Structure and Experiments to Recognition and Classification of Fake News about Citizens in Brazilian Government Positions

Igor Baesse de Araujo¹; Yuri Rezende Mustifaga¹, Luiz Melk de Carvalho¹, Diva de Souza e Silva Rodrigues¹, Flávio Henrique Batista de Souza¹

¹Centro Universitário de Belo Horizonte – UNIBH, Brazil

Abstract.

The high flow of information generated an effect, where the news started to have characteristics that disprove their reliability index, generating a problem called fake news. Based on this, a validation methodology was created using NLP (Natural Language Processing) and SVM (Support Vector Machines) pattern recognition algorithms to investigate news with accuracy measured through the analysis of AUCs (Areas Under the Curve - varying between 0 and 1) and represented via Chatbot. With a methodology focused on experimentation, firstly, for the collection of information and news, a webtool was developed to gather and synthetize the analysis of the main websites that publish news in Brazil, such as web crawler. An indexing function was developed for all website addresses, correlated to a research factor (which in this work is Brazilian citizens in government positions). The distribution network of the web crawler resembles a graph, where the vertices are the ten websites most relevant to the public and the edges are the interconnections between the websites. For the recognition of fake news patterns, during the website analysis, a list of words for research is generated, and the SVM algorithm performs a non-linear separation between fake news and true news. To disseminate the results of the fake news analysis to citizens, a chatbot was developed. In this chatbot the user sends a snippet of the news or the web address of the news with some keywords and the tool will handle the processing of the request. The assertiveness measured by the AUC was between 0.96 and 0.99.

Keywords: Fake news, Natural Language Processing, Support Vector Machines, Chatbot, Citizens in government positions,

1. Introduction

With the increase in access and availability of information caused by the technological revolution in the last 20 years, in addition to the exponential growth in demand for internet access (ten times expected in March 2020) due to the pandemic scenario by COVID 19, they
brought some known problems, but in proportions never before imagined (Saeed et al. 2020; Favale et al. 2020). The ease of access and the exacerbated lack of interest in fact-finding, presented to citizens, has caused alarming misinformation (Zhou et al., 2020).

The rapid exchange of information on social networks and the Internet has created a very common term among people. “Fake News” as it is called, is news that does not represent reality. This false news has been a major problem for companies and entrepreneurs who created these social networks (Allen et al., 2020).

In a brief analysis of Fake News, it is associated with the ability of the human being (technology user) to recognize the pattern of information (in the case of whether it is true or not). According to Mahajan et Al. (2016), the human being has the ability to recognize patterns the moment they see something. They capture all the information in the environment to isolate, associate and identify the patterns around it. This ability that human beings have, such as recognizing faces, handwriting, voices, and even a person's mood is called pattern recognition.

According to Huang et al. (2020), it can be said that patterns are means by which the world is interpreted and from these interpretations a certain attitude or decision is taken. This feature is simulated by computers using pattern recognition algorithms, such as the Support Vector Machine (SVM) and the Multilayer Perceptron (MLP).

The pattern recognition process sometimes needs a metric that supports the assessment of the accuracy of the classification performed. Therefore, the ROC (Receiver Operating Characteristic) curve is, according to Fawcett (2006), an efficient method to analyze and diagnose systems, which is a great way to evaluate machine learning performance through the amount of information and possible results in the analyzes. As an interpretation of these data, the AUC (Under the Curve Area) is the region below the ROC curvature, making it possible to generate graphs and obtain a predictive assessment of how close the model is to the best possible result.

With these premises in mind, the objective of this work is to develop a structure and experiments that assist in the process of accessing true information. The specific objectives are to: develop a method of collecting data from sources with a high rate of consultation through digital media and news portals; develop a methodology, based on pattern recognition and accuracy analysis to assess behavior and classify the information analyzed; present a form of validation with probability of certainty about the verified information; present a chatbot-based structure to interact with the user. This work is justified by the international context, where unsubstantiated information is considered alarming by nations, especially in relation to the election of government officials.

2. Methods

2.1 Selection of Development Tools

2.1.1 Pattern Recognition Algorithms

As already mentioned, pattern recognition algorithms are computational structures that enable the machine to recognize and distinguish some type of behavioral pattern. In this work, two structures were used: MLP and SVM.
According to Lama & Mishra (2016), Figure 1a describes the Perceptron, which is the mathematical description to represent a human neuron, where the values from $x_1$ to $x_n$ are the information inputs. The nerve impulses of a human neuron is represented by the values of the inputs in a perceptron model. The values from $w_1$ to $w_n$ (Figure 1a) are the weights that these inputs will have in the analyzed function and that later can be adjusted to obtain better results in the neural network. As an analogy to the neuron, the weights are the mathematical representation of the human synapses. The visual representation $\Sigma$ will be the result of the sum of all the inputs that culminated in the transfer function, so if and only if the value obtained by the function is greater than the value stipulated for the activation of this neuron.

In the concept of Multilayer Perceptrons (Figure 1b), instead of using networks of single Perceptrons, there are $N$ interconnected layers of $N$ neurons, thus forming a multilayer network (Multilayer Perceptrons), enabling the resolution of several computational problems.

**Figure 1: Perceptron and Multilayer Perceptron**

![Perceptron and Multilayer Perceptron](image)

*Source: (Lama & Mishra, 2016; Lopez, 2018)*

The SVM Support vector machine (SVM) is a classification method, where data entries are separated into groups based on their characteristics. This technique has been shown to be quite efficient when it comes to analyzes where the plane is n-dimensional as shown in Figure 2.

**Figure 2: Multidimensional support vector machine**

![Multidimensional support vector machine](image)

*Source: (Yeh, 2015)*

The efficiency of such structures can be assessed, as already mentioned, by the accuracy of the ROC curve, which generates the dimensionless value, from 0 to 1, called the AUC. The closer to 1, the better the accuracy of the algorithm (Fawcett, 2006).
2.1.2 Web Information Collection Tools

Web Crawlers are robots that access and download information from web sites in an automated way. Generally, they are systems that browse the internet in search of images, texts, links to make more access in the future, among others. For this purpose, it is necessary to configure these systems so that they access only one domain in order to have more filtered data. After a query is made, the spiders process the total information of the website and returns to the programmer several data in text format (Hamborg et al., 2017).

The Web Scraping process is the acquisition of these unstructured data, transforming them into structured text files. Based on this process, the user can access HTML attributes more specifically to extract some information as shown in Figure 3.

![Figure 3: Web crawler and web scraping](source: authors)

2.1.3 Web Information Collection Tools

For the development of this work, a set of computational tools was used: React Native, Python, MongoDB and Scrapy. A brief description is given in the following topics:

- **React Native**: is a tool for creating native mobile apps for Android and iOS platforms using the same JavaScript code.

- **Python**: To search for a base of information from various news sites, a web crawler with a Python framework called Selenium was developed. This library has modules that allow HTTP requests (Hypertext Transfer Protocol) to be extracted from the HTML (Hypertext Markup Language) content of websites. The process that a crawler performs is called web crawling or spidering. At each visit, this crawler collects the maximum amount of information and stores it in an unstructured database. With the
aid of the BeautifulSoup library, it was possible to extract more relevant information from this large mass of data previously entered by this crawler.

- **MongoDB**: is a database with information in JSON format (JavaScript Object Notation) without the restrictions of a relational database.

- **Scrapy**: one of the most used tools for Web Crawlers to collect information after a previously ready mapping system. Scrapy basically works with so-called Spiders. With this tool, it is possible to configure simultaneous access, navigation system of links within links, among other features. The operation of Scrapy is based on 3 main flows: Spider, Scheduler and Middleware Downloaders.

### 2.2 Experimental Methods

This work proposed the study and analysis of news published in national media, seeking the probability of its veracity, based on correlated news and the availability of a free research tool. Based on a methodology focused on experimentation, methods were first developed to collect information for later analysis and, finally, a process of disseminating the analyzed information is implemented through a mobile application.

#### 2.2.1 Information Collecting Process

To collect all the information necessary for the main objective of the work, a tool known as web crawler tool (or simply crawler) was developed for synthetic analysis and collection of the websites of the main news in the country.

Functioning as an aggregator of information, the crawler has the feature of indexing all addresses related to a research factor, which in this project was based on Brazilian citizens with some occupation in public positions. The distribution network of the developed crawler resembles a graph where the vertices are the websites most relevant to the public and the edges are the interconnections between the websites. With the use of this methodology it is possible to perform the collection of several news items on the same research factor and thus correlate them using natural language and artificial intelligence processing in order to achieve the research result.

#### 2.2.2 Information Analysis

After collecting the information, each text collected was processed, using the technique of tokenization and lexical treatment, seeking the computational representation of what comes to be a news item. With this treatment and the use of the bag of words technique, for the analysis of similarity and correlation among the research factors, it was possible to extract a common denominator among the texts that exemplifies the probability that they are the same factor and contain information self-evident. For the probability of a factor to be relevant, the number of factors that highlight it must be as large as possible to the detriment of those that do not. After the survey of the texts referring to a research factor, a weight will be given based on the number of references exercised on where the sum of its value will be the sum of the references as shown in equation 1.

\[
weight = \sum n^{-1}
\]  

(1)
With this technique (and a text previously not referenced by the network of referential denominators) it is possible to measure the discrepancy that the text under analysis has from the reference text; analyzing the values obtained from the use of this technique for the known text and the text to be measured for veracity; thus obtaining a probability of the unknown news matching or not the news present in the news with greater relevance.

2.2.3 Sharing Information Process

To help bring information quickly and concisely to citizens, a chatbot was developed to facilitate communication between the user and the database raised by analyzing the information. The operation of the chatbot is based on the user sending a snippet of the news or the web address of the news with some keywords and the tool will take care of processing the request. As the largest number of access to fake news occurs through smartphones and social networks of the population, an application for phone devices was developed and released for Android and iOS systems, where with a simple registration the user has access to the entire search tool.

3 Results and Discussion

3.1 Experimental Structure

The experimentation structure consisted of algorithms for collecting and editing data to seek for a better adaptation of a news item in natural language for computational processing. The data editing process consists of 5 steps, which must be followed subsequently to achieve the established model.

Process 1 - Tokenization: The tokenization process aims to establish a computational data structure known as Array. This process is characterized as follows: after a text news is collected, it must be processed and converted into blocks of words, thus generating a more computational friendly structure for the processes subsequent to it.

Process 2 - Stop Word: The stop word treatment seeks to remove words that do not help the process of defining the intention of a given news item in the Portuguese language. Words that only exist to define the grammatical structure of the language such as: o, a, os, are used in the Portuguese language, but in an intention identification process they must be removed to avoid the generation of false positives when processing a particular item. Thus, the stop word process removes these words and leaves only those that assist the process.

Process 3 - Stemming: it is the process of removing features from the language that do not help in the process. Characteristics such as verb tenses, gender and plural are not part of the intention process when dealing with news analysis, for this purpose it is not extracted the word itself, but the core of its meaning, always looking for the basic derivation of the word, as shown in figure 4. For instance, the word "women" after the stemming process would derive "woman", because for the process the plural of the word does not matter, but only its basic meaning.

Process 4 - Frequency: The frequency process uses the training base composed of 3,500 news items, half of which are proven to be false while the other half are proven to be true. Based on that, two data structures are obtained, being composed of all the fake news in one structure and the other with all the real news. These data structures are necessary for the information
identification and processing steps and will later be used to define the shape that should be used for the processing of the established neural network.

Process 5 - Bag of Words: Bag of words is the last step in the news serialization process and it consists of converting all the news text into a matrix of zeros and ones based on the most frequent words contained in each category as shown in Figure 5 as example.

For each word contained in a news item, it is checked whether it exists in the most frequent words according to the category. If the word exists, it is inserted and a matrix referring to that “news=1” (if there is no, 0), at the end of the process a matrix of zeros and ones.

This matrix should be used as an input value in the process of training and analyzing the news. For the model used, matrices of 100 x 100 were defined, which means that the training shape used in the defined neural network is a matrix of 100 lines per 100 columns of 0 and 1. The model was defined using the news pattern observed in the collection process.

Figure 4: Stemming for the word "play"

![Figure 4: Stemming for the word "play"](source)

Figure 5: News conversion process

![Figure 5: News conversion process](source)
3.2 Data Collection with Web Crawlers

A data collection structure was developed with the purpose to make the process dynamic and as close as possible to real time. The collection of news to be evaluated is done through the crawlers using the search system available on the listed sites. The process starts with the user referring a search term and through the developed platform, this search term is required on several sites. From this process it is possible to filter within the news sites, a certain term that is related to the search term required by the user within the system. With the news collected, the editing process begins and analyzes the HTML structure returned by these sites. At this point, the Scrapy tool is used with the aid of the BeautifulSoup library. Both tools enable greater agility in the processing of the information so that only the text that matches the news is extracted. After the extraction and processing of the data, the classification process begins using the pre-established neural network.

In order to index the information in a way that the algorithm can work, it is necessary that the information is delivered in an expected format. Thus, an analysis was carried out on the pages of the main information sites in Brazil, using pattern recognition. Each site has a unique pattern. When entering the evaluated website, as shown in Figure 6, it is possible to check some HTML tag patterns. However, not all sites present this same pattern, which makes it very difficult during the collection process. Each website needs a system for mapping information based on these tags. As shown in Figures 7a and 7b, each news site has a unique model of tag combinations to make the information available to society, making the collection of this information more difficult.
Translation of the highlighted news (from Portuguese to English): “Public official disguises himself as Federal Police and applies deception to smugglers”

Source: (authors)

Figure 7: HTML inspection comparison

Translation of the highlighted news (from Portuguese to English): “Self-proclaimed president, Juán Guaidó, called people to the streets against the regime and announced support from the military. Maduro’s government speaks in attempted coup, but says it has military loyalty”
Translation of the highlighted news (from Portuguese to English): Analysis: after first sign of military support to Guaidó, there is civil war on the horizon

Source: (authors)

3.3 Online Fake news Classifier

3.3.1 Tests Structure

A dataset developed by Monteiro et Al. (2018) was used to train and validate the pattern recognition algorithms. It has 3,500 records, 50% with fake news and 50% with real news. The focus is that the algorithms, already trained, come to classify each evaluated news online. For this to be possible, the classification algorithm based on SVM, needs that the data receive some treatment in relation to the captured content (after that, the structure of the neural network can work with computational data models).

The SVM model requires cataloged groups for distinction and classification to be possible. In the case of the developed algorithm, the classification of “true” or “false” was used to create this aggregation of content. Thus, it was possible to carry out the vectorization process of these data, which consists, in the first moment, in counting the number of most frequent words in each of the groups (thus obtaining a vector of more evident words in each type of content).

However, the vector extracted from this technique is not yet capable of being used in the subsequent classification process (MLP). This is due to the fact that it still contains complex structures given in text format.

In order to get around this situation, a conversion process, called bag of words, was necessary. This approach aims to convert each news item to a matrix of values 0 and 1 (a data structure that is computationally simpler to process).

This matrix is structured as follows: for each word contained in the news text, it is checked whether it exists in the most frequent word array, obtained through all elements. If it is true, the matrix for that news receives an element 1, otherwise the value for that position is 0. At the end of the processing, a matrix of values 0 and 1 will be obtained, which define the structure of the news in reference to its group. This resulting matrix is classified by the MLP.

3.3.2 Online Experiments

It was observed during the experiments that the size of the news, which circulates in the analyzed portals, is similar to that found in the dataset used. Based on this similarity and after
testing the algorithm implementation, it was defined that the dimensions of the matrix to be used in the neural network for data input (input layer) would be 100x100. With this structured data model, the definition of MLP layers was: two hidden layers, each with 8 Perceptrons, using the softmax algorithm to activate the non-linear regression. Thus, the assertiveness rates of the employed methodology were analyzed, based on AUC (measured, as already mentioned, from 0 to 1).

As shown in figure 8a, the assertiveness rate obtained with the test base reached 99%, which is to be expected when using the pre-processed test base and suitable for the test model. With this rate, obtaining results from news not previously known by the training model, it is necessary to evaluate and calculate its coefficient of variation in order to obtain the rate of distortion among these elements.

In addition to the assertiveness rate, referring to the test basis, the loss rate can also be obtained (Figure 8b). This rate refers to items, where completion and assertiveness are at intervals that are not considered satisfactory for the decision, by one option or another. The rate of loss is directly linked to the rate of assertiveness, which means that the less losses the model obtains the better its rate of assertiveness and consequently its performance.

![Figure 8: AUC x Text input](source: authors)

3.3.3 Iteration Structure Via Chatbot

A chatbot was developed to facilitate the interaction of the end user through the platform, as shown in figure 9.

With it, the user can interact with the algorithm so that the result of the data processed from an initial news is visualized.

The user informs the news title for the chatbot and then begins the actual processing with the algorithm.

In a second step, while the algorithm performs a scan (via crawler) and the validations as mentioned in section 3.2 the user is informed that his request is being processed.

At the end of the interaction, the user is informed of the veracity of this information based on the algorithm presented via Chatbot.
Translation from Portuguese to English:

White (Bot): Hello, for our conversation to be efficient, let me know the news link or the title, so that I can search for information about it.

Green (user): Hello, this news is true Agnelo Queiroz, ex-governor of DF, becomes defendant for tax crime

White (Bot): We will be processing the information in the next few minutes, the news accessible through the link http://g1.globo.com/df/distrito-federal/noticia/2019/06/10/agnelo-queiroz-ex-governador-do-df-vira-reu-por-pedalada-fiscal.ghtml. As soon as we have more information we will get back to you.

Green (user): Any feedback?

White (Bot): Sorry for the delay, we got some information about the news of your interest. It was referenced on 7 other sites containing a level of similarity between the sources, but we still didn’t get enough references to generate assertiveness about its content. Wait a little longer and we will return with more information.

Source: (authors)

4 Conclusion

According to what was discussed in this paper, considerable complexity was identified in the automatic collection of information made available on the main news sites in the country, taking into account that each site has a specific grammatical structure and the HTML structure of each site is structured in different forms among them. The variety of structures makes the process extremely manual to constantly keep the information collectors working properly.

The way news gets to Brazilian society today is increasingly fast and constant, so the availability of a tool along the lines that society is used to is extremely relevant, and access to the developed tool must be accessible as much as the news that it seeks to analyze.
Thus, the developed chatbot helped iterate with the user in the process of searching for the veracity of the news by the population, making the process faster and less bureaucratic.

Even with the help of tools such as the one developed during this work, fact finding is still complex and requires constant analysis, in addition to an adaptive approach to the social context. No matter how many tools a society has to corroborate the truth of the facts, it is necessary to understand it and how to do it and why to do it. As demonstrated by this work, society in the information age is bombarded with information without even having the means to validate it and in many cases influenced by political or commercial bias.

References


