step 5** and **6** previously described, it is assumed that at least one true news will be found. This assumption is feasible because the validation site needs it to unmask the fake news.

For deceiving sites, the procedure to extract fake news changes from the previous one because the deceiving sites combine true news with fake and it is important for us to identify first the fake news. The steps to gather a real-fake news pair from deceiving websites are described next.

**Step 1** Select a piece of news from a deceiving website.
**Step 2** Identify the keywords of the news by answering the questions What?, Who?, How?, When? and Where?.
**Step 3** Use the Google[4], Yahoo[5] and Duckduckgo[6] searching services to look for news by typing its headline or keywords.
**Step 4** From the results of the previous step, compare the keywords of the links in the first three pages of results with the news selected in Step 1.
**Step 5** If there are at least three links from reliable sources that confirm the information contained in the selected news then it is considered as true else if at least three links deny the information contained in the selected news or if no results were obtained from the searching services then the news is considered as fake.

**Step 5.1** Identify the keywords of the fake news recently found.
**Step 5.2** Use the Google search service to look for the real news counterpart by typing its headline or keywords.
**Step 5.3** From the results of the search performed in the previous step, try to find the news whose origin belong to a reliable site (see table 1) and best match with the headline and keywords. If it exists, the news is considered true.

The main drawback we faced for the compilation of the corpus was the search for the real-fake news pair of an event. In some cases, it was difficult to track real news on the Web that complemented the false news.

### *3.2. Corpus Normalization*

To prevent that some elements of the news (for example numbers, emails, authors name, among others) act as markers for the classes, we perform a normalization for the corpus.

[4] `www.google.com.mx`
[5] `mx.search.yahoo.com`
[6] `duckduckgo.com`

The normalization process eliminates elements that are common in the structure of news and can be used as markers. These elements are the name of the author or the name of the editor, dates, any footer or header that references the source website.

We use the standard *utf-8* codification and keep the punctuation marks. Special characters were eliminated and references to photos or videos were not included in the corpus.

The normalization process also seeks the following elements and mask them by a common identifier: 1) numbers that represent quantities, schedules or prices were masked using the *NUMBER* tag; 2) email addresses of authors or editors were masked using the *EMAIL* tag; 3) URLs of references was masked using the *URL* tag; 4) telephone numbers were masked using the *PHONE* tag; 5) dollar and euro symbols were masked using the *DOL* and *EUR* tags respectively.

The title of the news is detected and is saved as an element apart. The URL of the source is also considered as an element apart from the text of the news.

### *3.3. Corpus Statistics*

The corpus covers news from 9 different topics: Science, Sport, Economy, Education, Entertainment, Politics, Health, Security, and Society. The Table 2 shows the distribution of the collected news along the different categories.

Table 2
Spanish Fake News Corpus topic distribution

| Category | true | fake |
|---|---|---|
| **Science** | 46 | 43 |
| **Sport** | 66 | 58 |
| **Economy** | 24 | 19 |
| **Education** | 10 | 12 |
| **Entertainment** | 70 | 78 |
| **Politics** | 175 | 148 |
| **Health** | 23 | 23 |
| **Security** | 17 | 25 |
| **Society** | 60 | 74 |
| **Totals** | **491** | **480** |

To identify the vocabulary overlap between true and fake news we considered the lemmas and discarded the stop words. The overlap was calculated by dividing the intersection of both true and fake vocabulary between the joint vocabulary. The general vocabulary overlap between real and fake news is 27.68%. Table 3 presents the vocabulary for true and fake news along with the vocabulary overlap between categories.

Table 3
Vocabulary overlap within each category

| Category | True | Fake | Overlap |
|---|---|---|---|
| **Science** | 2543 | 1837 | 24.96% |
| **Sport** | 2261 | 1804 | 25.23% |
| **Economy** | 1774 | 1011 | 17.60% |
| **Education** | 917 | 800 | 18.41% |
| **Entertainment** | 3164 | 2523 | 27.68% |
| **Politics** | 5383 | 3844 | 29.95% |
| **Health** | 1858 | 1318 | 22.95% |
| **Security** | 917 | 994 | 20.94% |
| **Society** | 3255 | 2778 | 25.01% |
| **Average overlap** | - | - | **23.63%** |

The general vocabulary overlap between categories is presented in the Table 4.

For some pairs of categories, the vocabulary overlap was expected because the categories have topics in common. Cases like the pair of Politics and Entertainment, where it was not expected to have a high overlap, can be explained as a temporary phenomenon caused by the desire of entertainment personalities to participate in political positions for electoral processes in some countries. Other cases can be explained by the methodology of mixing the content of news from different categories used by sites dedicated to spreading fake news.

## 4. Fake News Detection

The corpus was split into train and test sets, using around the 70% of the corpus for train and the rest for test. We performed a hierarchical distribution of the corpus, i.e., all the categories keep the 70%-30% ratio. This distribution is described in the Table 5.

We followed a machine learning approach for automatically identifying fake news. We evaluated three feature representations. One of them is the standard bag-of-words model which is simple baseline for evaluating if there is an specific selection of words that can help us to identify fake news. The other two representations are the character $n$-grams and POS tags $n$-

Table 4
Vocabulary overlap within each category

| Category | Science | Sport | Economy | Education | Entertainment | Politics | Health | Security | Society |
|---|---|---|---|---|---|---|---|---|---|
| **Science** | - | 22.65% | 24.13% | 17.33% | 23.91% | 24.07% | 24.06% | 18.09% | 26.64% |
| **Sport** | 22.65% | - | 22.68% | 17.28% | 25.63% | 23.59% | 20.63% | 19.18% | 24.56% |
| **Economy** | 24.13% | 22.68% | - | 18.49% | 19.85% | 19.89% | 20.60% | 19.20% | 22.30% |
| **Education** | 17.33% | 17.28% | 18.49% | - | 15.11% | 13.06% | 17.68% | 17.30% | 16.02% |
| **Entertainment** | 23.91% | 25.63% | 19.85% | 15.11% | - | 26.13% | 19.43% | 18.33% | 26.53% |
| **Politics** | 24.07% | 23.59% | 19.89% | 13.06% | 26.13% | - | 18.25% | 16.13% | 29.64% |
| **Health** | 24.06% | 20.63% | 20.60% | 17.68% | 19.43% | 18.25% | - | 17.76% | 22.62% |
| **Security** | 18.09% | 19.18% | 19.20% | 17.30% | 18.33% | 16.13% | 17.76% | - | 17.62% |
| **Society** | 26.64% | 24.56% | 22.30% | 16.02% | 26.53% | 29.64% | 22.62% | 17.62% | - |

Table 5
Corpus distribution

| Category | Train | | Test | |
|---|---|---|---|---|
| | **true** | **fake** | **true** | **fake** |
| **Science** | 32 | 30 | 14 | 13 |
| **Sport** | 45 | 41 | 21 | 17 |
| **Economy** | 18 | 12 | 6 | 7 |
| **Education** | 6 | 9 | 4 | 3 |
| **Entertainment** | 48 | 55 | 22 | 23 |
| **Politics** | 121 | 105 | 54 | 43 |
| **Health** | 16 | 16 | 7 | 7 |
| **Security** | 11 | 18 | 6 | 7 |
| **Society** | 41 | 52 | 19 | 22 |
| **Totals** | **338** | **338** | **153** | **142** |

grams representation, both of them probed to be helpful for representing writing style of authors [14,15,16]. We evaluated the performance of word and character $n$-grams type when including and excluding stop words. An standard preprocessing of the corpus was performed by eliminating the stop words and punctuation marks using the Spacy [7] tool. The performance of each of the feature sets was evaluated separately and in combinations.

We trained a classifier to generate a model that can distinguish between real and fake news. We experimented with four machine learning classifiers: support vector machine (SVM) with linear kernel, logistic regression (LR), random forest (RF), and boosting (BO). All of them are widely used in many natural language problems and have achieved state-of-the-art results, like in authorship attribution [17], sentiment analysis [18] and opinion mining [19], and plagiarism [20], author profiling [21], among others.

We conducted experiments trying to identify real and fake news in the corpus regardless the category of the news (binary scenario). We trained the classifiers using their scikit-learn [22] implementation on the previously mentioned features set: bag of words (BOW), POS tags, and $n$-grams features (with $n$ varying from 3 to 4).

We established for this corpus as a baseline the strategy of assigning all the news in the test corpus to one of the classes because the train corpus is balance for both real and fake classes. We choose the real class since there are more news that belong to that class in the test corpus.

Table 6 presents the accuracy obtained in the test set when we trained the classifiers on individual features set such as BOW and POS. We also evaluated the performance of the classifiers when combining those feature sets.

Table 6
Results on the test set in terms of accuracy (%)

| Features set | SVM | LR | RF | BO |
|---|---|---|---|---|
| **BOW** | 71.52 | 72.20 | 76.27 | 72.54 |
| **POS** | 68.13 | 67.11 | 63.72 | 61.01 |
| **BOW + POS** | 70.84 | 73.55 | **76.94** | 72.20 |
| **Majority Baseline** | | | **51.86** | |

Table 7 presents the evaluation of $n$-gram as feature representation for detecting fake news. It can be

[7] Freely available at `https://spacy.io/usage/models`

observed the results of the different classifiers when trained on character and POS $n$-grams with sizes from 3 to 5.

The best results on the test set were obtained with character 4-grams without removing the stop words with the Boosting algorithm. Note that the exclusion of stop words decreased the performance of the classifiers.

Table 8 presents the performance of the classifiers on the test set when they are trained on the combination of $n$-grams features. We first evaluated the combination of sizes of $n$-grams and then the combination of types of $n$-grams. The SVM algorithm achieved the best results when combining all types and sizes of $n$-grams.

Analyzing the results from the experiments, we can see that the models based on character $n$-grams in general achieved better results. The case when combining the representations of $n$-grams of size 3 to 5 achieved lower results than using individual feature sets.

The use of a machine learning approach improved the accuracy of the proposed baseline for the corpus. Traditional techniques often used to solve natural language problems obtain good results in the proposed corpus.

## 5. Conclusions and Future Work

The detection of fake news is an emerging research area that is gaining a lot of attention. The development of new resources such as annotated corpora can help to increase the performance of automatic methods aiming at detecting this kind of news. In this work, we presented the Spanish Fake News corpus, which to the best of our knowledge is the first corpus consisting in news extracted from the internet and labeled as real or fake.

We presented the development procedure, that can be used to further increase the size of the corpus. We show the statistics of the complete corpus, describing the vocabulary size in the different news topics, and the vocabulary overlap between real and fake news. We aimed at increasing the vocabulary overlap, thus ensuring the classification algorithm is truly identifying fake news and not only thematic areas.

Concerning the fake news detection methodology, we trained well-known classification algorithms on lexical features BOW, POS tags, $n$-grams (with $n$ varying from 3 to 5), and $n$-grams combination. The classification results show that is possible to achieve very high accuracy, and that the corpus is a valuable resource for building fake news detection models.


## Acknowledgements

The work was done with partial support of CONACYT project 240844 and SIP-IPN projects 20181849 and 20171813.


## References


[1] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, Fake News Detection on Social Media: A Data Mining Perspective, *ACM SIGKDD Explorations Newsletter* **19**(1) (2017), 22–36.

[2] A. Balali, M. Asadpour and H. Faili, A Supervised Method to Predict the Popularity of News Articles, *Computación y Sistemas* **21** (2017), 703–716, ISSN 1405-5546.

[3] R. Satapathy, I. Chaturvedi, E. Cambria, S.S. Ho and J.C. Na, Subjectivity Detection in Nuclear Energy Tweets, *Computación y Sistemas* **21** (2017), 657–664, ISSN 1405-5546.

[4] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff and B. Stein, A Stylometric Inquiry into Hyperpartisan and Fake News, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2018, pp. 231–240. http://aclweb.org/anthology/P18-1022.

[5] T.H. Nazer, G. Xue, Y. Ji and H. Liu, Intelligent Disaster Response via Social Media Analysis A Survey, *ACM SIGKDD Explorations Newsletter* **19**(1) (2017), 46–59.

[6] N. Ruchansky, S. Seo and Y. Liu, Csi: A Hybrid Deep Model for Fake News Detection, in: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ACM, 2017, pp. 797–806.

[7] W. Ferreira and A. Vlachos, Emergent: a Novel Data-Set for Stance Classification, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1163–1168.

[8] P. Krejzl, B. Hourová and J. Steinberger, Stance Detection in Online Discussions, *arXiv preprint arXiv:1701.00504* (2017).

[9] W.Y. Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, *arXiv preprint arXiv:1705.00648* (2017).

[10] G. Sierra, *Introducción a los corpus lingüísticos*, Instituto de Ingeniería, UNAM: México, 2017, p. 210.

[11] G. Sierra, J. Solórzano Soto and A. Curiel Díaz, GECO, un Gestor de Corpus colaborativo basado en web, *Linguamática* **9**(2) (2017), 57–72.

[12] A. Dorantes, G. Sierra, T.Y.D. Pérez, G. Bel-Enguix and M.J. Rosales, Sociolinguistic Corpus of WhatsApp Chats in Spanish among College Students, in: *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, 2018, pp. 1–6.

[13] J. Martínez, G. Bel Enguix and L. Torres Flores, Observations on Phonetic and Metrical Patterns in Spanish-language Proverbs, in: *Proceedings of EUROPHRAS 2017*, Tradulex, 2017, pp. 182–189.

Table 7

Results of fake news detection on the test set in terms of accuracy (%) when classifiers are trained on $n$-gram features sets.

| $n$-gram size | | | Character $n$-grams removing stop words | | | | Character $n$-grams including stop words | | | | POS $n$-grams | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3** | **4** | **5** | **SVM** | **LR** | **RF** | **BO** | **SVM** | **LR** | **RF** | **BO** | **SVM** | **LR** | **RF** | **BO** |
| ✓ | | | 68.81 | 67.11 | 69.83 | 70.16 | 72.54 | 71.18 | 74.57 | 75.25 | 62.03 | 62.37 | **67.79** | 60.67 |
| | ✓ | | 72.20 | 69.15 | 71.18 | 71.52 | 75.59 | 76.61 | 75.25 | **77.28** | 58.98 | 61.35 | 65.76 | 56.61 |
| | | ✓ | **73.55** | 70.16 | **73.55** | 73.22 | 75.59 | 75.93 | 76.27 | 76.27 | 62.03 | 59.66 | 67.11 | 57.62 |

Table 8

Results of fake news detection on the test set in terms of accuracy (%) when classifiers are trained on combinations $n$-gram sizes and types.

| $n$-gram size | | | Type of $n$-gram | | | Character $n$-grams removing stop words | | | |
|---|---|---|---|---|---|---|---|---|---|
| **3** | **4** | **5** | **Word** | **Character** | **POS** | **SVM** | **LR** | **RF** | **BO** |
| ✓ | ✓ | ✓ | ✓ | | | 51.18 | 53.89 | 50.84 | 58.98 |
| ✓ | ✓ | ✓ | | ✓ | | **73.22** | 66.10 | 71.18 | **73.22** |
| ✓ | ✓ | ✓ | | | ✓ | 60.33 | 60.00 | 64.74 | 61.01 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **73.22** | 64.06 | 70.84 | 72.88 |


[14] M.A. Sanchez-Perez, I. Markov, H. Gómez-Adorno and G. Sidorov, Comparison of Character N-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2017, pp. 145–151.

[15] H. Gómez-Adorno, C. Martín-del-Campo-Rodríguez, G. Sidorov, Y. Alemán, D. Vilariño and D. Pinto, Hierarchical Clustering Analysis: The Best-Performing Approach at PAN 2017 Author Clustering Task, in: *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2018, pp. 216–223.

[16] H. Gómez-Adorno, G. Ríos-Toledo, J.P. Posadas-Durán, G. Sidorov and G. Sierra, Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts, *Computación y Sistemas* **22**(1) (2018).

[17] E. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for information Science and Technology* **60**(3) (2009), 538–556.

[18] B. Pang, L. Lee et al., Opinion mining and sentiment analysis, *Foundations and Trends® in Information Retrieval* **2**(1–2) (2008), 1–135.

[19] G. Sidorov, S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Trevino and J. Gordon, Empirical study of machine learning based approach for opinion mining in tweets, in: *Mexican international conference on Artificial intelligence*, Springer, 2012, pp. 1–14.

[20] M.A. Sanchez-Perez, G. Sidorov and A.F. Gelbukh, A Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014., in: *CLEF (Working Notes)*, Citeseer, 2014, pp. 1004–1011.

[21] I. Markov, H. Gómez-Adorno, G. Sidorov and A. Gelbukh, The winning approach to cross-genre gender identification in russian at rusprofiling 2017, in: *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation*, 2017, pp. 1–216.

[22] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt and G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.