# Brazilian Portuguese Hate Speech Classification using BERTimbau

**Félix Leonel Vasconcelos da Silva, Larissa A. de Freitas**
Federal University of Pelotas
Pelotas, Brazil
{flvdsilva,larissa}@inf.ufpel.edu.br

### Abstract

Hate speech is a language that attacks or denigrates a specific group based on their characteristics, such as their race, ethnicity, or sexual orientation. Hate speech became widespread and spread through social networks, blogs, videos, and other communication channels. With anonymity and a sense of impunity, people feel encouraged to spread their hatred on the internet. In this work, we used the BERT model for the Portuguese language called BERTimbau to classify hate speech in three datasets in Portuguese, available in the literature: OFFCOMBR-2, OFFCOMBR-3, and Fortuna et. al. (2019) dataset. Still, we performed some preprocessing and an oversampling technique on the datasets. Finally, we compared the results obtained with results obtained by works available in the literature. Experiments with BERTimbau, using preprocessing and oversampling obtained better results than other classification techniques.

## Introduction

Hate speech is a language that attacks or denigrates a specific group based on their race, ethnicity, religion, sex, age, or sexual orientation (Nobata et al. 2016). According to Politize (2020), hate speech is a type of verbal violence, and its basis is the non-acceptance of differences, that is, intolerance.

Hate speech became widespread and started to occur in social networks, blogs, videos, and other communication channels. With the anonymity and a sense of impunity, people feel encouraged to spread all sorts of offensive, and discriminatory comments on the internet (Politize 2020).

In 2020, following the assassination of George Floyd, a campaign led by prominent civil rights groups and non-profit organizations called "Stop Hate for Profit"[1], pushed brands to suspend paid Facebook ads[2] until Facebook took action to curb disinformation and counter hate speech. After the adhesion of large companies such as Coca-Cola and Unilever, Facebook started to take action to counter hate speech (Alright 2021).



[1]https://stophateforprofit.org

[2]https://pt.facebook.com

Classifying hate speech is a problem can be treated using linguistic rules, machine learning, and deep learning. In this work, we use deep learning (Alshalan and Al-Khalifa 2020).

Deep Learning (DL) is a type of machine learning that trains computers to perform tasks like human beings, including speech recognition, image identification, etc. (Data Science Academy 2021). The Transformer is a DL approach introduced in 2017 that uses the self-attention mechanism (Data Science Academy 2021). BERT (Bidirectional Encoder Representations from Transformers) is a pre-training methodology for Transformers, but it is also the name of the models pre-trained by this methodology (Data Science Academy 2021). It has reached the state of the art when applied to different Natural Language Processing (NLP) tasks (Data Science Academy 2021). In this work, we use BERT model (model for Portuguese language called BERTimbau) in the hate speech classification task.

The contribution of this work is to verify if BERTimbau is a good approach in hate speech classification in specific datasets and if there is improvement in the results with different preprocessing and oversampling configurations.

This paper is structured as follows: Section "Related Works" presents the related works; Section "Methodology" describes the methodology; Section "Analysis of the Results" shows the analysis of the results; and Section "Conclusions and Future Works" presents the conclusions and future works.

## Related Works

In the literature, we find some works about hate speech classification in the Portuguese language, they are: (de Pelle and Moreira 2017), (Fortuna et al. 2019), (Silva and Roman 2020) and (Leite et al. 2020).

de Pelle and Moreira (2017) created two datasets, OFFCOMBR-2 and OFFCOMBR-3. The source of the datasets was the site G1[3]. The authors selected 1250 comments. OFFCOMBR-2 contains 1250 comments using a majority vote to determine if the comments are classified as hate speech or not. And, OFFCOMBR-3 has only the comments that all judges agreed on whether or not the comment was offensive. The comments are converted to lowercase and tested comments in their original form in the experiments. Also,

[3]https://g1.globo.com/

the authors used n-grams (unigrams; unigrams and bigrams; unigrams, bigrams, and trigrams) as features and Support Vector Machine (SVM) and Naïve Bayes (NB) as classifiers with 10-fold-cross-validation. The results obtained were F-measure (average between the two datasets) of 0.80 for the SVM and 0.75 for the NB.

Fortuna et. al. (2019) created a dataset with tweets in Portuguese. The authors used a Twitter profile search API[4] to search for keywords and hashtags such as #dyke or #womensPlaceIsInTheKitchen collected between January and March 2017. The dataset contains 5668 tweets using a majority vote to determine if the tweets are classified as hate speech or not. Still, the dataset is divided into 90% of the data for training and 10% of the data for test. The authors used the Long Short-Term Memory (LSTM) as classifier, with cross-validation combined with holdout validation. Also, the authors converted tweets to lowercase and removed the stopwords and punctuation. The result obtained was micro-averaged F-measure of 0.78 for the LSTM.

Silva and Roman (2020) used the dataset and preprocessing created by Fortuna et. al. (2019). The dataset is divided into 90% of the data for training and 10% of the data for test. The authors used the NB, Logistic Regression (LR), SVM, and Multilayer Perceptron (MLP) as classifiers with 10-fold-cross-validation. The best result obtained was F-measure of 0.72 for the SVM with word-level BoW.

Leite et al. (2020) created a dataset called ToLD-BR. ToLD-BR contains tweets collected between July and August 2019 with the tool named GATE Cloud's Twitter Collector[5]. The dataset is divided into 80% of the data for training, 10% of the data for development, and 10% of the data for test. The authors used the AutoML model to build the model (BoW + AutoML), BERTimbau, and BERT Multilingual. For the BERT models they used the simpletransformers[6] library and default arguments for parameter tuning. The results obtained were micro-averaged F-Measure of 0.76 for BoW + AutoML, 0.75 for BERTimbau, and 0.76 for BERT Multilingual.

## Methodology

This section presents the description of the work developed to classify hate speech in Portuguese. Our work is composed of four main steps (Figure 1). Initially, the comments/tweets are pre-processed. After, some techniques of data augmentation are applied in each dataset. We use BERTimbau model and fine tuning in the hate speech classification task. Finally, the results are analyzed.

To perform the experiment, we use OFFCOMBR-2, OFFCOMBR-3 and (Fortuna et al. 2019) dataset.

OFFCOMBR-2 is composed of 1250 comments (agreement of 2 annotators, 419 offensive sentences, and 831 non-offensive sentences) and OFFCOMBR-3 is composed of 1033 comments (agreement of 3 annotators, 202 offensive sentences, and 831 non-offensive sentences). The Table 1 shows the examples of the OFFCOMBR-2 and OFFCOMBR-3.

Fortuna et. al. (2019) dataset is composed of 5668 tweets (agreement of 2 annotators, 1786 offensive sentences, and 3882 non-offensive sentences). The Table 2 shows the examples of the Fortuna et. al. (2019) dataset.

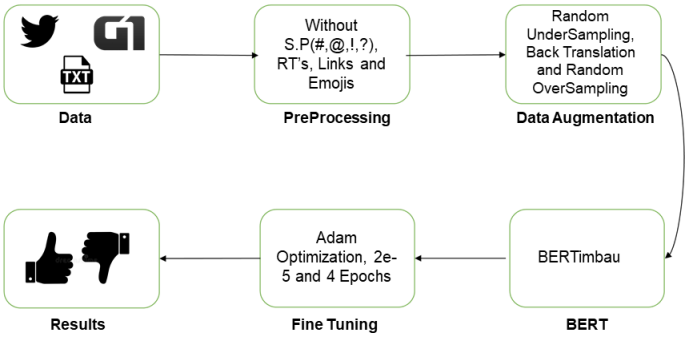


Figure 1: Methodology of this work.

Table 1: Examples of the OFFCOMBR-2 and OFFCOMBR-3.

| Comment | Hate Speech |
|---|---|
| "maria vai lavar uma louca que voce ganha mais" | Yes |
| "maria is going to wash some dishes that you earn more" | |
| "TODO ESPECIALISTA TEM QUE MENTIR PRA SE DAR BEM" | No |
| "EVERY SPECIALIST HAS TO LIE TO GET ALONG" | |

Table 2: Examples of the Fortuna et. al. (2019) dataset.

| Tweet | Hate Speech |
|---|---|
| "bom dia sapatao da minha vida" | Yes |
| "good morning dyke of my life" | |
| "E o sono... rs. Cheiros. Tb adorei!" | No |
| "And the sleep... lol. Smells. I loved it too!" | |

### Preprocessing

Preprocessing is an essential step in NLP, as it will determine the final quality of the data analyzed. It can even impact the prediction model generated from the data (Jurafsky and Martin 2009). In this work, we removed the special characters (for example: #, @, !, ?) in the three datasets. In the Fortuna et. al. (2019) dataset, we did preprocess usually done in tweets such as removing special characters, links, emojis, RT's, and hashtags.

### Data Augmentation

Data augmentation is a technique to generate new examples of training data to balance the datasets. There are some types

[4]https://developer.twitter.com/en/docs/twitter-api

[5]https://cloud.gate.ac.uk

[6]github.com/ThilinaRajapakse/simpletransformers

like Oversampling, Undersampling, Back Translation, Synonym Replacement, and others (Vladimir Lyashenko 2021). We used random undersampling, back translation, and random oversampling techniques in this work.

Undersampling is a technique to balance uneven datasets by keeping the data in the minority class and removing data from the majority class to equalize the data. Some types of undersampling include Random UnderSampling, Near-Miss UnderSampling, Condensed Nearest Neighbors UnderSampling, Tomek Links Undersampling, and others (Master's in Data Science 2021).

In this work, we used random undersampling to match the majority class (non-offensive) with the minority class (offensive) in the three datasets. Applying this technique, OFFCOMBR-2 contain 419 offensive and 419 non-offensive comments, OFFCOMBR-3 contain 202 offensive and 202 non-offensive comments, and Fortuna et. al. (2019) dataset contain 1786 offensive and 1786 non-offensive tweets.

Back Translation is the process of re-translating, data from the target language back to its source language in literal terms (Patrycja Jenkner 2020). In this work, we used English as a target language, for example: "Falei alguma mentira" (Portuguese) - "I said some lie" (English) - "Eu disse alguma mentira" (Portuguese).

Oversampling is a typical data analysis technique used to adjust the class distribution of data. Oversampling is done by applying a transformation to existing data instances to generate new data instances to modify the class imbalance (Won, Jap, and Bhasin 2020). There are some types of oversampling such as Random Oversampling, Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE, Adaptive Synthetic Sampling (ADASYN), and others[7].

In this work, we used random oversampling to equalize the number of comments and tweets of the minority class (offensive) with the majority class (non-offensive) in the three datasets. Applying this technique, OFFCOMBR-2 and OFFCOMBR-3 contain 831 offensive and 831 non-offensive comments, and Fortuna et. al. (2019) dataset contain 3882 offensive and 3882 non-offensive tweets.

### BERT

BERT is a Transformers pre-training method, but so is the name of the models pre-trained by this method (Devlin et al. 2019). BERT trains the language models based on the complete set of words in a query or phrase known as bidirectional training, making the language models able to discern the context of words based on the surrounding words rather than words that follow or precede it (Data Science Academy 2021).

Nowadays, many BERT models are available in the literature, such as BERT Base, which has 12 layers, and BERT Large, which has 24 layers. In this work, we use BERTimbau Base Cased[8], a pre-trained BERT model for Brazilian Portuguese.

### Fine Tuning

Fine Tuning means making small adjustments to a process to achieve the desired output or performance. In the case of DL, it involves using weights from a previous DL algorithm to program another similar DL process (Pratik Bhavsar 2019).

In this work, we use the Adam optimizer, the learning rate of 2e-5, batch size of 32, and 4 epochs. Where: (i) Adam optimizer is a first-order gradient-based algorithm of stochastic objective functions based on adaptive estimation of lower-order moments (Kingma and Ba 2014); (ii) learning rate indicates at what rate the weights are updated (Hafidz Zulkifli 2018); (iii) batch size is the number of training examples used in an iteration (Sagar Sharma 2017); (iv) epoch is how many full passes through the datasets should be used (Sagar Sharma 2017).

## Analysis of the Results

In this work the data was divided between 80% for training, 10% for validation and 10% for testing and cross-validation was not used.

In the experiments, we use eight configurations: (#1) Original, which is when we kept the sentences in their original format; (#2) No Special Characters (No S.C), which is when we removed special characters, links, emojis, RT's and hashtags; (#3) OverSampling, which is when we use the Random Oversampling technique to balance the classes of the datasets; (#4) OverSampling and No Special Characters, which is when we use the Random Oversampling technique and removed special characters, links, emojis, RT's and hashtags; (#5) Undersampling, which is when we use the Random Undersampling technique to balance the classes of the datasets; (#6) Undersampling and No Special Characters, which is when we use the Random Oversampling technique and removed special characters, links, emojis, RT's and hashtags; (#7) Back Translation, which is when the tweet or comment is translated into a different language and then translated back to its source language; (#8) Back Translation and No Special Characters, which is when the tweet or comment is translated into a different language and then translated back to its source language and removed special characters, links, emojis, RT's and hashtags.

We used three types of data augmentation, oversampling, undersampling and back translation, with the use of oversampling the datasets OFFCOMBR-2 and OFFCOMBR-3 have 1662 comments with 831 target as offensive and 831 target as non-offensive and the dataset Fortuna et al. (2019) has 7764 tweets with 3882 target as offensive and 3882 target as non-offensive, with the use of undersampling the dataset OFFCOMBR-2 have 838 comments with 419 target as offensive and 419 target as non-offensive and the dataset Fortuna et. al. (2019) has 3572 tweets with 1786 target as offensive and 1786 target as non-offensive, for the use of back translation we selected English as the target language.

We used the metrics Precision, Recall, Accuracy, Balanced Accuracy (Bacc), and F-Measure to evaluate the results. Precision indicates how many of the model's positive

[7]https://towardsdatascience.com/7-over-sampling-techniques-to-handle-imbalanced-data-ec51c8db349f

[8]https://huggingface.co/neuralmind/bert-base-portuguese-cased

classifications were correct. Recall indicates how many of the existing positive samples the model could classify. Accuracy is the number of hits divided by the total number of examples. Bacc is the calculation of all hits divided by all hits plus errors. F-measure is the harmonic mean between Precision and Recall (Müller and Guido 2018).

In this work, true positive occurs when the model classifies a comment or tweet correctly as offensive, false positive occurs when the model classifies a comment or tweet incorrectly as offensive, true negative occurs when the model classifies a comment or tweet correctly as not offensive, and false negative occurs when the model classifies a comment or tweet incorrectly as not offensive.

As can be seen in Table 3, for the Fortuna et. al. (2019) dataset, in the eight configurations in most cases there is a better performance of the model when the tweets are used in the original way they were written by the users, without removing the special characters, links, emojis, RT's and hashtags.

The examples below of the Fortuna et. al. (2019) were classified as true positive by configurations #1, #3, #5, #7 and classified as false negative by configurations #2, #4, #6, #8. Examples: "@user O cara mora em outro mundo Não no mundo real REFUGIADOS são os que moram nas favelas vizinhas às suas fortalezas" (Portuguese) - "@user The guy lives in another world Not in the real world REFUGEES are the ones who live In the slums neighboring their fortresses" (English), "@user bom dia sapatão da minha vida" (Portuguese) - "@user good morning dyke of my life" (English) and "@user Eu não me importo se os brancos, em Angola, foram assinados, quando a metrópole fez ainda pior com os nativos" (Portuguese) - "@user like I don't care if the whites, in Angola, were murdered, when the metropolis did even worse to the natives" (English).

In general, the best results by Fortuna et. al. (2019) dataset are obtained with configurations #1 and #3. This dataset obtaining 0.86 for Precision, Recall, Accuracy, F-Measure and 0.83 for Balanced Accuracy for configuration #1 and 0.86 for Recall and Accuracy for configuration #3. These configurations have 52 true positives with configuration #1 and 55 with configuration #3, 132 true negatives with configuration #1 and 34 with configuration #3, 16 false negatives with 16 configuration #1 and 13 with configuration #3, and 12 false positives with configuration #1 and 10 with configuration #3. Still, the worst results are obtained with configuration #8. Fortuna et. al. (2019) dataset obtaining 0.80 for Precision, Recall, Accuracy and F-Measure and 0.77 for Balanced Accuracy. This configuration has 48 true positives, 123 true negatives, 21 false negatives, and 20 false positives.

OFFCOMBR-2 was the dataset that had the best performance, reaching the best results, with the configuration #4 with 0.91 for Precision, Recall, Accuracy, and F-Measure and 0.90 for Balanced Accuracy and the worst results, with the configuration #5 with 0.85 for Precision and 0.84 for Recall, Accuracy, Balanced Accuracy, and F-Measure. In this dataset, it can be seen in Table 3, there is a difference between the use of oversampling and undersampling, the configuration #4 obtained 14 true positives, 29 true negatives, 2 false negatives, and 1 false positives and the configuration #5 obtained 8 true positives, 20 true negatives, 8 false negatives, and 10 false positives.

The examples below of OFFCOMBR-2 were classified as true negatives by configuration #4 and classified as false positives by configuration #5. Examples: "Uma visita aos familiares vitimados por esta escória seria louvável" (Portuguese) - "A visit to the family members victimized by this scum would be commendable" (English), "Usando o cálculo diferencial e integral, se você tem anos de contribuição e você tem anos de idade, hoje você tem anos para a aposentadoria, vai cair no pedágio até faltarem anos, talvez você trabalhe mais alguns anos, se você tivesse menos anos, cairia na aposentadoria integral" (Portuguese) - "Using the differential and integral calculus, if you have years of contribution and you started at the age of years, today you have years for the retirement, you will fall into the toll of until there are years left, maybe you work a few more years, if you had less years, you would fall into the full retirement" (English) and "Não pode ser sério o que acabei de ler" (Portuguese) - "It can't be serious what I just read" (English).

OFFCOMBR-3 reaching the best results using oversampling, which were configurations #3 and #4 with 0.88 for Precision and F-Measure, 0.89 for Recall and Accuracy, and the worst results using undersampling, which were configurations #5 and #6, 0.83 for Precision, 0.80 for F-Measure, 0.79 for Recall and Accuracy and 0.78 for Balanced Accuracy for configurations #5, and 0.85 for Recall, 0.80 for Balanced Accuracy, 0.76 for F-Measure and 0.74 for Recall and Accuracy for configuration #6. There is a difference between the use of oversampling and undersampling, for example configuration #3 achieved 15 true positives, 28 true negatives, 2 false negatives, and 1 false positives, while configuration #6 achieved 5 true positives, 16 true negatives, 13 false negatives, and 12 false positives.

The examples below of OFFCOMBR-3 were classified as true positive by configuration #3 and #4 and classified as false negatives by configuration #5 and #6. Examples: "Coitada e no mínimo para definí-la eu a chamaria de PODRE" (Portuguese) - "Poor thing and the least to define her I would call her ROTTEN" (English), "Basta olhar para o número de dislikes do seu comentário inútil e você verá que você não é aquele Palhaço moral" (Portuguese) - "Just look at the number of dislikes of your useless comment and you will see that you are not that moral Clown" (English) and "E PIAUI QUERENDO ASSUMIR O MUNDO COM ESSA FIGURA OU SOMENTE PESSOAS FEIAS NASCEM LÁ" (Portuguese) - "AND PIAUI WANTING TO SCARE THE WORLD WITH THIS FIGURE OR ONLY UGLY PEOPLE ARE BORN THERE" (English).

## Comparison of Results

In Table 4, we compare the results obtained by the proposed experiments and the works available in the literature from (de Pelle and Moreira 2017), (Fortuna et al. 2019), (Silva and Roman 2020), and (Leite et al. 2020).

There are different configurations between each classifier, so we will only generalize the results of the datasets in each classifier used in the related works. Our work using BERTimbau, showed promising results compared to related

Table 3: The results obtained for each dataset.

| Configuration | Dataset | Precision | Recall | Accuracy | Bacc | F-Measure |
|---|---|---|---|---|---|---|
| (#1) Original | OFFCOMBR-2 | 0.89 | 0.89 | 0.89 | 0.87 | 0.89 |
| (#1) Original | OFFCOMBR-3 | 0.87 | 0.88 | 0.88 | 0.79 | 0.87 |
| (#1) Original | Fortuna et. al. (2019) | **0.86** | **0.86** | **0.86** | **0.83** | **0.86** |
| (#2) No S.C | OFFCOMBR-2 | 0.88 | 0.88 | 0.88 | 0.86 | 0.88 |
| (#2) No S.C | OFFCOMBR-3 | 0.87 | 0.88 | 0.88 | 0.79 | 0.87 |
| (#2) No S.C | Fortuna et. al. (2019) | 0.83 | 0.83 | 0.83 | 0.77 | 0.82 |
| (#3) Oversampling | OFFCOMBR-2 | 0.87 | 0.87 | 0.87 | 0.84 | 0.86 |
| (#3) Oversampling | OFFCOMBR-3 | **0.88** | **0.89** | **0.89** | 0.83 | **0.88** |
| (#3) Oversampling | Fortuna et. al. (2019) | 0.85 | **0.86** | **0.86** | 0.82 | 0.85 |
| (#4) OverSampling and No S.C | OFFCOMBR-2 | **0.91** | **0.91** | **0.91** | **0.90** | **0.91** |
| (#4) OverSampling and No S.C | OFFCOMBR-3 | **0.88** | **0.89** | **0.89** | 0.81 | **0.88** |
| (#4) OverSampling and No S.C | Fortuna et. al. (2019) | 0.84 | 0.85 | 0.85 | 0.80 | 0.84 |
| (#5) Undersampling | OFFCOMBR-2 | **0.85** | **0.84** | **0.84** | **0.84** | **0.84** |
| (#5) Undersampling | OFFCOMBR-3 | **0.83** | **0.79** | **0.79** | **0.78** | **0.80** |
| (#5) Undersampling | Fortuna et. al. (2019) | 0.83 | 0.82 | 0.82 | 0.81 | 0.82 |
| (#6) Undersampling and No S.C | OFFCOMBR-2 | 0.85 | 0.85 | 0.85 | 0.85 | 0.85 |
| (#6) Undersampling and No S.C | OFFCOMBR-3 | **0.85** | **0.74** | **0.74** | **0.80** | **0.76** |
| (#6) Undersampling and No S.C | Fortuna et. al. (2019) | 0.82 | 0.81 | 0.81 | 0.81 | 0.81 |
| (#7) Back Translation | OFFCOMBR-2 | 0.86 | 0.87 | 0.87 | 0.83 | 0.86 |
| (#7) Back Translation | OFFCOMBR-3 | 0.85 | 0.86 | 0.86 | 0.72 | 0.84 |
| (#7) Back Translation | Fortuna et. al. (2019) | 0.82 | 0.82 | 0.82 | 0.78 | 0.82 |
| (#8) Back Translation and No S.C | OFFCOMBR-2 | 0.87 | 0.87 | 0.87 | 0.84 | 0.86 |
| (#8) Back Translation and No S.C | OFFCOMBR-3 | 0.85 | 0.86 | 0.86 | 0.72 | 0.84 |
| (#8) Back Translation and No S.C | Fortuna et. al. (2019) | **0.80** | **0.80** | **0.80** | **0.77** | **0.80** |

works, reaching an F-Measure of 0.91 for OFFCOMBR-2, while the best result for related works was 0.77 for SVM in the paper by (de Pelle and Moreira 2017), for OFFCOMBR-3 our work reached an F-Measure of 0.88, while the best result for a related work was 0.82 in the paper by (de Pelle and Moreira 2017), and for the Fortuna et. al. (2019) dataset, our work achieved 0.86 while the best result for a related work was 0.78 in the paper by Fortuna et. al. (2019).

## Conclusions and Future Works

In this work, we use BERTimbau to identify hate speech in the OFFCOMBR-2, OFFCOMBR-3, and Fortuna et. al. (2019) dataset. The best results for Accuracy, Bacc, and F-Measure were for the OFFCOMBR-2 using the Random Oversampling technique and removed special characters and retweets, configuration #4.

Even though there is no great gain, the results obtained with our work are promising when compared with those in the literature, such as: (de Pelle and Moreira 2017), (Fortuna et al. 2019), (Silva and Roman 2020), and (Leite et al. 2020). Our results were better compared to works in the literature that use NB, SVM, LSTM, LR, MLP, BoW + AutoML, and BERT Multilingual as classifiers in hate speech tasks.

As future works, we intend to use other types of preprocessing (for example: remove the stopwords), oversampling (for example: SMOTE, Borderline SMOTE, ADASYN), BERT models (for example: BertPT and AlbertPT, pretrained models in Portuguese which are freely available at https://github.com/diego-feijo/bertpt/) and check if there is more gain with different languages.

## Acknowledgments

This research was supported by FAPERGS - EDITAL 02/2021 - PROBIC/PROBITI in the project entitled Aspect-Based Sentiment Analysis Using Deep Learning: A Proposal Applied to the Portuguese Language.

## References

Alright. 2021. O impacto do boicote ao facebook no brasil,. https://alright.com.br/live-impacto-do-boicote-ao-facebook. Last accessed 15 Jul 2021.

Alshalan, R., and Al-Khalifa, H. 2020. A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Applied Sciences* 10(23).

Data Science Academy. 2021. Deep learning book,. https://www.deeplearningbook.com.br. Last accessed 26 Jul 2021.

de Pelle, R., and Moreira, V. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Fortuna, P.; Rocha da Silva, J.; Soler-Company, J.; Wanner, L.; and Nunes, S. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, 94–104. Florence, Italy: Association for Computational Linguistics.

Table 4: Comparison of results obtained with the literature works.

| Literature Work | Dataset | Classifier | F-Measure |
|---|---|---|---|
| de Pelle and Moreira (2017) | OFFCOMBR-2 | NB | 0.71 |
| de Pelle and Moreira (2017) | OFFCOMBR-2 | SVM | **0.77** |
| de Pelle and Moreira (2017) | OFFCOMBR-3 | NB | 0.79 |
| de Pelle and Moreira (2017) | OFFCOMBR-3 | SVM | **0.82** |
| Fortuna et al. (2019) | Fortuna et. al. (2019) | LSTM | **0.78** |
| Silva and Roman (2020) | Fortuna et. al. (2019) | SVM | 0.72 |
| Silva and Roman (2020) | Fortuna et. al. (2019) | LR | 0.69 |
| Silva and Roman (2020) | Fortuna et. al. (2019) | MLP | 0.69 |
| Silva and Roman (2020) | Fortuna et. al. (2019) | NB | 0.45 |
| Leite et al. | ToLD-BR | BoW + AutoML | 0.76 |
| Leite et al. | ToLD-BR | BERTimbau | 0.75 |
| Leite et al. | ToLD-BR | BERT Multilingual | 0.76 |
| Proposed Experiments | OFFCOMBR-2 | BERTimbau | **0.91** |
| Proposed Experiments | OFFCOMBR-3 | BERTimbau | **0.88** |
| Proposed Experiments | Fortuna et. al. (2019) | BERTimbau | **0.86** |


Hafidz Zulkifli. 2018. Understanding learning rates and how it improves performance in deep learning,. https://medium.com/modern-nlp/transfer-learning-in-nlp-f5035cc3f62f. Last accessed 22 Oct 2021.

Jurafsky, D., and Martin, J. H. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Leite, J. A.; Silva, D. F.; Bontcheva, K.; and Scarton, C. 2020. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *CoRR* abs/2010.04543.

Master's in Data Science. 2021. What is undersampling? https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/. Last accessed 01 Jan 2022.

Müller, A., and Guido, S. 2018. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Incorporated.

Nobata, C.; Tetreault, J. R.; Thomas, A. O.; Mehdad, Y.; and Chang, Y. 2016. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.

Patrycja Jenkner. 2020. Data augmentation in nlp: Best practices from a kaggle master. Disponível em: https://dev.to/patrycjajenkner/data-augmentation-in-nlp-best-practices-from-a-kaggle-master-1b4e. Acesso em: 04 Jan 2022.

Politize. 2020. Discurso de odio, o que é?,. https://www.politize.com.br/discurso-de-odio-o-que-e. Last accessed 27 Jul 2021.

Pratik Bhavsar. 2019. Transfer learning in nlp,. https://medium.com/modern-nlp/transfer-learning-in-nlp-f5035cc3f62f. Last accessed 22 Oct 2021.

Sagar Sharma. 2017. Epoch vs batch size vs iterations,. https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9. Last accessed 22 Oct 2021.

Silva, A., and Roman, N. 2020. Hate speech detection in portuguese with naïve bayes, svm, mlp and logistic regression. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, 1–12. Porto Alegre, RS, Brasil: SBC.

Vladimir Lyashenko. 2021. Data augmentation in python: Everything you need to know. Disponível em: https://neptune.ai/blog/data-augmentation-in-python. Acesso em: 03 Jan 2022.

Won, Y.-S.; Jap, D.; and Bhasin, S. 2020. Push for more: On comparison of data augmentation and smote with optimised deep learning architecture for side-channel. Cryptology ePrint Archive, Report 2020/655. https://ia.cr/2020/655.