THE USE OF DATA AUGMENTATION AS A TECHNIQUE FOR IMPROVING NEURAL NETWORK ACCURACY IN DETECTING FAKE NEWS ABOUT COVID-19

Wilton O. Júnior, Mauricio S. da Cruz, André Brasil Vieira Wyzykowski, Arnaldo Bispo de Jesus
Instituto de Informática – Universidade Católica do Salvador (UCSAL)
Salvador - Bahia - Brazil
{wilton.junior, mauricio.cruz}@ucsal.edu.br

ABSTRACT
This paper aims to present how the application of Natural Language Processing (NLP) and data augmentation techniques can improve the performance of a neural network for better detection of fake news in the Portuguese language. Fake news is one of the main controversies during the growth of the internet in the last decade. Verifying what is fact and what is false has proven to be a difficult task, while the dissemination of false news is much faster, which leads to the need for the creation of tools that, automated, assist in the process of verification of what is fact and what is false. In order to bring a solution, an experiment was developed with neural network using news, real and fake, which were never seen by artificial intelligence (AI). There was a significant performance in the news classification after the application of the mentioned techniques.

Keywords Fake News · COVID-19 · AI · Data Augmentation · NLP

1 Introduction
The COVID-19 pandemic emerged within a context in which people are informed more by headlines on social media like Facebook than by the content of the newspaper article, opening up gaps for the phenomenon of fake news. The discovery of a new virus demands news dealing with the symptoms, severity, and forms of transmission. Furthermore, as everything is new and volatile, it is a fertile ground for the insertion of disinformation, especially in a scenario in which the habit of checking the veracity of information does not happen often. The production of fake news happens at a speed that is impossible to be checked by a small group of journalists, and these professionals must also cover the news coming from reliable sources. Given this context, the automation of this activity seems to be the only effective way to combat fake news through Artificial Intelligence (AI).

It is worth mentioning that the fight against fake news has a huge importance and responsibility of determining something as being false or true. Given that, there is the ability to innovate in producing new frauds. That’s why works like this and others mentioned in this article have a lot of relevance in this pandemic context.

Several authors such as [1], [2] and [3] have already proposed algorithms, many using modern AI techniques such as Neural Processing Language (NLP) or multi-layered Perceptron (MLP). Most of this work is related to contenders promoted on the Kaggle platform[1]. Some authors like [4] and [2] use the term infodemic to describe this phenomenon of fake news about the coronavirus.

Some social networks, like Twitter and Facebook, have released tools to combat fake news. In the Facebook’s case, a series of authentications to ensure that the person responsible for the post is the real owner of the account [5] and Twitter’s Birdwatch[2], a tool that allows the user to report posts as fake news. However, there are several reasons to

1Kaggle: an online community of data scientists and machine learning professionals subsidiary of Google.
2Source: https://tinyurl.com/ferramentamessenger
suspect the interest of these corporations in combating this practice since, when verifying the source, the user leaves the platform, in contrast with the business model of a social network that focus on engagement.

2 Background

With the advance of the COVID-19 pandemic, the discussion of whether access to internet is one of the fundamental rights for survival (as well as water, electricity, and basic sanitation) is no longer controversial and has become a consensus, given the fact that the dynamics of modern society demands that, for a citizen to be included, he must be connected. [6], talks about how the democratization of the internet is fundamental for the maintenance of the basic right to education provided in the Brazilian constitution of 1988. According to the Collins Dictionary, lockdown is the word of the year for 2020, considering that the dictionary recorded around 250 thousand uses against only 4 thousand the previous year [7].

For many years, the media's credibility, so be it print, radio or television, was practically irrefutable. Today, with the democratization of access, and especially the production of information and content, new vehicles, websites, blogs, YouTube channels, and other platforms have also started to act as press [8]. [8] goes on to say that, even after the fake news is denied, it still influences the population, as it attacks not only the facts but also the credibility of important institutions such as press vehicles. In addition, social networks have a tremendous impact on the way people think and show themselves as a powerful weapon of manipulation, leading political groups to create fake news or spread half-truths in order to create or influence a movement.

In 2020, some authors analyzed the impact and danger of fake news during the pandemic. Kalsnes cites in his work that the main reasons behind fake news include political, financial, and social issues [8]. [9] Shows how a series of fake news can be at the service of an ideology such as the privatization of the Brazilian Unified Health System (Sistema Único de Saúde - SUS) since these false articles try to discredit an organ of such tradition and importance.

Still, in his analysis [9], says that fake news in the Brazilian context about COVID-19 is divided into five categories. They are information related to health authorities, therapy, prevention measures, disease prognosis, and vaccination. These are the five main points, and their primary vehicles are social networks, especially WhatsApp.

The power of fake news has been extremely underestimated in the past. From the 2016 elections in the United States to the elections in Brazil in 2018 and now during the coronavirus pandemic, where President Bolsonaro accused press vehicles like Globo Television, and the Folha de SP newspaper in 2018 [10]. This behavior is similar to that of the President of the United States, Donald Trump, whom several times classified information coming from CNN and New York Times as fake news [8].

Efforts to combat fake news have proved quite challenging for traditional methods, as producing fake news takes less time and effort than debunking it. One of these efforts is Lupa, a project carried out by the Folha de São Paulo newspaper, which is a Fact-Checking agency, the first in Brazil [11]. However, it has been observed that there is a pattern in fake news created by the context of the pandemic, and through it, it's possible to automate this validation.

3 Related Works

Several researchers worldwide have been dedicated to creating solutions for the detection of fake news. However, there is still a lot of debate regarding the best architecture for this purpose. [12] He did a work comparing several techniques of modern artificial intelligence such as Convolutional Neural Network (CNN), Artificial Neural Network (ANN), and Recurrent Neural Network (RNN), and the accuracy varied a lot according to the dataset and the preprocessing the data before training.

The fact that it is a classification, it is necessary to portray the similarity of the data through a vector representation. This consists of analyzing the curve between the expected results with the results obtained and thus obtaining the accuracy and the LOSS, that is, how much is lost, of a training.

[12] in his work, he made a comparison between the techniques Word2Vet, One-Hot Encoding, Doc2Vec and TF-IDF, used text vectorization, used based on the CNN, ANN and RNN architectures. He concluded that the data do not vary much according to the architecture, but it does trough the dataset, which when using the TF-IDF technique with the dataset Kaggle were obtained the best results in the range of 0.96.

In 2021, [13] participated in the TUDublin team at Constraint@AAAI2021 - COVID-19 Fake News Detection, a competition where researchers with the same dataset seek to obtain the best possible accuracy. In 2021, the theme was precisely the pandemic of the new coronavirus. At the end of the project, they obtained between 0.91 and 0.95 accuracy, which are encouraging results given the urgency of this category of work.
For this, the researchers analyzed the most frequent terms using TF-IDF and removed the most common affixes, morphological and inflectional endings to keep only the central idea of the word with the PorterStemmer algorithm and, finally, changed the entire text to lowercase. After pre-processing, the model used the following architecture: logistic regression, SVM [footnote 1], and a combination of Naive Bayes [footnote 2] with logistic regression.

In another related work, researchers at Accenture [footnote 3] created the Cross-SEAN, which is based on the analysis of messages coming from social media, especially Twitter. For this, they divided the process into four stages. The first was to separate the tweets related to the coronavirus, based on the posts available in the dataset Kaggle, the IDs of these tweets were taken and searched in the API [footnote 4]. Then they collected reliable information about the pandemic using existing fact-checking tools (Snopes, PolitiFact, FactCheck, and TruthOrFiction). From there, they extract URL [footnote 5] with topics related to COVID-19, classifying the tweets between genuine and fake regarding the pandemic using text manipulation techniques such as BERT [footnote 6] and ROBERTa [footnote 7]. The last step performed was manually validating a sample of the results obtained. After the entire process, accuracy of 92% was found among the false posts detected automatically.

Despite excellent initiatives, we miss in all these works their applicability in the news, since it is through them that public opinion is based, given the fact of trusting the credibility of the press agency.

4 Materials and methods

This research was adapted based on a work written by another author [footnote 8]. His repository can be found in the references of this article. Furthermore, we propose to increase the accuracy of this algorithm by applying data augmentation techniques to texts, in addition to translating the dataset in order to try to bring the research application to the national territory due to the scarcity of datasets to be used on the proposed topic.

Although there are good works such as Fake Br Corpus [footnote 9] and [footnote 10] in Brazilian Portuguese for fake news. They do not address the context of the new coronavirus pandemic. There is a gap between the works available in Portuguese and the information on COVID-19, and this ends up becoming a challenge in this area. The only alternative found was LATAM Coronavirus [footnote 11], a constantly updated database used to check, validate and explain information about the new coronavirus. However, this database is not well-structured for news analysis. All the code is available on Google Colab [footnote 12] The link is in the references of this article, as well as the link to GitHub [footnote 13] with the author’s project, which we use as a base.

5 Technologies used

In this topic, we will cover all the technologies applied during the development process of this work.

5.1 Python

Launched by Guido Van Rossum in 1991, Python is a high-level programming language used for object-oriented programming (OOP) and structured programming. The language was designed with the philosophy of emphasizing the importance of programmer effort over computational effort.

There are several reasons why data scientists use Python. These need to create data visualizations to communicate results and predictions at any business level. With that in mind, the Python language has a great advantage because it contains libraries, framework [footnote 14], and exclusive packages to be used in the data area, such as sci-kit learn for Machine Learning, Numpy, Pandas for data analysis, among others [?].

---

3 Logarithm Regression: a statistical technique that aims to produce a model that allows the prediction of values taken by a categorical variable.
4 Naive Bayes: family of simple "probabilistic classifiers" based on the application of the Bayes’ theorem.
5 API: set of routines and programming patterns for accessing Twitter’s own software or web-based platform.
6 URL: virtual address of a page or website.
7 BERT model: a technique used for pre-training and natural language processing developed by Google for machine learning.
8 RoBERTa: pre-training model based on highly optimized BERT.
9 Google Colab: Collaboratory or "Colab" allows you to write Python code in your browser without any necessary configuration and with free access to GPUs.
10 GitHub: hosting platform for source code and files that allows any user registered on the platform to contribute to private and Open Source projects.
11 Framework: a set of ready-made codes that can be used in application development.
5.2 Dataset

Dataset is the term used for a collection of tabulated information (data), where each column represents a variable, and the particular value of each row corresponds to a dataset in question.

datasets can contain information, such as medical records or insurance records, for use by a program running on the system. These are also used to store information needed by applications or the operating system itself, such as source programs and macro libraries or system variables or parameters. These can be cataloged, which allows the dataset to be referred to by name without specifying where it is stored.

5.3 Natural Language Processing (NLP)

It is a subfield of Artificial Intelligence that gives the computer the ability to understand, analyze, manipulate and reproduce human speech/writing with considerable accuracy. It is also used to study and understand natural human language in its entirety.

Currently, Google’s search engine uses NLP to find and return results similar to the keywords entered the search field. In addition, it uses NLP in Google Translate, making the translation from one language to another much more efficient, not letting the text lose the sense of context.

5.4 Tokenization

Tokenization is the technique used to break the sequence of characters in a text by locating the limit of each word, separating each term of a sentence into tokens, a term used for computational purposes. Tokenization can normalize a text, for example, by mapping words to lowercase-only versions, expanding contractions, and even extracting the root of each word.

5.5 SpaCy

SpaCy is an open-source library used in the Python programming language for NLP (Natural Language Processing). This library already has pre-built models that act as ‘trained brains’ for each language, making it easy to use for various purposes.

In addition to being used in NLP, this library contains processing tools POS-Tagging (Part-Of-Speech Tagging (or understood as “Grammar Analysis”) for detecting adjectives, nouns, among others. Grammatical classes in the text, as well as being used as NER (Named Entity Recognition) to determine if a term refers to a person, date, place, organization, among others, in order to improve the identification accuracy of correct terms.

5.6 PyDictionary

PyDictionary is a library for getting meanings, translations, synonyms, and antonyms. It uses WordNet for meanings, Google for translations, and synonym.com for synonyms and antonyms.

5.7 Neural Network

These are computational models inspired by neurons in the human brain. With proper training, they can recognize patterns and correlations in raw data, be able to group and classify them, and over time, be able to learn continuously. Its first appearance was in 1943 by Warren McCulloch and Walter Pitts. From an article created on how neurons worked, they modeled a simple neural network using electronic circuits.

5.8 Data Augmentation

The data augmentation technique consists of manipulating the dataset in order to create similar data for training in order to increase the amount of data, adding slightly modified copies of existing data or newly created synthetic data from existing data. However, for a data augmentation to be efficient, the manipulation cannot mischaracterize the dataset. In other words: a recommendation from the World Health Organization (WHO) must still be true and correspond with the context of COVID-19.

12Macro: sequences of programmed events (such as keystrokes, clicks, and delays) that assist with repetitive tasks.
6 Development

The critical point of this research is to improve a set of pre-formatted data in order to increase the amount of information contained in it through the data augmentation technique so that, after training, the neural network has a percentage of accuracy superior to the previous one, increasing its effectiveness. However, there is still another factor that we need to consider: the dataset chosen is in the English language. That means the whole approach would have to be done by obeying the grammatical rules of the language. There was an attempt to search for a dataset in the native language (Portuguese). However, Brazil lacks a set of pre-defined data to be freely analyzed in this context. Because of this, the choice of the foreign dataset was the fastest and closest option found for the research development.

In total, 04 datasets were separated: one for the real training (Real Train), one containing real news that will be used for testing (Real Test), one for the training fake (Fake Train) and the last one containing fake news (Fake Test), without having to be used in that order, where all datasets “Train” has 800 news and datasets “Test” has 200.

The challenge was to translate all the dataset into Portuguese in order to keep the entire grammatical structure intelligible so that the neural network could maintain accuracy or have the least possible loss because the validation will occur with national (real or false), translation from one language to another can significantly impact the accuracy of neural network analysis. For this, the library Google Translate was used to do the translation properly. The translation process took 36 hours to complete. This factor was aggravated because the chosen library did not translate texts that contained more than 5,000 (five thousand) characters. Which led to the following specific instructions:

1. Create a loop to loop through the entire dataset
2. Check if the current text is longer than 5000 characters
3. If yes, we translate the text sentence by sentence, detecting through a period (.) in the text.
4. If not, the library translates all unprocessed text.

```python
def translate_text(word):
    translator = google_translator()
    tword = translator.translate(word, lang_src='en', lang_tgt='pt')
    time.sleep(1)
    return tword
```

Listing 1: Source code in Python. Source: The Author

After the translation process carried out by the function shown in Listing 1, it was necessary to visualize the most common words in real or fake news, based on the datasets used, to discover any pattern that differentiated them. For that, the library WordCloud was used, efficient to visualize the most used words in both situations. A word cloud was generated for the dataset translated “Real Train” and another word cloud for the dataset translated “Fake Train”. The datasets “Test” are for validation purposes only.

![Image](image1.png)

(a) Most common words in real news  
(b) Most common words in fake news

Figure 1: Word clouds after processing.
Source: The author

Figure 1(a) indicates the most common words in real news based on the dataset “Real Train”. However, it shows a critical flaw: when translating the dataset, the word cloud appears to have a lack of words which, considering the size of the chosen dataset, does not reflect the ideal amount of words, while the word cloud shown in Figure 1 (b), from
"Fake Train", has a much higher consistency of words. It is noticed that, when the texts are translated, the words with neutral contexts of the Portuguese language tend to stand out as the most used words in real news. In the case of dataset fake, the words in common tend to have more diversities, such as proper names, countries, and political contexts. In short, what can be highlighted is:

1. The word cloud of the dataset “Real Train” indicates that the most common words are those with neutral contexts ("to", "that", "as", "more than"...)

2. The word cloud of the dataset “Fake Train” already covers contextualized keywords, such as “Bill Gates”, “Donald Trump”, “Bats”, “human cells”, among other words.

However, it is necessary to have in mind that the word cloud of the dataset “Real Train” maintains a false representation of the dataset used. With the size of both datasets, the figures would need to present several words as close to each other as possible. However, Figure 1 (a) shows a huge lack of words compared to Figure 1 (b).

Figure 2: Process of data augmentation and translation of dataset.
Source: The Author

Figure 3: Example of text translated without data augmentation and translated with the technique applied.
Source: The Author

With that in mind, a process was built, as shown in Figure 2, where data augmentation is applied even before the translation takes place. With this process, both datasets would be much more abundant and very contextualized. To apply this technique, two libraries will be used: SpaCy and PyDictionary, to apply the NLP and search for synonyms of the chosen words, respectively. This example can also be seen in Figure 3, which shows how a sentence would look if translated in its entirety and how the same sentence would look if we applied data augmentation first.

To not misdirects the news, we applied the first rule that all words exchanged would be nouns. This is because verbs, pronouns, and adjectives can quickly go out of context if synonyms replace them.

The second rule is that the noun is only changed if at least 40 percent similar to the chosen synonym. This number was taken based on the parameters of the SpaCy library, known as the “similarity index”. Although this number may
If the value is too high, the data augmentation will not be robust, so it will have no impact during training.

For the first rule, the POS-Tagging (Part Of Speech Tagging) technique was used, as shown in the example of Figure 4, where the library had already a pre-trained model to be used, separating each word into a token to be parsed in a unitary way. The POS-Tagging technique identifies which grammatical class that word belongs to (pronoun, noun, verb, adjective, among others). Only words in the “noun” class will be identified.

For the second rule, we apply the PyDictionary library to search for synonyms for the tokenized noun, as shown in Figure 4. As this library can return a list containing several nouns, we use the similarity index from the SpaCy’s library depending on the word to identify if it has a similarity above 40 percent to the token, replacing it with the synonym with the highest percentage if so. There is also the possibility that a noun has no synonyms. In this case, we continue without changing the word.

This process is repeated for each news item in a dataset. In all, the process took about 60 hours to complete. After completing the data augmentation on the datasets “Train” and the translation process, it was generated a new word cloud of both datasets (Real and Fake) to see if there has been any significant change in the most common words for each. In both Figures 5 (a) and (b), it is possible to identify that there was significant growth in the most common terms presented in the word cloud of the datasets. It is also possible to notice that Figure 5 (b) in particular, word cloud from “Fake Train”, is a little more crowded with words than Figure 5 (a). So, it can be indicated that the data augmentation process impacted significant changes in both datasets.

6.1 Modeling and Training

An architecture based on the BERT classification model was used for the training phase. This model has been widely used for text classification because of its success in several architectural categories. According to BERT’s creators [28] this model is pre-trained on a large body of text and then fine-tuned for use in specific tasks.

---

13 Token: Word of a text separated in an independent unit.
14 SpaCy: spacy.io
For this work, in addition to BERT, Dropout\textsuperscript{15} and Dense Layer\textsuperscript{16} with five layers. All training takes place in 10 epochs\textsuperscript{17} and three epochs of patience. Patience is an indicator referring to the number of epochs with no improvement in accuracy before training stops. If there is no improvement in one season and the next one has a negative performance, the training stops. This prevents LOSS from growing and maintains accuracy at its best.

![Neural network neurons architecture](image)

**Figure 6:** Neural network neurons architecture.

Source: The Author

Figure 6 represents how the structure of neurons is designed. As the BERT algorithm can only accept sentences of up to 512 words, it is necessary to preprocess the data (long news) to feed the algorithm. The author follows this idea and segments each part of the text into multiple subtexts of a maximum of 150 words. Subtexts will be overlapped where, specifically, the last 30 words of the first subtext will be the first 30 words of the second subtext.

Then the original text is preprocessed into input resources that BERT can read. The process consists of tokenizing and converting the original text into token IDs, so the algorithm can read them. Words are tokenized based on the vocabulary dictionary they were pre-trained on (about 30,000 words), and unfamiliar words are broken down into smaller words contained in the dictionary. The maximum string length is also specified to fill all strings with the same length\textsuperscript{18}.

After all the texts are prepared, we move on to training the neural network using the translated and augmented datasets. The training structure made by the author used five epochs. However, as the datasets had their size almost doubled, the training was changed to 10 epochs so that there is no loss of pro efficiency in case of a short training time limit.

| BEFORE DATA AUGMENTATION | AFTER DATA AUGMENTATION |
|--------------------------|-------------------------|
| VALIDATION LOSS          | 0.3638                  | 0.3297                  |
| VALIDATION ACCURACY      | 0.9487                  | 0.9172                  |
| GENERAL ACCURACY         | 0.9516                  | 0.9913                  |
| LOSS                     | 0.3611                  | 0.0293                  |

Table 1: Comparative table indicating the values before and after applying the technique.

Source: The Author

The data plotted in Table 1 indicate that there was a significant increase in training accuracy and a decrease in the loss percentage, also showing a significant drop in validation accuracy, one of the indicators that correspond to real data compared to the predictions.

\textsuperscript{15} Dropout techniques were used: Technique to reduce complex co-adaptations of neurons, forcing a neuron to learn more robust features that are useful in conjunction with many random subsets of other neurons.

\textsuperscript{16} Dense Layer: It is a layer in which each input neuron is connected to the output neuron, and the parameter units only inform the dimension of your output.

\textsuperscript{17} Training epoch: refers to a training cycle with the entire dataset. Typically, training a neural network takes more than a few epochs.

\textsuperscript{18} The final string length would be longer than specified, as the BERT tokenizer will split unknown words into several known small words.
6.2 Validation and Results

After training, precision was used as the primary metric for model validation, in addition to comparing values prior to the data augmentation process, to analyze whether there was any change in three essential parameters: “Training Accuracy”, “Accuracy of Validation,” and “Accuracy of Test”, where each item represents the precision of each step (“training”, “test” and “validation”).

| BEFORE DATA AUGMENTATION | AFTER DATA AUGMENTATION |
|--------------------------|-------------------------|
| TRAINING ACCURACY        | 0.9596157073974609      | 0.9944672584533691 |
| VALIDATION ACCURACY      | 0.954023003578186       | 0.9166180491447449 |
| TEST ACCURACY            | 0.9438775777816772      | 0.9254571199417114 |

Table 2: Comparative table indicating the values before and after applying the technique

Source: The Author

It can be seen in Table 2 that there was a drop in the values of “Validation Accuracy” and “Test Accuracy” after the application of data augmentation. Despite the considerable increase in the “Training Accuracy” indicator, the drop in other indicators may be associated with the translation of the news into Portuguese. In any case, the post data augmentation values remained firmly above 91 percent.

For validation and in order to verify hit rates in the news that the neural network has never seen before, a case study was approached with the following metrics:

1. Analyze 20 real news and 20 fake news, both taken from the datasets “Real Test” and “Fake Test”.
2. Analyze 20 real news and 20 fake news, both taken from genuine sites and posts in Brazilian territory.

First validation: in the test with the dataset “Fake Test”, the neural network got 95 percent correct (19 out of 20 news), in addition to significant improvements in respect of accuracy. In the test with the news real, the neural network obtained 95% accuracy (19 of the 20 news), showing a significant improvement in accuracy values.

Second validation: in the test with real news, the neural network was 95% correct (19 out of 20 news), showing good stability in accuracy between the news. In the test with fake news, the neural network was 70 percent correct (14 out of 20 news). Looking at the last test of this validation step, it was decided to generate a word cloud with the most common words to understand why the neural network made a mistake in certain news.

Figure 7: Most common words in the news are misclassified.
Source: The Author

In Figure 7, it is possible to notice that words like “Pfizer”, “chip” and “Microsoft” appear as the most common keywords among the sentences where the neural network applied a wrong classification. Because, in the dataset, there is no consistent information containing these words (the closest to “Microsoft” would be “Bill Gates”, but the neural network has not been trained to relate terms). As the dataset is of foreign origin, it is possible that, in other countries, fake news is created with some different characteristics from Brazilians ones, and adapting to the Brazilian context is a real challenge.
It is also possible to explain the case of the first validation: both analyzed news came from datasets that were also translated. This may explain the percentage difference between the two validation steps. In addition, even before the translation of the datasets the AI was already classifying false positives in tests, information that can be considered through the percentages of accuracy shown in Tables 1 and 2.

All this data can be consulted on GitHub.\(^{19}\)

7 Future works

In the course of this work, some functions emerged that could help in the AI training process, as well as in the final classification accuracy:

1. Correlational analysis of words: as it was possible to notice in the second stage of validation of new classification, the neural network could not relate the word “Microsoft” with “Bill Gates” because it was not trained to do so. This practice can be implemented during the data augmentation process as a way of looking for relationships between words in the same context (not just by synonyms) and building an alternative dataset so that the neural network can classify news with greater accuracy.

2. Addition of Brazilian news to datasets: real or fake, adding more messages in the Brazilian context, not only translating foreign news, can have a significant impact on the neural network during the classification of national news.

3. Classification of the type of fake news: implement a structure where the neural network can identify beyond just real or fake news, being able to sub classify fake news as misleading, partially false, unsustainable (when there is no statistical evidence to prove it, for example), and other kinds of frauds.

8 Conclusions

As the tests demonstrate, the data augmentation was effective in increasing the accuracy of the analyses. It was possible to detect that the complete translation of a dataset into another language implies a loss of precision when applied in the context of the country. However, with the translation combined with the data augmentation technique, it was possible to use the neural network to analyze true and false news in the Brazilian language, presenting a very significant percentage of accuracy.

On the other hand, it is understandable to note that there was a considerable drop in detecting fake news in the Brazilian language, which the neural network never had contact with concerning the news in the dataset. This is because all the training was based on a translated dataset, where the sentences and contexts, even efficiently translated, generally did not correspond to real situations in Brazil.

However, this research project proved to be possible, improving existing functionalities, and implementing new functions in an already developed project, presenting a significant increase in the margin of precision. The entire project is available on GitHub.\(^{20}\)

References

[1] Youness Madani, Mohammed Erritali, and Belaid Bouikhalene. Using artificial intelligence techniques for detecting covid-19 epidemic fake news in moroccan tweets. *Results in Physics*, page 104266, 2021.

[2] Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tannoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. *arXiv preprint arXiv:2011.03327*, 2020.

[3] Pakpoom Mookdarsanit and Lawankorn Mookdarsanit. The covid-19 fake news detection in thai social texts. *Bulletin of Electrical Engineering and Informatics*, 10(2):988–998, 2021.

[4] Kaize Ding, Kai Shu, Yichuan Li, Amrita Bhattacharjee, and Huan Liu. Challenges in combating covid-19 infodemic–data, tools, and ethics. *arXiv preprint arXiv:2005.13691*, 2020.

\(^{19}\)Validation test: https://github.com/mauriciosena/tccdataaugmentation/tree/main/Tests

\(^{20}\)Project GitHub: https://github.com/mauriciosena/
[5] CanalTech. "facebook messenger ganha novas ferramentas para combater assédio e fake news". https://tinyurl.com/ferramentatwitterfacebook 2021. Accessed: 2021-05-23.

[6] Thaís Farias de Almeida. A pandemia de covid-19: Reflexos na garantia do direito à educação. Pensar Acadêmico, 18(5):881–894, 2020.

[7] Collins. Collins dictionary 2020. https://www.collinsdictionary.com/woty 2020. Accessed: 2021-05-15.

[8] Bente Kalsnes. Fake news. In Oxford Research Encyclopedia of Communication. 2018.

[9] Mercedes Neto, Tatiana de Oliveira Gomes, Fernando Rocha Porto, Ricardo de Mattos Russo Rafael, Mary Helhem Silva Fonseca, and Julia Nascimento. Fake news no cenário da pandemia de covid-19. Cogitare Enfermagem, 25, 2020.

[10] Matheus Ribeiro Pereira. A desinformação como estratégia política: uma análise dos tweets de ataque à imprensa postados por jair messias bolsonaro no ano de 2019. Aquila, (24):97–110, 2021.

[11] Folha. Lupa, a primeira agência fact-checking do brasil. https://piaui.folha.uol.com.br/lupa/?utm_source=Search&utm_medium= CPC&utm_campaign=LupaAnos 2021. Accessed: 2021-05-16.

[12] Lovedeep Singh. Fake news detection: a comparison between available deep learning techniques in vector space. In 2020 IEEE 4th Conference on Information & Communication Technology (CICT), pages 1–4. IEEE, 2020.

[13] Elena Shushkevich and John Cardiff. Tudublin team at constraint@ aaai2021–covid19 fake news detection. arXiv preprint arXiv:2101.05701, 2021.

[14] William Scott Paka, Rachit Bansal, Abhay Kaushik, Shubhashis Sengupta, and Tanmoy Chakraborty. Cross-sean: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. Applied Soft Computing, page 107393, 2021.

[15] singularity014. Singularity014 - autor do projeto base, 2021. Disponível em: https://github.com/singularity014/BERT_FakeNews_Detection_Challenge.

[16] Roney Santos. Fake.br corpus. https://github.com/roneysco/Fake.br-Corpus 2021. Accessed: 2021-05-19.

[17] Cristiano Ruschel Marques Dias. Towards fake news detection in portuguese: New dataset and a claim-based approach for automated detection. 2019.

[18] Chequeado. Latam coronavírus. https://chequeado.com/latamcoronavirusportugues/ 2021. Accessed: 2021-05-22.

[19] PYTHON. Python, 2021.

[20] IBM. Dataset, 2021. Disponível em: https://www.ibm.com/docs/en/zos-basic-skills?topic=more-what-is-data-set.

[21] NLP. Nlp, 2021. Disponível em: https://www.nlp.com/what-is-nlp/.

[22] Google. Google research, 2021. Disponível em: https://research.google/teams/language/.

[23] NLTK. Tokenize, 2021. Disponível em: http://nltk.sourceforge.net/doc/pt-br/tokenize.html.

[24] PyPi. Spacy, 2021. Disponível em: https://pypi.org/project/SpaCy/.

[25] PyPi. Pydiction, 2021. Disponível em: https://pypi.org/project/PyDictionary/.

[26] Stanford. História da rede neural, 2021. Disponível em: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/history/history1.html.

[27] TensorFlow. Tensorflow, 2021. Disponível em: https://www.tensorflow.org/tutorials/images/data_augmentation.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.