Development of a system for automatic classification of fake news in Portuguese language

Roger Oliveira Monteiro[1] , Rodrigo Ramos Nogueira [2]
Centro Universitário Leonardo da Vinci[1] , Universidade de Coimbra
Indaial – SC - Brazil[1], Coimbra - Portugal
roger.o.monteiro@gmail.com, rodrigonogueira@dei.uc.pt

Abstract: With the rapid advancement of technology and easy access and dissemination of information, the term fake news has gained commanding attention, and research in several areas has been developed. This paper introduces the use of machine learning methods to discover, classify and store fake news texts for later ETL application of a Data Warehouse and a query environment that will contribute to future research. For this, a dataset was created and the Logistic Regression, Naive Bayes and SVM methods were evaluated. Finalizing the work with the selection of the best method that was inserted in an automatic classification system of fake news.

Key words: Machine Learning, Text Mining, Web Mining.

1 Introduction

In 1990, Tim Bernes-Lee created the well-known WWW, which allowed the manipulation of the graphical interface guaranteeing its expansion in the following years through the emergence and development of browsers, access providers and various portals of services.

According to a report by the International Telecommunication Union (ITU), a member of the UN, in 2018, the percentage of people connected to the Internet exceeded 3.9 million people, representing 51.9% of the world population.

Brazil's 2018 data show that the percentage of people over 10 years old connected online in Brazil rose from 64.7% to 69.8% from the end of 2016 to the end of 2017. As a result, there was an increase of almost 10 million users within one year.

At the beginning of the twenty-first century there emerged an enormous tendency that follow until the present day: the use of social networks. More and more people are connected to each other through these platforms. Currently we have as examples of great cases of success of social networks: Facebook, Instagram, WhatsApp and Twitter. According to the Statistician Facebook had 608 thousand people connected at the end of 2010. At the beginning of 2019, the number of users reached 2,375 million, as shown in the chart below:

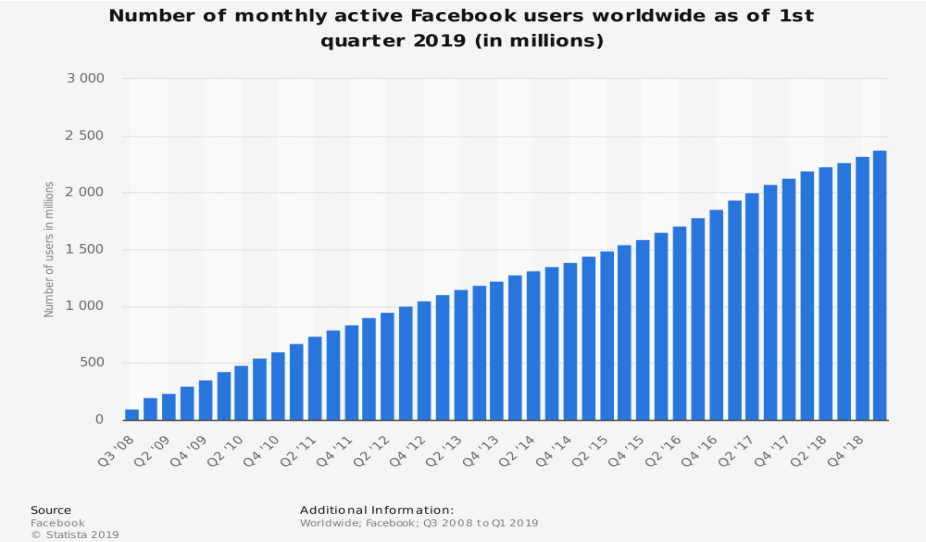

Figure 1 Facebook Users

Since the beginning of the Web, the volume of data that is in the repositories in the world network has grown exponentially, there are currently about 200 million active websites on the Internet, of which only the social network Twitter generates, on average, 500 million of posts per day. Such an explosion of data led to a study by the IDC (Institute Data Corporation) that estimates in the year 2020 we will have 44 data zettabytes generated in worldwide [1].

In the different niches of social networks that have emerged, different ways of writing critiques, propitiated by the characteristics of the applications, have been observed. Specific sites, such as movie reviews, allow users to write relatively long texts. Microblogs, on the other hand, impose limits on the amount of message characters and are not exclusively intended for critical publishing. In the process of discovery and research that continued in social networks, the need to express opinions more directly arose [2].

New sites are the third largest vehicle of information accessed by the Internet, second only to messaging applications and social networks. This information reflects the importance of using news sites and their impact on people's daily lives [3]. Along with the importance of news stories and their sharing in social networks, comes the rise and spread of fake news. Since the middle of 2017, the number of events and debates about this phenomenon that has been called fake news has grown so. Fake news can be defined as news articles that are intentionally and verbatim fake and can fool readers. In this definition of fake news includes intentionally fabricated news articles, such as a widely shared article from the now defunct website denverguardian.com with the headline "FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide" Hillary's e-mail leak found dead in apparent murder-suicide) [4].

Faced with the ease with which anyone can now access information, and with the ease of its use, we experience an era of great advances and solutions, followed by even greater problems, such as fake news. Because of its attractive nature, fake news spread rapidly, influencing people's behavior on a wide range of subjects, ranging from healthy issues (eg revealing miracle drugs) to politics and economics (as in the recent Cambridge Analytica scandal / Facebook and Brexit situation) [5]. Given its prominence, several multidisciplinary studies on the subject have been carried out. Aiming to contribute to such researches, this paper aims to couple to the ETL (Extract, Transform, Load) of a News Data Warehouse the semantic enrichment through classification of the type of news: real or false.

2 Related Works

About fake news and the application of Machine Learning, [6] constructed a dataset with news, in Portuguese and English, aiming to construct a classifier to predict whether the news source is reliable. Using an SVM algorithm with a linear kernel, it was possible to establish the most important characteristics as well as their classification. As a result, the classification algorithm obtained an accuracy of 85% for Brazilian datasets and 72% for American datasets. In a contribution to the news classification area, [5] use the dataset Fake.br with the objective of evaluating the main methods of preprocessing texts to evaluate the performance of the SVM method. The best results were obtained with the combination of bag-of-words with feelings, as well as the use of all attributes, both with accuracy of 90%.

On your work, [7] has collected news from truthful news sites and websites with fake news and / or satirical news, in order to find the best method for detecting fake news. As part of the pre-processing of the data, we used the Gensim framework to remove non-alphabetic characters, substitute spacing and line breaks for single spaces, remove words shorter than 3 characters, and convert capital letters to lowercase letters. The keras framework for data tokenization was also used. With the application of the classification algorithms LSTM and SVM, an accuracy was obtained above 90%.

Overfitting is a major problem in the case of a textual database. Therefore, FENG [8], used the algorithm AdaBoost, known to obtain great success for reduction of overfitting in face detection, character recognition (OCR) and vehicle classification. In his experiments, we used datasets from 20 newsgroups, Reuters dataset, which consists of 22 files with a total of 21,758 documents, and a BioMed dataset, which is divided into 10 topics, each containing between 1966 and 5022 news articles. The results were an average of 86% accuracy in the AdaBoost algorithm (Bonzaiboost).

3 System Architecture

After searching for a database with fake news, we found that there are few resources available in the Portuguese language, in which the most used dataset is Fake.br [5]. This present proposal aims to provide an environment with consistent and clean data in the form of a multidimensional corpus for consumption by external applications and users. The multidimensional corpus is a set of texts stored according to a multidimensional model, which allows to explore the multidimensionality in different levels of abstraction: time, category of news, type (true or fake news).

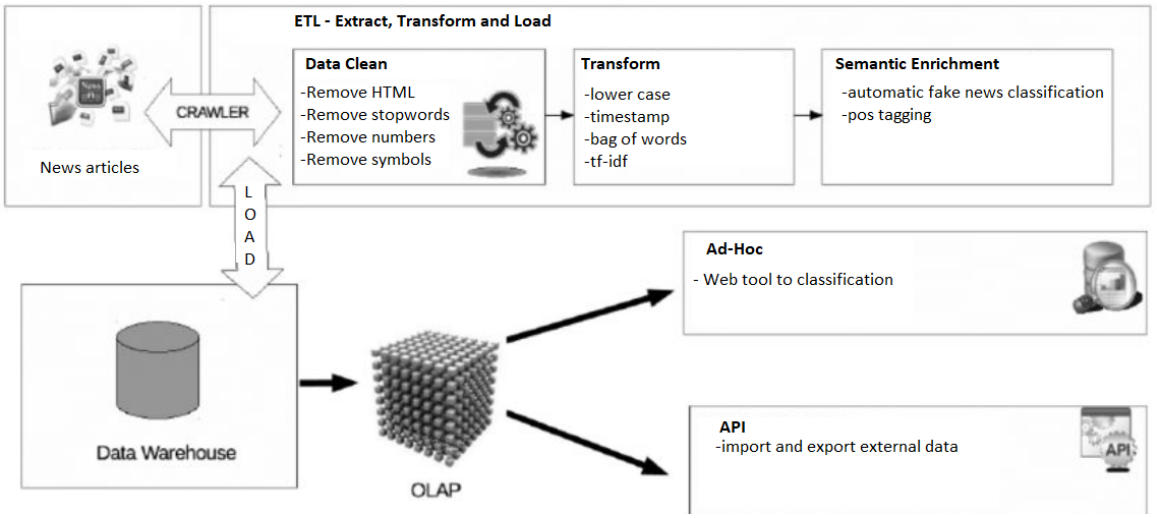

Figure 2 System Architecture

To perform the experiments, we developed a web crawler, using a python language, Including a library, beautiful soup for an initial data collection. A set of data was constructed by 1744 titles and fake corporations collected from the sites <boatos.org> and <g1.globo.com/fato-ou-fake>, and 3185 titles and body of real news collected from the site brasil.elpais.com. Initially will be tested the tests applied only to news, afternoon and body titles with the title and make a comparison between both. For this, the algorithms of machine learning, Logistic Regression, AdaBoost, Naive Bayes and SVM are used.

From the creation of a collection system, with an algorithm coupled to the ETL step, it will automatically classify the collected data, thus increasing the accuracy of the classifier, and generating a larger data base for future work to combat fake news. Also, a web interface was built, where the user will be able to submit a link and verify if this is true news, serving as a prototype before being submitted to ETL stage (this being the general purpose of this work).

4 Results and Discussions

The data obtained received treatment of null values, noise (special characters such as commas, periods, parentheses, etc.) and transformation to lowercase letters. Each dataset received a new column, called label, where Boolean value 0 was given for true news, and 1 for fake news. With this, the data was combined into a single dataset. Initially, the dataset used only contained the news headlines, being then divided between training and test, in the proportion of 75% and 25% respectively.

The first portion serves to train the algorithm, while the second, to verify the accuracy of the algorithm. Afterwards, they received tokenization treatment, using the NLTK package, with the bag of words in Brazilian Portuguese. Tests performed using the algorithms Logistic Regression, AdaBoost, Naive Bayes and SVM (linear kernel), obtained the accuracy of 88.85%, 81.37%, 86.22% and 87.45%, respectively, in the model of tests. As a technique for evaluating the models used, cross-validation was used with the k-fold = 10 method.

Again, the dataset was divided between training and test, joining now the headings to the news body. They received the same treatment mentioned above, obtaining the accuracy of 90.88%, 84.23%, 91.19% and 91.16% in the Logistic Regression, AdaBoost, Naive Bayes and SVM algorithms, respectively. The application of the cross-validation method revealed an overfitting in some cases. Finally, the dataset was broken down to use only the news bodies.

The same methods used previously were used in relation to the treatment and cleaning of the data. The application of the algorithms resulted in 90.88%, 94.23%, 91.19% and 91.16% accuracy in the algorithms Logistic Regression, AdaBoost, Naive Bayes and SVM respectively.

Table 1 – Experiments Results

	Regressão Logística	AdaBoost	Naive Bayes	SVM (kernel Linear)
Título	88,85%	81,37%	86,22%	87,45%
K-fold	0,88	0,75	0,86	0,55
Corpo	97,40%	95,12%	97,80%	98,62%
K-fold	0,97	0,95	0,97	0,64
Título + Corpo	90,88%	84,23%	91,19%	91,16%
K-fold	0,90	0,84	0,91	0,54

From the results analysis, the Naive Bayes method was selected as the best method, due to its high accuracy, complemented by being an incremental learning method (online). After the coupling, the fake news classification interface was developed, the Figure 2 shows the system interface and is available on the server https://detectorfakenews.herokuapp.com/. The tool expects as a parameter the link of a news site, and returns whether or not it is a fake news (fake news).

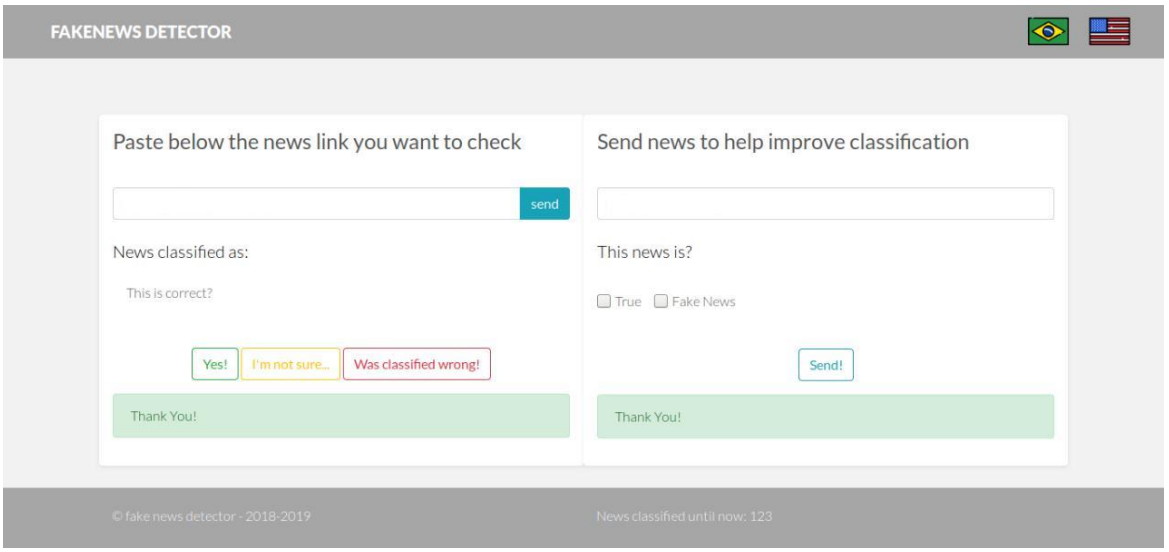

Figure 2 Fake News system classification interface.

5 Conclusion

Fake news has existed since the beginning of human civilization and has always generated problems in our society. This fake news spread slower in the past, today we have the Internet that makes it easier to share this news to the world in less than a minute. This news causes a great evil to our society, generating deaths and traumas.

People are accused of crimes they never committed. There are innumerable cases of people being charged without having committed at least one crime, and in many cases these charges have been made without any basis, thus making the victims of such charges innocent. These false accusations can end up causing damage, aggression, trauma, death and even suicide. People also fail to take vaccines and make medical treatment, in addition to the policy.

In this paper we developed a system for automatic classification of fake news, which obtained a result of 97% efficiency. It still becomes necessary to test with larger databases, as well as to put the system to be validated in the real world.

Overfitting is a recurring problem on a textual basis. Some algorithms have achieved very relevant results, but when we applied cross-validation with k = 10, we noticed a large overfitting in some cases. Thus, it was observed that the Naive Bayes algorithm obtained besides the high accuracy, tolerance to overfitting.

References

[1] IDC. Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012), 1-16.

[2] Von Lochter, J. (2015). Máquinas de classificação para detectar polaridade de mensagens de texto em redes sociais.

[3] Nogueira, R. R. (2017). Newsminer: um sistema de data warehouse baseado em texto de notícias.

[4] Delmazo, Caroline; VALENTE, Jonas CL. Fake news nas redes sociais online: propagação e reações à desinformação em busca de cliques. Media & Jornalismo, v. 18, n. 32, p. 155-169, 2018.

[5] MONTEIRO, Rafael A.; SANTOS, Roney L. S.; PARDO, Thiago A. S.; ALMEIDA, Tiago A. de; RUIZ, Evandro E. S.; VALE, Oto A.. "Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results." In: International Conference on Computational Processing of the Portuguese Language. Springer, Cham, 2018. p. 324-334.

[6] GRUPPI, Maurício; HORNE, Benjamin D.; ADALI, Sibel. "An Exploration of Unreliable News Classification in Brazil and The U.S." Rensselaer Polytechnic Institute, Troy, New York, USA.2018.

[7] MARUMO, Fabiano Shiiti. "Deep Learning para classificação de Fake News por sumarização de texto." - Londrina, 2018.

[8] FENG, Xiaoyue; LIANG, Yanchun; SHI, Xiaohu; XU, Dong; WANG, Xu; GUAN, Renchu. "Overfitting Reduction of Text Classification Based on AdaBELM", 2017