UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

CARLOS ABEL CÓRDOVA SÁENZ

Understanding stance classification of BERT models: an attention-based mechanism

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Profa. Dra. Karin Becker

Porto Alegre
June 2022

CIP — CATALOGING-IN-PUBLICATION

Sáenz, Carlos Abel Córdova

Understanding stance classification of BERT models: an attention-based mechanism / Carlos Abel Córdova Sáenz. – Porto Alegre: PPGC da UFRGS, 2022.

84 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022. Advisor: Karin Becker.

1. Interpretability. 2. BERT. 3. Attention. 4. Stance classification. I. Becker, Karin. II. Título.

ACKNOWLEDGMENT

I want to begin expressing my gratitude to my brother Enrique and my parents, Tatiana and Carlos, for the support and motivation they have given me day by day before and during the completion of this master's degree.

I am deeply grateful to my dear counselor, Prof. Dr. Karin Becker, because her constant motivation, teachings, guidance, advice, and good humor have been fundamental during all these years, even in times as complicated as the start of the pandemic.

Special thanks to my colleagues, Régis Ebeling and Marcelo Dias. Their collaboration in the different investigations carried out together allowed me to learn a lot and share good experiences.

I am also grateful to my friends. Especially to my friend Hugo, who has been an inexhaustible source of activities and continuous motivation in my professional career. To Ingrid, Miguel, Alonso, Joseph, Arnold and Rolando, with whom I could always count throughout this academic phase.

I'd like to thank the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the scholarship and financial assistance that supported me during my postgraduate studies.

Finally, I would like to thank all the professors and members of the Informatics Institute at the UFRGS because their teachings and professional quality allowed me to learn and become a better professional.

ABSTRACT

BERT produces state-of-the-art solutions for many natural language processing tasks at the cost of interpretability. As works discuss the value of BERT's attention weights to this purpose, we contribute with an attention-based interpretability framework to identify the most influential words for stance classification using BERT-based models. Unlike related work, we develop a broader level of interpretability focused on the overall model behavior instead of single instances. We aggregate tokens' attentions into words' attention weights that are more meaningful and can be semantically related to the domain. We propose attention metrics to assess words' influence in the correct classification of stances. We use three case studies related to COVID-19 to assess the proposed framework in a broad experimental setting encompassing six datasets and four BERT pre-trained models for Portuguese and English languages, resulting in sixteen stance classification models. Through establishing five different research questions, we obtained valuable insights on the usefulness of attention weights to interpret stance classification that allowed us to generalize our findings. Our results are independent of a particular pre-trained BERT model and comparable to those obtained using an alternative baseline method. High attention scores improve the probability of finding words that positively impact the model performance and influence the correct classification (up to 82% of identified influential words contribute to correct predictions). The influential words represent the domain and can be used to identify how the model leverages the arguments expressed to predict a stance.

Keywords: Interpretability. BERT. Attention. Stance classification.

LIST OF ABBREVIATIONS AND ACRONYMS

AA	Absolute Attention
BERT	Bidirecional Encoder Representations from Transformers
FP	False Positive
FN	False Negative
IDF	Inverse Document Frequence
IW	Influential Word
LIME	Local interpretable model-agnostic explanations
LOO	Leave-One-Out
LSTM	Long Short-Term Memory
MPA	Minimum Positive Attention
NLP	Natural Language Processing
PAW	Proportional Attention Weight
PIW	Positive Influential Word
PLOO	Positive Leave-One-Out
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machines
TF	Term Frequence
TP	True Positive
TN	True Negative
XAI	Explainable Artificial Intelligence

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Models based on the Transformer architecture (VASWANI et al., 2017), particularly BERT (DEVLIN et al., 2019), have obtained state-of-the-art results with different natural-language processing (NLP) tasks, such as text classification, question answering, or translation (YILMAZ et al., 2019; GHOSH et al., 2019; GIORGIONI et al., 2020; KAWINTIRANON; SINGH, 2021; DAUDERT, 2021; WANG et al., 2021). BERT has significantly changed the NLP landscape. The great variety of pre-trained models with corpora in various languages (e.g., Portuguese, English), of different sizes (e.g., base, large), and shapes (e.g., cased, uncased), has allowed the development of BERT models that, through fine-tuning using specific domain datasets, can achieve high-performance results.

Nevertheless, BERT performance benefits have come at the cost of interpretability (TENNEY; DAS; PAVLICK, 2019; ROGERS; KOVALEVA; RUMSHISKY, 2020). According to Molnar (2019), interpretability is the degree to which a person can understand the reasons for a prediction produced by a Machine Learning (ML) model. Interpretability intends to provide the users with insights to understand the results obtained by a model, which can further help perform modifications. There have been attempts to adapt existing ML interpretability techniques to BERT, such as TranSHAP (KOKALJ et al., 2021) or TSESE (AYOUB; YANG; ZHOU, 2021), based on SHAP (LUNDBERG; LEE, 2017); and Captum[1], which relies on Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017).

Another trend has been to leverage BERT's attention weights for interpretability purposes. The attention weights are provided by the multiple internal attention heads that are central to BERT's underlying Transformer architecture. Several studies have expressed contradictory opinions, highlighting the pros and cons of using these values for interpretability (JAIN; WALLACE, 2019; WIEGREFFE; PINTER, 2019; SERRANO; SMITH, 2019; VASHISHTH et al., 2019; BAI et al., 2021). In addition, BERT has a large number of attention weights internally, which makes it difficult to interpret them. Some proposals have tried to consolidate these values (ABNAR; ZUIDEMA, 2020; CHEFER; GUR; WOLF, 2021) or provide a way to visualize them intuitively (VIG, 2019). However, these works have two main limitations. First, they are targeted at instance-level interpretability, making it hard to identify patterns in the overall predictions made by the

[1] https://captum.ai/

model. Second, these techniques focus on tokens, which are often meaningless parts of words, making it difficult to make sense of the attention weights in terms of the real-world semantics.

This work addresses stance classification, i.e., the task of identifying the position (e.g., in favor, against) expressed by a person on an issue under evaluation (ALDAYEL; MAGDY, 2021), in which BERT-based models have achieved state-of-the-art results (GIORGIONI et al., 2020; KAWINTIRANON; SINGH, 2021). Our research group has investigated stances regarding issues underlying the COVID-19 pandemics (i.e., vaccination, social isolation) and how they are influenced by political polarization (EBELING et al., 2020a; EBELING et al., 2021b; EBELING et al., 2022). The main objective of this work is to investigate how attention mechanisms can be leveraged to understand the stances predictions made by BERT models.

The present work proposes an interpretability framework to identify the most influential words for stances predicted using BERT models. The framework is focused on the overall model and thus, relates an attention score (*Absolute Attention*) to words that are significant within a set of documents in order to identify the most important ones for the classification (*Influential Words*). We also propose a metric (*Proportional Attention Weight*) to identify the influential words that contribute the most to the correct classification of instances (*Positive Influential Words*). Our framework starts from the tokens' weights collected for each instance according to the method in (CHEFER; GUR; WOLF, 2021), and develops a broader level of interpretability by:

- relating tokens' attention weights to their original words in each instance;
- aggregating the words' attention scores in individual instances to create an overall word-influence measure regarding the predictions made by the model.

To assess the proposed interpretability framework, we considered a wide experimental setting involving three case studies of stances expressed on Twitter on issues about the COVID-19 pandemic. We derived six datasets and deployed four BERT pre-trained models to address the English and Portuguese languages. Our experiments aim to answer the following research questions:

- RQ1: Does the choice of a BERT pre-trained model influence the results?
- RQ2: Do the *Influential words* contribute to the correct predictions?
- RQ3: Are the *Influential words* representative of the domain and stances?

- RQ4: Does the vocabulary in BERT pre-trained models affect the quality of the results?

- RQ5: How does the proposed interpretability framework compare to Captum's Sequence Classification Explainer?

Our results were encouraging. Very similar words with the highest absolute attention were found when comparing pairs of models trained in the same dataset, revealing that the results are not dependent on a particular pre-trained BERT model. We also found that the words with a high absolute attention score tend to positively influence the correct classification. We observed that each model's influential words represent the dataset's domain and can be related to the topics and arguments expressed by the respective polarized stances. We also confirmed that the influential words for the (correct) classification were not affected by the characteristics of BERT-models vocabularies (i.e., size, content). Finally, we found a statistically significant alignment between our interpretability framework's results and the ones obtained using Captum's *Sequence Classification Explainer*, an alternative interpretability model.

With preliminary results presented in (SÁENZ; BECKER, 2021) and (SáENZ; BECKER, 2021), the main contributions of this dissertation are:

- An interpretability framework to identify the most influential words for stances predicted using BERT models, leveraging internal attention weights. Unlike related work (CHEFER; GUR; WOLF, 2021; ABNAR; ZUIDEMA, 2020; VIG, 2019), it provides a broader level of interpretability focused on the overall model behavior against a test dataset. It also aggregates tokens into words that can be semantically related to the domain, and propose metrics to measure the influence of words in (correct) predictions;

- A broad set of quantitative and statistical experiments involving different case studies, datasets, BERT pre-trained models, and metrics to assess the proposed attention-based framework. The results provide valuable insights and patterns that allow us to generalize our findings, contributing with further evidence on the value of attention weights for the interpretability of BERT models for stance classification.

The rest of this work is structured as follows. Chapter 2 presents the theoretical foundation necessary to understand different aspects of our research. Chapter 3 reviews

related work in this area. Chapter 4 describes the case studies of stances on issues about COVID-19. Chapter 5 details the proposed interpretability framework and illustrates its use with a proof of concept. Chapter 6 describes the configuration, method, and results of the experiments performed. Finally, Chapter 7 draws conclusions, limitations and points out to future work.

2 THEORETICAL FOUNDATION

This chapter describes the aspects of BERT and attention weights that are relevant to this work. It also addresses existing interpretability methods for ML. Finally, it details the metrics used to assess our interpretability framework.

2.1 BERT

According to Devlin et al. (2019), BERT is a bidirectional model based on the Transformer architecture (VASWANI et al., 2017) that replaces the sequential nature of recurrent networks with a much faster attention-based approach. BERT works as a Masked-Language model (i.e., a language representation model) allowing to perform different NLP tasks and obtaining state-of-the-art results (KAWINTIRANON; SINGH, 2021; GIORGIONI et al., 2020).

2.1.1 Pre-trained BERT models

Pre-trained models are BERT instances that have passed for a Masked-language modeling process where they were trained in general NLP tasks (e.g., sentence prediction) using an extensive corpus of texts (e.g., the original BERT model was trained using approximately $3,300$M words, according to its authors). The result is a language representation model that can be later *fine-tuned* using a smaller dataset to perform an specific task. This approach, known as a type of *transfer learning*, makes BERT fine-tuning simpler and faster than traditional neural networks training methods, as the most complex model settings are already configured during pre-training. This frees users from substantial task-specific architecture modifications and allows them to use smaller training data sets when developing ML models, requiring them only to design the fine-tuning process.

The present study evaluates the proposed interpretability framework considering the English and Portuguese languages. It uses pre-trained models in those languages and variations in terms of word-casing.

The pre-trained models used in our work are:

- **BERT** (“bert-base-uncased")[1]: Presented in (DEVLIN et al., 2019), it is a non-

[1] https://huggingface.co/bert-base-uncased

case-sensitive model for the English language pre-trained using the BookCorpus[2] and the English Wikipedia[3]. This model is used in this work for the case study focused on English tweets.

- **BERT Multilingual cased** ("bert-base-multilingual-cased")[4]: Presented in (DEVLIN et al., 2019), it is a case-sensitive model pre-trained on the 104 languages with the largest Wikipedias[5]. This model is used in this research for the case studies focused on Portuguese tweets.

- **BERT Multilingual uncased** ("bert-base-multilingual-uncased")[6]: Presented in (DEVLIN et al., 2019), it is a non-case-sensitive model pre-trained on the 104 languages with the largest Wikipedias. As the datasets used to evaluate the presented framework were lower-cased, this model is used in this study to gather insights if its results are different from its cased counterpart.

- **BERTimbau** ("bert-base-portuguese-cased")[7]: Proposed in (SOUZA; NOGUEIRA; LOTUFO, 2020), it is a case-sensitive model for Brazilian Portuguese, pre-trained with the "BrWaC (Brazilian Web as Corpus)" (FILHO et al., 2018). This research uses this model since it reported state-of-the-art results for different NLP tasks on Brazilian Portuguese.

2.1.2 Tokenizers

BERT models need numerical data to process the inputs they receive. They use a tokenizer that converts the input texts into sets of tokens and then associates them with numerical values. The pre-trained BERT models have a fixed vocabulary that contains various tokens identified in the pre-training corpus (i.e., dictionary). If a word in an input text received by BERT is contained in the vocabulary, it is considered a single token. Otherwise, the word is divided into several tokens using the *WordPiece* algorithm (WU et al., 2016).

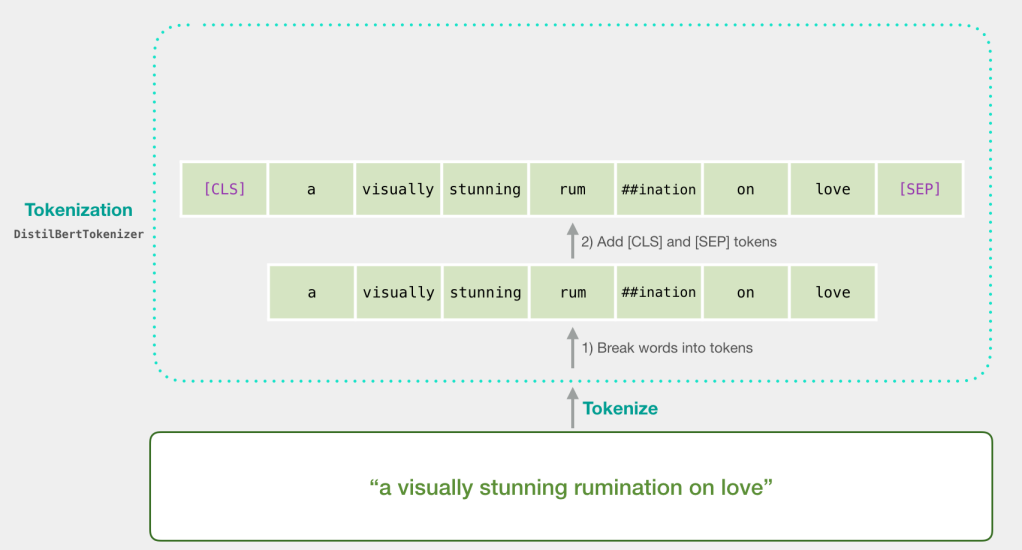
The use of tokens has some limitations. First, tokens do not necessarily have a

[2]https://yknzhu.wixsite.com/mbweb
[3]https://en.wikipedia.org/wiki/English_Wikipedia
[4]https://huggingface.co/bert-base-multilingual-cased
[5]https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages
[6]https://huggingface.co/bert-base-multilingual-uncased
[7]https://huggingface.co/neuralmind/bert-base-portuguese-cased

meaning per se. Their relevance in the domain can only be analyzed if the context of the word they were contained is identified. Also, the same token can belong to more than one word. For instance, in the BERTimbau pre-trained model's dictionary, the word "vacina" (vaccine) is not present. Hence, it is decomposed and subsequently explored by BERT as three tokens: "va", "##ci' ' and "##na". Notice that those tokens can be part of different words, such as "vacinação" (vaccination) and other words.

Figure 2.1 shows the tokenization process of a textual input, which ends with a numerical set of "token ids". Notice that the tokenization process generates two special tokens: [SEP] and [CLS]. The former is used as the separator of sentences for other tasks such as Sentence Prediction. The latter summarizes the whole input obtaining a sentence-level representation used for the text classification task.

Figure 2.1: BERT Tokenization process



Source: (ALAMMAR, 2018)

This work addresses the tokens limitations by proposing an interpretability framework that analyzes the influence of meaningful words instead of tokens, in the scope of the stance classification task.

2.1.3 Text classification using BERT

One of the NLP tasks for which BERT is very effective is text classification. As already described, fine-tuning BERT models allows obtaining vectorial representations of the texts received as input for the model. These representations can be used to train and evaluate different traditional ML algorithms for text classification through the token with

input text information focused on it ([CLS]).

One possibility is to use traditional ML algorithms, such as Random Forest, Logistic Regression, K-Nearest Neighbors, among others. For instance, in (SÁENZ; DIAS; BECKER, 2020; SÁENZ; DIAS; BECKER, 2021), we use BERT's [CLS] embeddings, obtained from the last layer of the network, and those traditional algorithms for classifying fake news.

Another alternative is to use *BertForSequenceClassification*, a BERT model specialized in text classification provided by the *transformers*[8] package in Python. This model has the same architecture as its original version, with a final additional classification layer. It receives the same input as the traditional BERT model and outputs the scores for the prediction labels. The label with the greatest score is the predicted label for each instance. Our previous work applying BERT for stance classification (SÁENZ; BECKER, 2021) proved that the fine-tuned BertForSequenceClassification model obtains better results compared to a set of traditional ML algorithms.

In this work, we use *BertForSequenceClassification* for the stance classification task.

2.2 Attention mechanism

BERT models are based on the Transformer architecture (VASHISHTH et al., 2019). Its use of the attention mechanism is responsible for the great performance and speed to train these type of models. Attention mechanisms model the dependencies between the parts of a text regardless of the distance between them. As shown in Figure 2.2, the Transformer model internally has several sets of attention mechanisms called “heads". These are distributed throughout all the layers of the architecture.

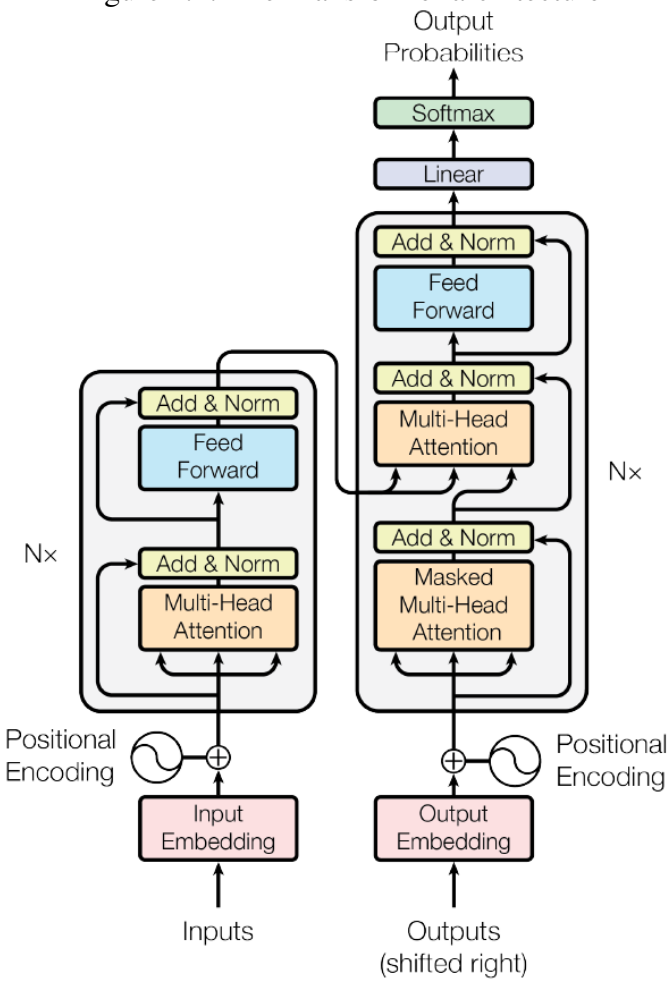
In BERT, the attention mechanisms perform the so-called "self-attention", which analyzes the interdependence between all the possible pairs of tokens of each textual input, assigning some "attention weight". BERT defines and adjusts attention weights in each layer and head, giving a higher weight to the tokens that it considers the most important ones for the task. These attention weights have been analyzed by various works, obtaining insights about the knowledge they have coded internally (ROGERS; KOVALEVA; RUMSHISKY, 2020).

In this work, our proposed interpretability framework leverages the attention

[8] https://huggingface.co/docs/transformers/

weights assigned to each input of a BERT model for stance classification.

Figure 2.2: The Transformer architecture



Source: (VASHISHTH et al., 2019)

2.3 ML/DL Interpretability

Molnar (2019) describes interpretability as “the degree to which a user can understand the results obtained from an ML model". Considering an ML engineer developing a model as the user, we can affirm that it is more interpretable than others if it is easier for the user to understand its behavior. The same concept can be applied to DL (Deep Learning) algorithms, which are harder to understand due to the many layers and settings.

Interpretability is sometimes referenced as “explainability", a term also commonly used concerning mechanisms that allow understanding the predictions of ML/DL models (e.g., XAI: Explainable Artificial Intelligence[9]). Despite the different definitions for both

[9]https://www.darpa.mil/program/explainable-artificial-intelligence

terms in the literature (RUDIN, 2019; LIPTON, 2018; MARCINKEVIČS; VOGT, 2020), in this research, we follow the decision made in other studies (CARVALHO; PEREIRA; CARDOSO, 2019; MILLER, 2019; MOLNAR, 2019), considering both terms as equivalent and that can be used interchangeably.

Molnar (2019) also proposes a taxonomy for interpretability based on the scope of the mechanisms to provide explanations:

- *Local Interpretability*: mechanisms targeted at answering the question “Why did the model make a certain prediction for an instance?". It analyzes the trained model's behavior focused only on single instances;

- *Global Interpretability*: mechanisms targeted at answering the question “How does the trained model make predictions?". It analyzes the trained model's behavior using a whole evaluation dataset (i.e., multiple instances).

In this work, *Local Interpretability* is called *instance-level interpretability*, while *Global Interpretability* is called *model-level interpretability*. The latter is the type of interpretability addressed in the proposed framework for BERT based stance classification.

In the remaining of this section, we describe some well-known interpretability methods.

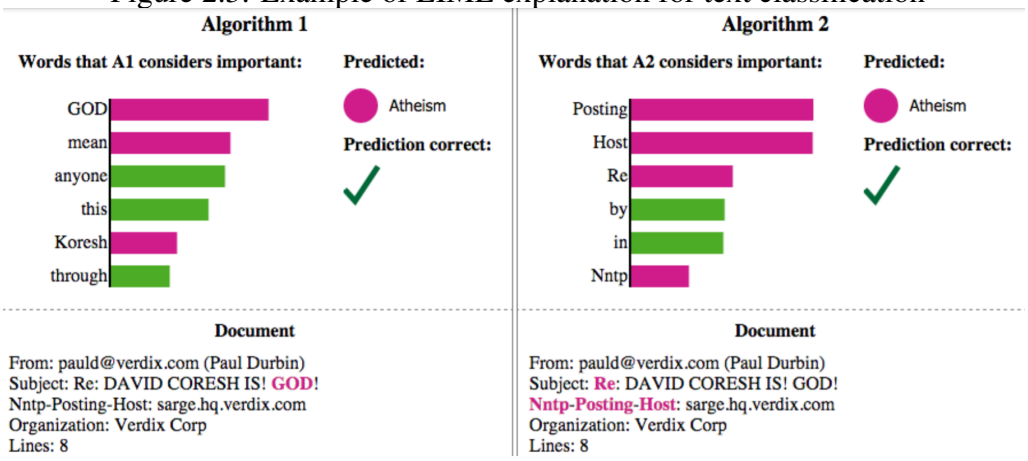
2.3.1 LIME

LIME (Local interpretable model-agnostic explanations) (RIBEIRO; SINGH; GUESTRIN, 2016) is a “surrogate model" (i.e., a model used to obtain an explanation from a black-box model) targeted at interpreting individual predictions. It generates a new dataset composed of multiple modifications made to the input instance under analysis and the predictions obtained by the black-box model to be interpreted. This dataset will be used to train an interpretable model (e.g., a decision tree), weighted by a measure of similarity between each disturbed instance and the original instance. Finally, the interpretability of the whole black-box model is reduced to the interpretation of the surrogate model, making it easier to understand the prediction.

LIME is an example of a mechanism for instance-level interpretability. By reducing the interpretability to the space of the surrogate model, this technique allows obtaining a good local interpretation of the model's behavior against individual instances. Figure 2.3

presents an example of a LIME explanation for a text classification instance predicted by two algorithms (on each side). The visualization makes it possible to identify the words most/least associated with the predicted class and which ones contributed the most.

Figure 2.3: Example of LIME explanation for text classification



Source: Ribeiro, Singh e Guestrin (2016)

LIME does not allow the analysis of the model as a whole (i.e., model-level interpretability). Another limitation is that this method is not always trustworthy, as the explanations obtained for two very similar disturbed instances can vary greatly (ALVAREZ-MELIS; JAAKKOLA, 2018), and they can be manually modified to hide biases (SLACK et al., 2020).

2.3.2 SHAP

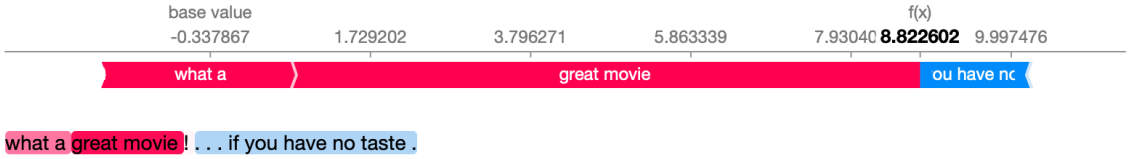
SHAP (SHapley Additive exPlanations) (LUNDBERG; LEE, 2017) is also an interpretability method for individual predictions, which calculates the contribution of each input attribute in an obtained prediction.

SHAP analyzes single instances through the calculation of *Shapley values* (SHAPLEY, 1988), which rely on the coalitional game theory. These values provide information about the contribution of each instance's features to the obtained prediction by verifying the variations of the results when introducing modifications in the input data. In contrast to LIME, SHAP settles in solid game theory, making it a more reliable method.

Although SHAP is also targeted at instance-level interpretability, it can also be used to do a global analysis of models by aggregating the Shapley values calculated for every instance. This allows creating model-level interpretability methods, leveraging features' importance, dependence, interactions, clustering, and summary plots.

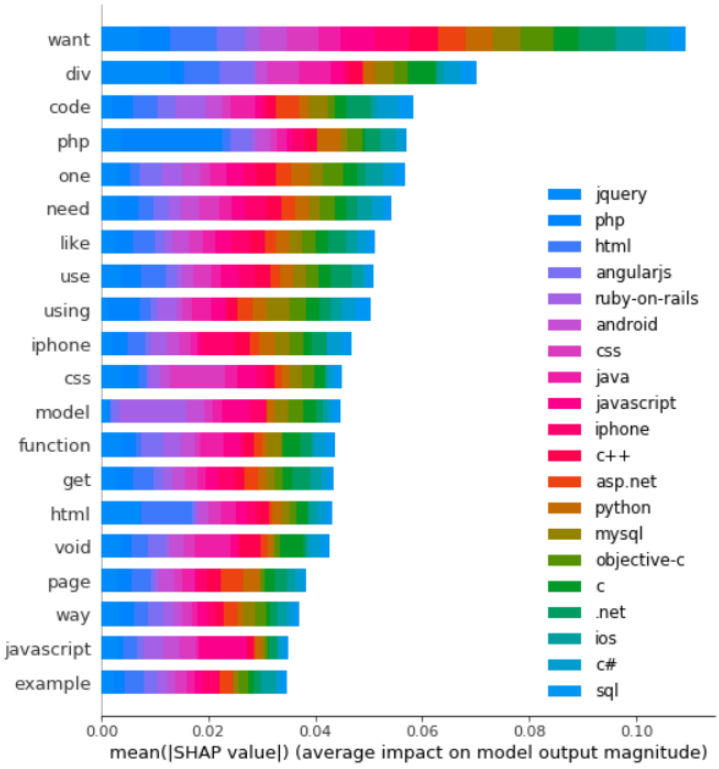
Figure 2.4.(a) presents a visualization for the interpretability of a single instance using SHAP, using colors to indicate a positive or negative association of the words in the text with the predicted class. On the other hand, Figure 2.4.(b) shows a model-level interpretability example, where each word in the bar chart has a degree of association (i.e., the average impact on model output magnitude), represented by a color, to a class predicted by the model.

Figure 2.4: SHAP instance-level and model-level interpretability examples



(a) Example of SHAP for instance-level interpretability

Source: (LUNDBERG; LEE, 2017)[10]



(b) Example of SHAP for model-level interpretability

Source: (LI, 2019)

A frequently mentioned disadvantage is that SHAP is slow, as the Shapley values on which it relies are computationally expensive. It can also be intentionally manipulated to hide biases (SLACK et al., 2020), which affects the user receiving the explanation.

In this work, we also aggregate the coefficients related to the instance-level to be able to interpret the behavior of the classification model. Since we work with BERT models, we use attention weights instead of Shapley values.

2.3.3 Captum

Captum (“comprehension" in Latin) is a Python library for the interpretability of ML/DL models. Captum’s models can be used for many applications, including BERT’s interpretability.

The Captum package for the interpretability of BERT is called *transformers-interpret*[11]. It relies on Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017) and Layer Integrated Gradients, a variation of the former, as its core attribution methods used to assign an attribution score to each input feature based on its influence on the prediction. The *transformers-interpret* package presents different interpretability mechanisms for various BERT models depending on the task.

The *Sequence Classification Explainer* is the interpretability model intended to be used for *BertForSequenceClassification*. It assigns attributions scores to the words of a certain input in a prediction obtained by a *BertForSequenceClassification* model. The scores can be positive or negative, where high positive values mean that the word is closely associated with the predicted class (i.e., positive contribution), and low negative values mean that the word is closely associated with a class different from the predicted one (i.e., negative contribution).

Figure 2.5 shows an example of the *Sequence Classification Explainer* for instance-level interpretability. Considering the green and red colors, one can notice the positive and negative attributions scores, in the fourth column, for each of the words in the sentence allowing the identification of the ones most associated with the predicted class.

Figure 2.5: Captum’s *Sequence Classification Explainer* example

Source: (CAPTUM, 2022)

Captum’s *Sequence Classification Explainer* handles the problem of analyzing tokens that are meaningless for interpreting the classification, enabling one to focus on

[11] https://github.com/cdpierse/transformers-interpret

words instead. However, it is still restricted to instance-level interpretability, not allowing the model's global behavior interpretation. In addition, Integrated Gradients, central to this method, have been criticized as they require a specific baseline configuration to provide reliable feature contribution insights (MADSEN et al., 2021).

The *Sequence Classification Explainer* from Captum is a baseline method used to compare our results obtained using attention weights.

2.4 Evaluation metrics

This section describes the metrics used in our experiments.

2.4.1 Performance metrics

In classification tasks, such as the one addressed in this work, models are traditionally evaluated using metrics that rely on a confusion matrix. For example, as illustrated in Table 2.4.1 for a binary classification problem, the confusion matrix is composed of:

- True positives (TP): Total predictions where the predicted class is positive, equal to the expected class.
- False Positives (FP): Total predictions where the predicted class is positive, different from the expected negative class.
- True negatives (TN): Total predictions where the predicted class is negative, equal to the expected class.
- False negatives (FN): Total predictions where the predicted class is negative, different from the expected class.

Table 2.1: Confusion matrix example for binary classification

		Predicted Value	
		Class A	**Class B**
Expected Value	**Class A**	TP	FN
	Class B	FP	TN

Good models are the ones that maximize the TP and TN obtained, minimizing the FP and FN. For classification problems with more than one class, these values are relativized to each specific class, considering it as positive and the other classes as negative.

For example, for a given class i, TP_i is the number of predictions where the predicted class i was equal to the expected class i. At the same time, FN_i is the number of predictions where the predicted class differed from the expected class i.

From the confusion matrix, and taking as n the total number of instances and i as the class under evaluation, different metrics can be calculated:

- **Accuracy**: is the total number of hits of the model, and it is used to evaluate its behavior as a whole. Its calculation is presented in the Equation 2.1;

$$Acc = \frac{\sum TP_i + \sum TN_i}{n} \tag{2.1}$$

- **Precision**: is the proportion of times the model was correct when predicting a class i. It allows recognizing how accurate the model is with respect to all the predictions it makes about a given class i. It is calculated using the Equation 2.2;

$$Prec_i = \frac{TP_i}{TP_i + FP_i} \tag{2.2}$$

- **Recall**: is the proportion of times the model hit a class i out of all the times it was expected to predict that class. It serves to recognize how accurate the model is with respect to all the expected predictions about a class. Its calculation is described in the Equation 2.3;

$$Rec_i = \frac{TP_i}{TP_i + FN_i} \tag{2.3}$$

- **F1-measure**: is the weighted harmonic mean of the precision and recall. It combines these two metrics in order to evaluate both aspects addressed by these metrics at the same time. The Equation 2.4 describes its calculation.

$$F1_i = \frac{2 * Prec_i * Rec_i}{Prec_i + Rec_i} \tag{2.4}$$

For all these metrics, values closer to 1 indicate better performance.

The Precision, Recall, and the F1-measure metrics allow the analysis of the model's performance against each class i individually. There are aggregation metrics to evaluate the model's performance considering all classes, such as *weighted average*, *micro average* and *macro average*. The *weighted average* calculates the average of the

results of each class and adds them, weighting each value with the proportion of instances of the class and dividing the result by the total number of instances. This way of aggregating the results also allows dealing with datasets where the number of instances per class is unbalanced. We adopted this aggregation metric in our work.

2.4.2 Ranking metrics

The analyses performed also compare how our interpretability framework ranks influential words and compares them with other word influence/relevance techniques (e.g., TF-IDF, Captum scores). For this purpose, some common metrics in information retrieval and ranking algorithms performance evaluations are used (MANNING; RAGHAVAN; SCHüTZE, 2008).

- **Precision@k**: is the number of at most k documents that are considered relevant according to a proposed method and a baseline method simultaneously, divided by k. The Equation 2.5 describes its calculation.

$$Prec@k = \frac{\#(\text{Proposed method relevant documents}) \cap \#(\text{Baseline relevant documents})}{k} \tag{2.5}$$

- **R-Precision**: is the number of documents that are considered relevant according to a proposed method and a baseline method simultaneously, divided by the number R of influential documents based on the baseline. This measure is very similar to the Precision@k, but uses a restricted number of R relevant documents obtained by a baseline used to compare a method. The calculation of this metric is the same as Equation 2.5 with R taking the place of k;

- **NDCG@k**: The Normalized Discounted Cumulative Gain (NDCG) is a ranking measure that compares documents rankings according to a proposed method with a baseline. This measure does not consider the absence of relevant documents in the ranking under evaluation. It only verifies that documents' positions in a ranking according to one method are equal to or very close to their position according to a baseline method. The NDCG@k performs the comparison described for a set of k ranked documents, allowing them to be compared for different cut-offs defined

by k. The calculation of this metric can be seen in Equation 2.6, there, Q represents the set of k ranked documents based on a method on evaluation, $R(j,d)$ is the ranking d received by a document j from a baseline ranking method, and Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG is 1.

$$\text{NDCG}(Q,k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2(1+m)} \tag{2.6}$$

For all of these ranking metrics, values closer to 1 indicate that the results are more similar, while values closer to 0 mean they are different.

The experiments in this work focus on (influential) words rather than (relevant) documents. Thus, in our context:

- the Precision@k metric is used to obtain the proportion of words that are simultaneously influential according to the proposed technique and a baseline technique to verify how similar the obtained results are;
- the R-Precision metric is used with baseline methods where the number of influential words is minimal and varies based on the different models under analysis, as is the case of the comparison made with the BERTopic's "Topic words";
- the NDCG@k metric complements the Precision@k results, determining if the order of the words ranked by the proposed method is similar to those obtained through a baseline method.

3 RELATED WORK

This chapter generally describes the works in stance classification, BERT interpretability, and usefulness of attention weights to give a better idea of the scenario in which this research arises.

3.1 Stance classification

Stance classification has been addressed using ML/DL supervised algorithms such as Logistic Regression (TSAKALIDIS et al., 2018; KUCHER et al., 2020), SVM (LAI et al., 2020), artificial neural networks (ZHANG et al., 2019), LSTM (WEI; LIN; MAO, 2018; VANTA; AONO, 2020), and Gaussian Processes (LUKASIK et al., 2019), among others.

However, state-of-the-art results have been achieved using BERT (POPAT et al., 2019; GHOSH et al., 2019; GIORGIONI et al., 2020; KAWINTIRANON; SINGH, 2021). As described in Section 2.1.3, BERT models are pre-trained using large unlabeled corpora, which makes them able to create embeddings (i.e., vector representations) that summarize syntactical and semantical relationships on the input received without making special configurations on the model. These vector representations can then be used to train different classification algorithms for stance classification. This ease of use and their power to represent knowledge from text allow BERT models to obtain outstanding results.

3.2 BERT's Intepretability

Despite the excellent results in many NLP applications, BERT-based models are black boxes, and thus it is not easy to identify the influential features for classification. Moreover, the works with state-of-the-art results in applications apart from stance classification, such as document retrieval (YILMAZ et al., 2019), fake news detection (SÁENZ; DIAS; BECKER, 2021), or sentiment analysis (DAUDERT, 2021; WANG et al., 2021), do not make efforts to tackle this limitation.

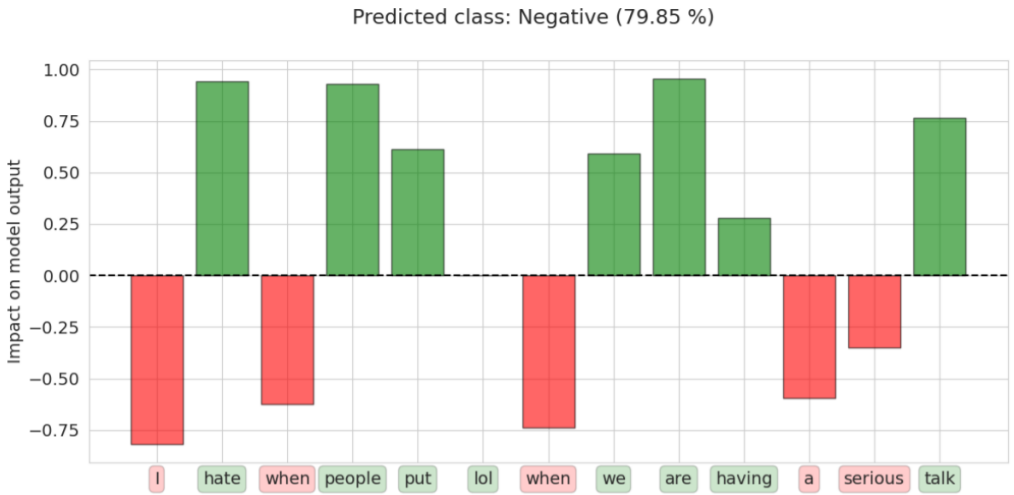
The remaining of this section will describe some methods that have been applied to provide interpretability to BERT models.

3.2.1 Feature-based Interpretability

One method applied to understand BERT is the Integrated Gradients (SUNDARARAJAN; TALY; YAN, 2017), which are present in modern interpretability methods such as the Captum package, previously described in Section 2.3.3. As detailed, Captum's *Sequence Classification Explainer* model allows analyzing the contribution of each input word in an instance for BERT's prediction. This model can be used to develop visualizations of the words contributions scores as presented in Figure 2.5. However, Integrated Gradients have been criticized as they require a specific baseline configuration to provide reliable feature contribution insights (MADSEN et al., 2021), making them harder to implement.

Other studies use SHAP (LUNDBERG; LEE, 2017) to develop their interpretability mechanisms. TransSHAP (KOKALJ et al., 2021) performs feature contribution analysis of BERT models' predictions using the SHAP technique. They present an enhanced visualization (Figure 3.1) for a better understanding of the results, which is then assessed using a survey for the users.

Figure 3.1: TransSHAP visualization of prediction explanations for negative sentiment

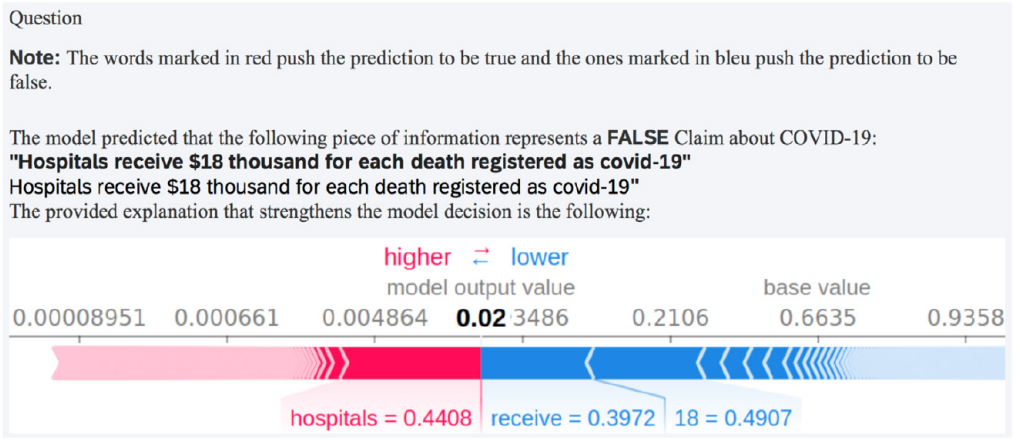


Source: (KOKALJ et al., 2021)

Ayoub, Yang e Zhou (2021) also uses SHAP to provide interpretability to BERT models. Through the technique they proposed, they evaluate the contribution of each feature (i.e., word) analyzed at instance-level (i.e., a misinformation claim). SHAP results, the predicted instance, and the prediction itself are presented in a visualization, as illustrated in Figure 3.2.

SHAP has the limitation that it tends to be computationally expensive (MADSEN et al., 2021). Additionally, all these works, including Captum's *Sequence Classification*

Figure 3.2: Text + SHAP Explanation



Source: Ayoub, Yang e Zhou (2021)

Explainer are limited to analyzing only individual instances and not the entire model behavior.

3.2.2 Attention-based Interpretability

3.2.2.1 Attention weights and Interpretability

Attention mechanisms are central to the Transformer's architecture and critical to the excellent performance of BERT-based models. Nevertheless, there is no consensus on the value of attention weights for interpretability, mostly due to the insufficient knowledge of how BERT understands and represents the syntactic, semantic, and linguistic patterns it leverages for NLP (TENNEY; DAS; PAVLICK, 2019; ROGERS; KOVALEVA; RUMSHISKY, 2020).

Works such as (JAIN; WALLACE, 2019; SERRANO; SMITH, 2019) developed experiments on the variation of the results when altering the attention weights, concluding that these mechanisms do not contribute to interpretability.

In the opposite direction, Wiegreffe e Pinter (2019) strongly argue that experiments rejecting the value of attention weights for interpretability are not sufficient evidence to rule out their value. Vashishth et al. (2019) performed different NLP experiments and found proof that the attention weights do serve for interpretability and are correlated with feature-importance metrics.

Bai et al. (2021) claim that the combinatorial shortcuts in the Transformers architecture make it hard to provide interpretability through attention mechanisms, proposing two methods to mitigate this issue, which unfortunately requires advanced knowledge of

the intrinsics of BERT neural networks. This limits the usability of this method for users less skilled in the inner workings of BERT.

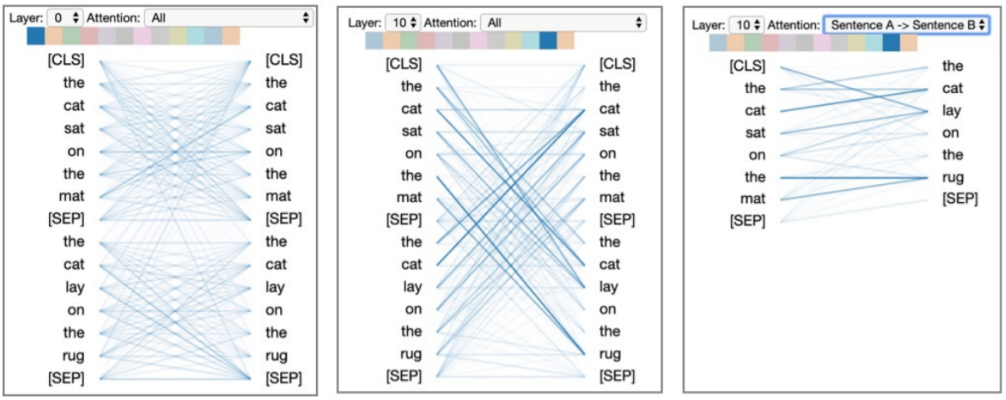
All these works motivate us to analyze the value of attention weights for the interpretability of BERT-based stance classification models.

3.2.2.2 BERT's interpretability methods using attention weights

A significant concern is how the attention weights are assigned to tokens during the training. First, the Transformer architecture relies on various sets of attention weights (i.e., "heads"), distributed throughout the network layers pointing to different parts of the inputs that are finally combined to produce a final attention weight (i.e. "multi-head attention"). Thus, the relationship between the input, the attention weights, and the outcomes of the model is not straightforward (ROGERS; KOVALEVA; RUMSHISKY, 2020). Second, BERT processes the input text as tokens rather than words, as discussed in Section 2.1.2. Thus the attention weights are assigned over these items, which are more difficult to understand outside the model.

Related work addresses these challenges by providing visualization tools and mechanisms to consolidate the weights. *Bertviz* (VIG, 2019), as illustrated in Figure 3.3, allows visualizing the attention of the tokens in a text under different perspectives but always considers each attention weight relative to a given layer and head, which is difficult to be understood by a non-expert user.

Figure 3.3: Example of Attention-head view for BERT



Source: (VIG, 2019)

Some studies have made efforts to condense the multiple attention weights in BERT to produce a value that summarizes the information produced by these coefficients. Abnar e Zuidema (2020) proposed two strategies to condense the attention values obtained by each token in the whole network from the weights in each part of the network

(i.e., attention rollout and attention flow), assuming that attentions can be linearly combined. Nevertheless, this linearity cannot be guaranteed.

The technique proposed by Chefer, Gur e Wolf (2021) leverages LRP (Layer-wise Relevance Propagation) to overcome this limitation and summarize the attention weights using information related to both the relevance and gradient. This method provides a sound consolidation for attention weights. However, it has two limitations:

a) it highlights the tokens that most contributed to classifying a single instance (i.e., instance-based interpretability) without mechanisms for an aggregated model-oriented interpretation;

b) as the tokens may not have precise semantics in the domain and possibly compose many different words, it may be hard to understand their meaning in the domain.

3.2.3 Final considerations

Table 3.2.3 summarizes the different described studies providing interpretability mechanisms to BERT models. Notice that all of those works provide only instance-level interpretability, except ours.

Table 3.1: Works proposing interpretability mechanisms for BERT models

Study	**Technique**	**Consolidation**	**Visualization**	**Interpretability Level**
Sequence Classification Explainer	Integrated Gradients	✓	✓	Instance
TransSHAP (KOKALJ et al., 2021)	SHAP	✓	✓	Instance
Text + SHAP Explanation (AYOUB; YANG; ZHOU, 2021)	SHAP	✓	✓	Instance
LRP (CHEFER; GUR; WOLF, 2021)	Attention-based	✓		Instance
Attention rolllout/fallback (ABNAR; ZUIDEMA, 2020)	Attention-based	✓		Instance
BERTViz (VIG, 2019)	Attention-based		✓	Instance
Our work	Attention-based	✓	✓	Model

This dissertation contributes to a better understanding of the role of attention weights for interpretability by proposing a framework to assess the behavior of a stance classifier based on BERT in terms of domain words' contribution to classification. It builds on the consolidated weights proposed in (CHEFER; GUR; WOLF, 2021) to provide model-level interpretability by revealing influential words that are meaningful in the domain and proposing metrics to assess their influence on correct prediction.

4 CASE STUDIES OF STANCES ON ISSUES ABOUT THE COVID-19

This chapter describes three case studies of stances on Twitter used to assess the proposed method in Chapter 6.

4.1 Introduction

The COVID-19 pandemic brought many discussions on social media. People used these platforms to get informed, express their sentiments and emotions, or defend and promote stances on topics related to COVID-19. Our research group investigated the influence of political polarization on different COVID-19's related issues in Brazil, such as social isolation (EBELING et al., 2020b; EBELING et al., 2020a; EBELING et al., 2021b) and vaccination (EBELING et al., 2021a; EBELING et al., 2022), expressed in Twitter.

Considering the deep understanding developed through that research, we adopt datasets related to COVID-19 issues in our experiments.

4.2 Case study 1: Social Isolation

This case study was presented and described in detail in (EBELING et al., 2020b; EBELING et al., 2020a; EBELING et al., 2021b). It is based on the Brazilian scenario by late March 2020, at the beginning of the pandemic in Brazil. It addresses the stances of Brazilians about social isolation, the only known effective control action available back then. The Brazilian Ministry of Health was in favor of social isolation, while President Jair Bolsonaro was in favor of less strict measures, promoting the use of medicines without scientifically proven efficacy (e.g., chloroquine). By March 2020, he promoted campaigns on social networks against social isolation, such as "Brazil cannot stop", arguing that the damages to the economy were more extensive than the health benefits.

The dataset in this case study is composed of tweets in Portuguese collected by the end of March 2020. The tweets were labeled using representative hashtags for each stance. The polarized labels (or classes) are:

- *Chloroquinners*: represented by hashtag #OBrasilNãoPodeParar ("Brazil cannot stop"). This group is characterized by arguments stating that social isolation is

not an effective solution due to the negative economic consequences.

- *Quarenteners*: represented by hashtag #OBrasilTemQuePararBolsonaro (“Brazil must stop Bolsonaro"). This group is against the campaign “Brazil can not stop" promoted by President Jair Bolsonaro. They believe that social isolation, as promoted in other countries, is the only available solution known to mitigate the pandemic.

- *Neutrals*: represented by hashtags #FiqueEmCasa and #FicaEmCasa (variations of “StayAtHome"). This group endorses social isolation in general. However, contrary to the Qurenteners group, they do not express a political bias.

We performed a topic modeling process using BERTopic over the dataset to identify the topics and arguments most frequently discussed by each stance. These results are summarized in Table 4.1, along with some statistics of the dataset in this case study. Further details can be found in the original papers (EBELING et al., 2020b; EBELING et al., 2020a; EBELING et al., 2021b)

Table 4.1: Social Isolation case study: number of tweets and arguments of the stances

Stance	# Tweets	Main arguments
Chloroquinners	74,395	concern on the economic impact of social isolation praise and support for the president minimization of COVID-19 health risks rejection of the governors in favor of social isolation rejection of the prospective presidential candidate João Dória
Quarenteners	31,060	fear of the fatal consequences of COVID rejection of the president and his actions criticism of pro-economy businessmen rejection of the president’s supporters
Neutrals	201,499	discussion on measures to combat COVID-19 concern about the implications of social isolation random discussions about everyday aspects of the pandemic (e.g., entertainment, sentiment, etc)

4.3 Case study 2: Vaccination

This case study was presented in (EBELING et al., 2021a; EBELING et al., 2022). It is set in the context of emerging news about the various phases of the COVID-19 vaccine. Throughout 2020, when vaccines were under development and testing, there was an interest in several countries in securing their supply in the future. In Brazil, there were partnerships between international laboratories and Brazilian research Institutes, such as

the Chinese pharmaceutical company Sinovac and the Brazilian Butantan Institute, to produce the vaccine “Coronavac". The governor of São Paulo, João Dória, endorsed this cooperation initiative, some say for political reasons. As Bolsonaro and Dória were potential candidates for the 2022 Presidential elections, the antagonism between them and their followers grew enormously on social networks.

The dataset in this case study comprises tweets written in Portuguese between January 2020 and April 2021, covering all phases of vaccine development, up to their approval and application to the population. These tweets were labeled using representative hashtags for each stance. These stances are described below:

- *Pro-vaxxers*: represented by the hashtags #EuVouTomarVacina, #VacinaBrasil, #VacinaÉAmorAoPróximo, #VacinaJá, #VacinaNoBrasil, #VacinaParaTodos, #VacinasPelaVida, #VemVacina and #VacinaUrgenteParaTodos. This stance is in favor of vaccination and uses hashtags to support vaccination programs (e.g., VaccinesForLife) and to raise awareness about its urgency (e.g., VaccineNow)

- *Anti-vaxxers*: represented by hashtags #EuNãoVouTomarVacina, #VacinaNão, #VacinaObrigatóriaNão and #NãoVouTomarVacina. This stance is against COVID-19 vaccination and uses hashtags that express no intentions to get vaccinated (IWontTakeVaccine) or against mandatory vaccination to reach community immunization.

- *Anti-sinovaxxers*: represented by hashtags #VachinaNão, #VacinaChinesaNão, #VachinaObrigatóriaNão, #VachinaNãoPresidente. This stance is specifically against Coronavac, referred to as “the Chinese vaccine” or the diminishing expression “vacchina”. Although it could be regarded as an anti-vax stance, the arguments endorsing this position are more politically motivated.

We performed a topic modeling process also using BERTopic to identify the main arguments discussed by each stance. These stances’ arguments, along with the number of tweets, are summarized in Table 4.2. Further details can be found in the original studies (EBELING et al., 2022; EBELING et al., 2021a).

Table 4.2: Vaccination case study: number of tweets and arguments of the stances

Stance	# Tweets	Main arguments
Pro-Vaxxers	19,363	joy and gratitude for vaccines expectation for getting vaccinated as soon as possible praise for science and Brazilian Public Health System strong criticism of Bolsonaro and the government's actions
Anti-Vaxxers	25,371	individual choice opposition to mandatory vaccination criticism towards the governors' "dictatorship" rage against STF ruling (constitutionality) support to the president and Federal Government
Anti-Sinovaxxers	15,010	opposition to mandatory vaccination distrust and rejection of Coronavac mistrust/prejudice against the "Chinese" origin opposition to Dória praise to Bolsonaro

4.4 Case study 3: Chloroquine/Hydroxychloroquine

This case study comes from a dataset presented in (MUTLU et al., 2020). It is focused on discussions on Twitter about the COVID-19 pandemic, particularly chloroquine/hydroxychloroquine-related topics. On the one hand, there were people in favor of the use of hydroxychloroquine to combat COVID-19, motivated by different events such as the publication of preliminary results on its use as a treatment for COVID-19 patients, news in the USA about the purchases of hydroxychloroquine sulfate tablets by the Department of Veterans Affairs, publications of questionable origin about the effectiveness of the drug or even fake news widely shared on social networks. On the other hand, there were people against using those drugs. They rejected the news shared on social networks, grounded on the FDA's claims and scientific recommendations, expressing clear opposition to chloroquine/hydroxychloroquine.

This dataset was collected in April 2020 and contains tweets in English. The labeling process was manual, following a set of guidelines described in the original paper. The stances in this case study are:

- *Pro-Chloroquine*: this stance expresses support for hydroxychloroquine either directly (e.g., by promoting its use) or indirectly (e.g., by commenting on its benefits).
- *Anti-Chloroquine*: this stance is against the use of hydroxychloroquine. These people do not usually express a personal opinion but tend to include URLs of news or studies on the ineffectiveness of this drug.
- *Neutrals*: this stance has not a clearly defined position for or against the drug. They

usually express doubts about opinions on social networks, ask questions to learn more about the subjects, or express themselves in a neutral tone.

The original study has only included a frequency analysis of the most frequently used words for each stance to describe the behavior underlying these stances. As a complement, we applied a topic modeling process using BERTopic to gain better insights on these issues, reproducing the method previously applied in our previous research. Table 4.3 describes the number of tweets and the main arguments found for each stance.

Table 4.3: Chloroquine/Hydroxychloroquine case study: number of tweets and arguments of the stances

Stance	# Tweets	Main arguments
Pro-Chloroquine	3,713	support to the use of hydroxychloroquine praise and support for Trump distrust of the FDA rejection to journalists disagreeing with Trump
Neutrals	2,112	concern on published clinical trials concern about the drugs' promotion by the White House references to India, which made an export ban discussion about the drugs' use in malaria treatments
Anti-Chloroquine	3,901	rejection of hydroxychloroquine endorsement to FDA rejection to Trump and Fox News allegations of secret financial interests

4.5 Final considerations

This chapter described the three case studies used in this paper. Each of them presents stances on various issues around the COVID-19 pandemic. In our research, we derived different data sets for each case study to assess our solution through a number of experiments, further detailed in Chapter 6.

5 FRAMEWORK FOR INTERPRETABILITY OF BERT-BASED STANCE CLASSIFICATION

This chapter presents the proposed framework for the interpretability of BERT-based models for stance classification. In the following sections, we first provide an overview of the framework, highlighting its main contributions. Then, each component is presented in detail. Finally, we illustrate how the proposed metrics could be used to interpret a stance classification model.

5.1 Overview

The main goal of the proposed framework is to interpret the predictions obtained at model-level, by identifying the words that contributed the most to the correct prediction of stances based on the consolidation and aggregation of the attention weights assigned to tokens at the instance-level. This framework targets users who do not necessarily have a strong knowledge of the internal workings of BERT but want to understand why the model is making such predictions. The striking features of our solution for interpretability are:

- assignment of attention weights to the original words, based on the tokens scores, to be able to interpret the model in terms of domain-related concepts;
- model-level interpretability by the aggregation of attention weights in a set of predictions, through the concept of *Absolute Attention* (AA);
- quantification of the contribution of words with high attention scores to correct classification, through the concepts of *Influential Words* (IWs), *Positive Influential Words* (PIWs) and *Proportional Attention Weight* (PAW).

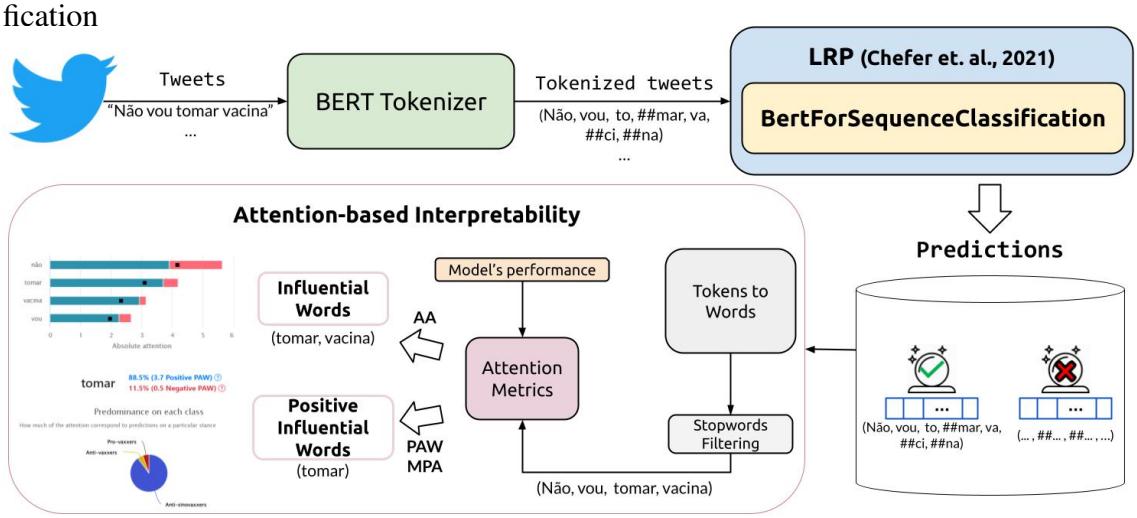
As shown in Figure 5.1, the framework is divided into two phases:

a) prediction of stances using a BERT-based model and collection of the consolidated tokens' attention weights;

b) identification of the most influential words for the models' predictions through the calculation of proposed attention metrics (i.e., AA, PAW).

The proposed framework requires a labeled dataset. Part of this dataset (i.e., the training and validation sets) will be used to create a stance classification model. The other

Figure 5.1: Overview of the framework for interpretability of BERT-based stance classification

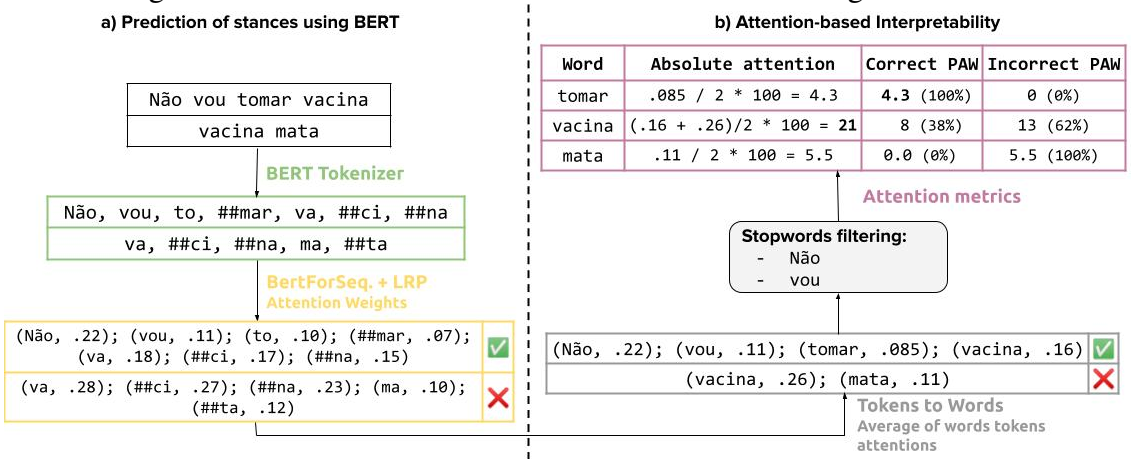


part (i.e., the test set) will be used to make predictions and gather attention weights that will serve to assess the model's performance and identify the words that were more influential for the classification (i.e., Influential Words), particularly the ones that influenced the correct predictions (i.e., Positive Influential Words).

The framework was developed and assessed considering tweets. Its generalization regarding other documents needs to be further investigated.

Figure 5.2 presents a running example to follow during the explanation of our framework. The remaining of this chapter will describe in detail each phase and illustrate the use of our proposal through a proof of concept.

Figure 5.2: Transformations from tweets to attention weights of words



5.2 Prediction of stances using BERT

The main objective in this phase is to collect, using a test dataset, the input tokens, their attention weights, and predicted labels for them. It is divided into two main tasks:

- Training a BERT stance classification model, as a requirement for applying the proposed method (Figure 5.3);
- Collection of the evaluation data (i.e., predictions) to identify the influential words for the stance classification (Figure 5.4).

This phase is illustrated through the example in Figure 5.2.(a), using the tweets “não vou tomar vacina" and “vacina mata".

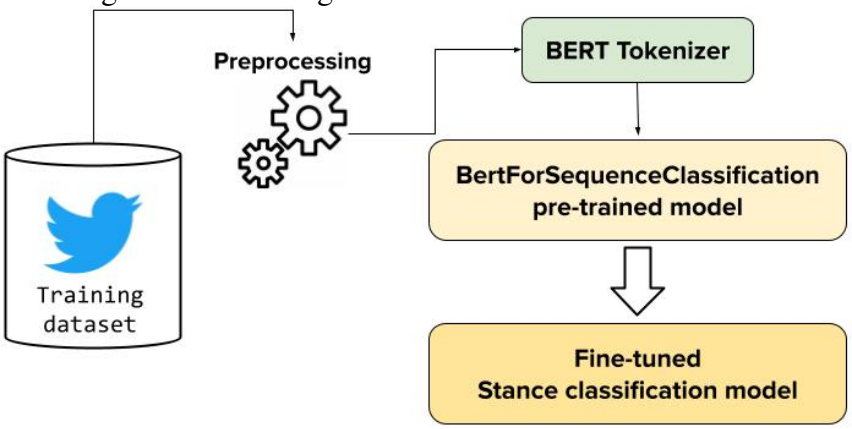
5.2.1 Training a BERT stance classification model

First, a dataset of labeled tweets containing stances around a topic in discussion is required. This dataset needs to be preprocessed to avoid introducing bias and not corrupting the model. Preprocessing can be done using typical actions such as removing mentions/URLs/special characters, lower-casing, and discarding short tweets. If hashtags were used to crawl the data and determine the stances, they should also be deleted to avoid bias. Then, this dataset should be divided into two subsets: the training/validation dataset and the test dataset.

Each instance in the training/validation dataset should be transformed into a set of tokens using *BERT Tokenizer*. As described in Section 2.1.2, each BERT pre-trained model has its tokenizer that uses its own vocabulary of words. If a word in the input text is also contained in the model’s vocabulary, it is considered a single token; otherwise, it is divided into tokens (e.g., “vacina” or “va”, “##ci” and “##na”). For example, in Figure 5.2.(a), “vacina mata" is transformed into the tokens “va", “##ci”, “##na”, “ma”, “##ta”.

As depicted in Figure 5.3, the *BertForSequenceClassification* model is fine-tuned/trained using this dataset of tokens and labels to produce a more robust model for stance classification on the domain of the training dataset.

Figure 5.3: Training a BERT stance classification model



5.2.2 Collection of the evaluation data

The fine-tuned stance classification model is integrated into the LRP model previously described in Section 3.2.2.2. As presented in Figure 5.4, this integration is done by passing the stance classification model's configuration details and fine-tuned embeddings to the LRP model, which internally possesses a *BertForSequenceClassification* instance that will use this information to perform the stance classification.

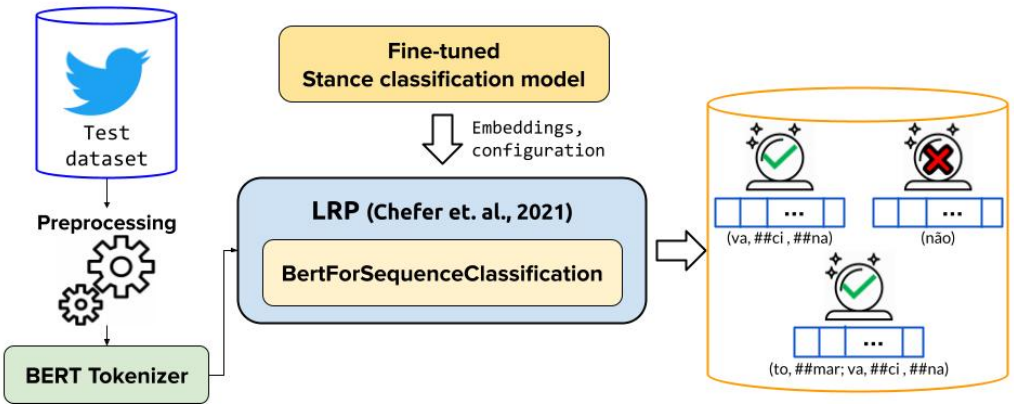
This task consists in obtaining a set of predictions. To that end, a test set needs to be preprocessed and tokenized in the same way as described in Section 5.2.1. Then, it is passed to the LRP model, which will output both the predictions and the attention weights of each token in the input instances of the test dataset. Those values accompany the original input instance (i.e., a set of tokens), so they can be used for the next phase of the framework.

The result of this task is depicted at the bottom of the running example in Figure 5.2.(a), where there is a correct and an incorrect prediction. It is also possible to see that each token is accompanied by its attention weight. All the tokens and weights in an instance are accompanied by their respective predicted label.

5.3 Attention-based Interpretability

In this phase, our framework aims to identify the most influential words for the classification of instances, particularly those that contributed the most to the correct predictions performed by the model. Unlike related work (ABNAR; ZUIDEMA, 2020; CHEFER; GUR; WOLF, 2021; VIG, 2019; KOKALJ et al., 2021; AYOUB; YANG;

Figure 5.4: Collection of the evaluation data



ZHOU, 2021), this identification is made at model-level by aggregating the instances collected attention weights in the previous phase of the framework. The tasks in this phase are:

- Aggregation of tokens' attention (Figure 5.5);
- Calculation of model-level word attention (Figure 5.6);
- Identification of IWs and PIWs (Figure 5.7).

The inputs are the tokenized tweets of the test dataset, accompanied by their respective attention weights and the predicted labels for each instance, obtained as described in Section 5.2.2. This process can also be followed through the running example in Figure 5.2.(b) from bottom to top.

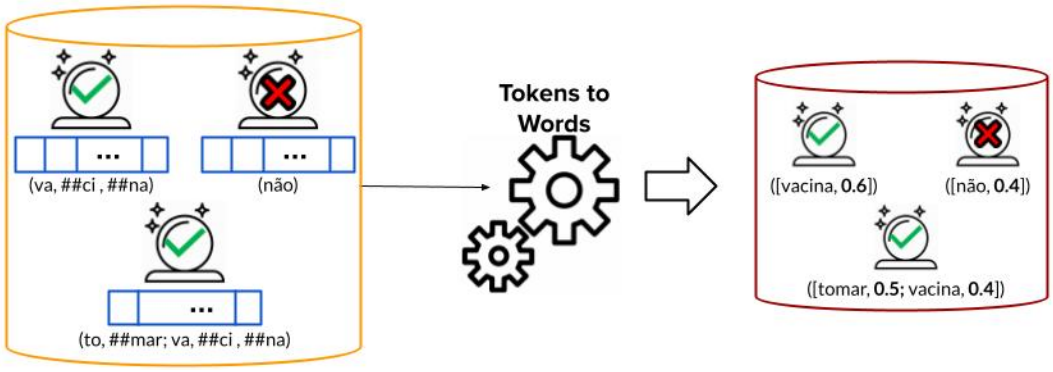
5.3.1 Aggregation of tokens' attention

First, it is necessary to aggregate tokens into the words from which they were initially extracted. Also, each recomposed word should have an attention weight calculated from the attention weights of those tokens that compose them. Figure 5.5 illustrates this process. This step is necessary because, contrary to words, tokens are not necessarily meaningful. Their relevance to the domain can only be analyzed when the context of the word containing them is identified. In addition, the same token can be part of more than one word. For instance, none of the BERT pre-trained models for the Portuguese language contained the word "vacina" (vaccine), and tokens such as "va" or "ci" were assigned high attention weights. Thus, each tweet calculates the attention of its words based on the average of the tokens composing each one of its words. Considering the

tokenized tweets of the running example (Figure 5.2.(a)), tokens are traced back to the original words (“vacina", “mata"), using the token’s average attention weight in the respective tweets (bottom of Figure 5.2.(b)).

Notice that the scores assigned to words are restricted to their respective tweets (i.e., there are not model-related scores for each word at this point). Consider the example in Figure 5.5, in which there are three tweets, two of them using the word “vacina”. This word has a particular score in each tweet, based on the attention weights assigned to its tokens in each prediction.

Figure 5.5: Tokens to words: attention aggregation



5.3.2 Calculation of model-level word attention

In order to evaluate the model’s behavior in terms of the most influential words, it is necessary to aggregate the instance-level attention weights of each word into a single word-related attention weight regarding the entire test dataset. This process is depicted in Figure 5.6. We propose the metric *Absolute attention* (AA), which uses the average of the words’ attention weights in all of the instances of the test dataset, multiplied by one hundred. High values of absolute attention could be due to the representativeness of words in terms of frequency, which is assumed as important for classification pattern identification[1].

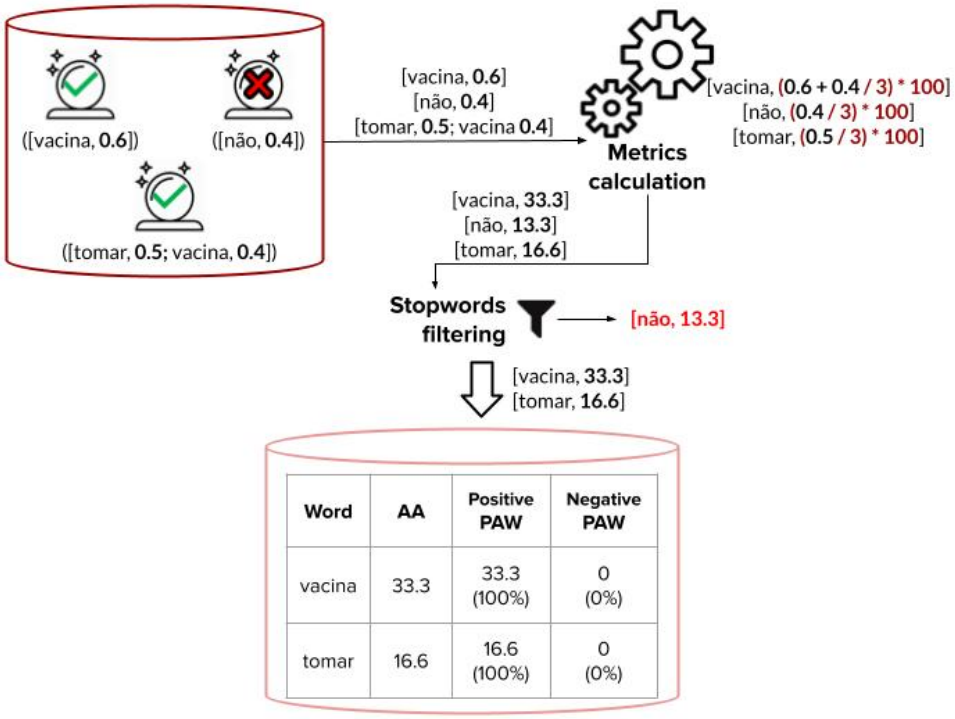
Another proposed metric is the *Proportional Attention Weight* (PAW), used to evaluate the contribution of words in the incorrect and correct predictions. The PAW for correct predictions (i.e., Positive PAW) considers the sum of the attention weights only in the correctly classified instances divided by the number of predictions multiplied by one hundred. Likewise, the PAW for incorrect predictions (i.e., Negative PAW) only considers

[1] In a preliminary work (SÁENZ; BECKER, 2021) we also experimented with *Relative attention* using only the tweets that contained the words, but it did not achieve acceptable results

the sum of weights in the misclassified instances. If the Positive PAW of a given word meets some threshold, we can assert it contributes to correct classification. Otherwise, despite its high value, it confuses the classifier, contributing to the misclassification of instances in practice. Section 5.3.3 further discuss this threshold and how it is calculated.

Given that high AA scores can be assigned to stop words, and those do not contribute to the model's interpretability, they should be removed using some pre-defined list (e.g., NLTK Python library[2]). In the running example of Figure 5.2.(b), it is possible to see that words "não" e "vou" are discarded. Notice that the word "vacina" is the one with the greatest AA with a score of 21. Its Positive PAW is 8, and its Negative PAW is 13, adding up to its AA.

Figure 5.6: Calculation of model-level word attention



5.3.3 Identification of IWs and PIWs

The IWs (i.e., Influential Words) are the words with the highest AA values. They are *influential* as they received the highest aggregated attention weights (i.e., AA) among all the words in the test set by the BERT model for stance classification. In the example in Figure 5.2.(b), "vacina" is the IW that has the highest absolute attention. The IWs are influential words for the model's decision-making, despite not necessarily positively in-

[2]https://www.nltk.org/

fluencing the correct classification, since it is possible to find them predominantly among wrong predictions.

A key issue is whether there is a minimum value for the AA of a word to consider it positively influences the correct predictions. The previously described PAW metric defines the proportion of the AA in correct and incorrect predictions and can help determine which words contributed positively to the correct classification or negatively to misclassification. However, given that Positive and Negative PAW add up to the AA, it is necessary to determine in what proportion, at least, should the Positive PAW be part of the AA to state that the word has a positive influence. This measure of minimum Positive PAW tolerance can be considered as a threshold. We named this threshold the *Minimum Positive Attention* (MPA).

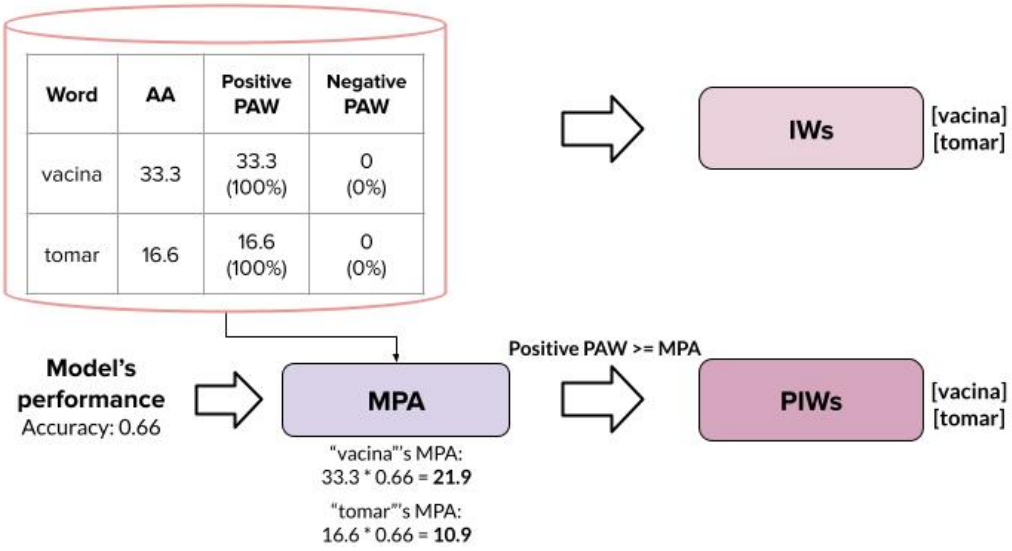
The MPA is different for each word, given that each words' attention metrics (e.g., AA, Positive/Negative PAW) are also different. It also needs to consider the distribution of the correct and incorrect predicted instances. Thus, it should be defined according to the properties of the model as a whole. In that sense, the threshold for a model with good performance should be higher compared to models with inferior performance. Therefore, the MPA is defined according to a performance metric of the model, such as the F1 measure or accuracy.

Consider the example in Figure 5.7, where the model has accuracy of 66%. Hence, the MPA of "vacina" and "tomar" are 21.9 and 10.9, respectively. Those values are obtained by taking the proportion defined by the accuracy (0.66) of the AA of those words (33.3 for "vacina" and 16.6 for "tomar"). Words with a Positive PAW that equals or exceeds this value are called *Positive Influential Words* (PIWs).

Back to the running example of Figure 5.2, we observe the model has an accuracy of 50%, and hence the MPA for each word is 50% of the AA. In this example, only the word "tomar" is considered a PIW, given that its PAW (4.3) exceeds its MPA threshold (2.15). The word "vacina", although being an IW with high AA and meaningful in the domain, proportionally contributed more to the misclassification.

By identifying IWs and PIWs, users can better understand the decisions made by BERT models for stance classification and determine which terms characterize and differentiate the stances the most. The IWs are the most influential words for the classifier model, helping it make the predictions. At the same time, the PIWs are the most influential words for correctly classifying instances, helping to distinguish between the polarized classes.

Figure 5.7: Identification of IWs and PIWs



5.4 Proof of concept

This section illustrates how a user could use the proposed metrics to assess the words' influence on the correct predictions of stances. Consider a dataset representing stances on COVID-19 vaccination, like the one described in Section 4.3. The user, a person developing a stance classification model, has to perform the following tasks:

- prepare a training/validation dataset to fine-tuning a stance classification model, as described in Section 5.2.1;
- prepare a test dataset and collect the evaluation data, as described in Section 5.2.2.

Figures 5.8, 5.9 and 5.10 presents some visualization tools that could be explored by the user to assess and interpret the model. The user can select the number of words to assess in all cases. In our examples, the user has selected the top-10 words according to the metrics examined.

Figure 5.8 presents a stacked horizontal bar chart with the top-10 IWs in this case study. Each bar represents a word's AA divided by its respective Positive PAW (blue) and Negative PAW (red). A black square inside each bar represents the MPA threshold. In this way, the user can identify the PIWs by looking for the bars where the MPA threshold square is within the blue region. For example, in the Figure, the word "tomar" (take), despite being an IW with the highest AA, does not positively contribute to the predictions, as its MPA threshold square is inside the red region of the bar. All the other words are considered PIWs, such as "presidente" (president), "china", "foradoria" ("Doria out"),

"vachina" ("vacchina"), etc. These results are consistent with our case study explained in Section 4.3.

Figure 5.8: Stacked horizontal bar chart visualization

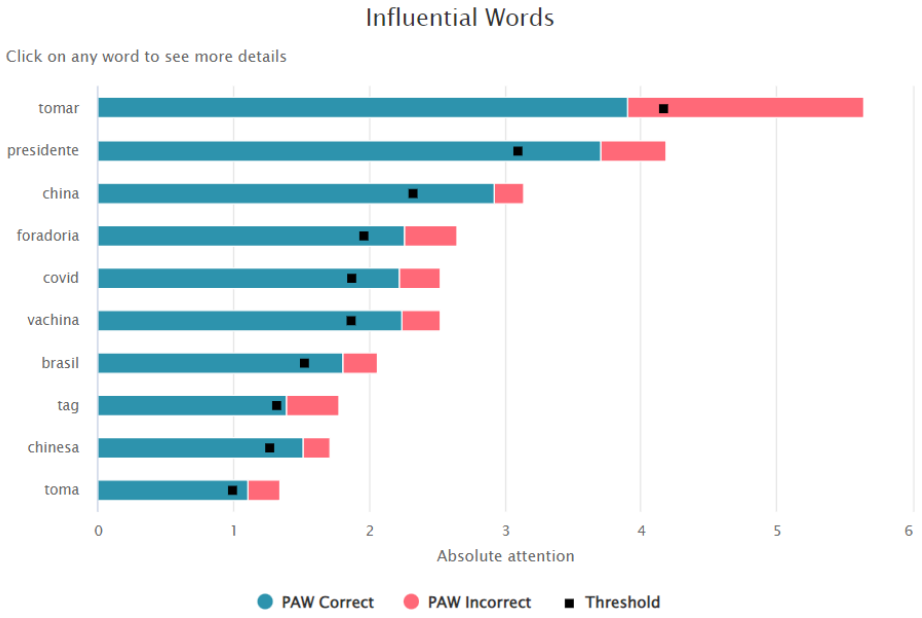
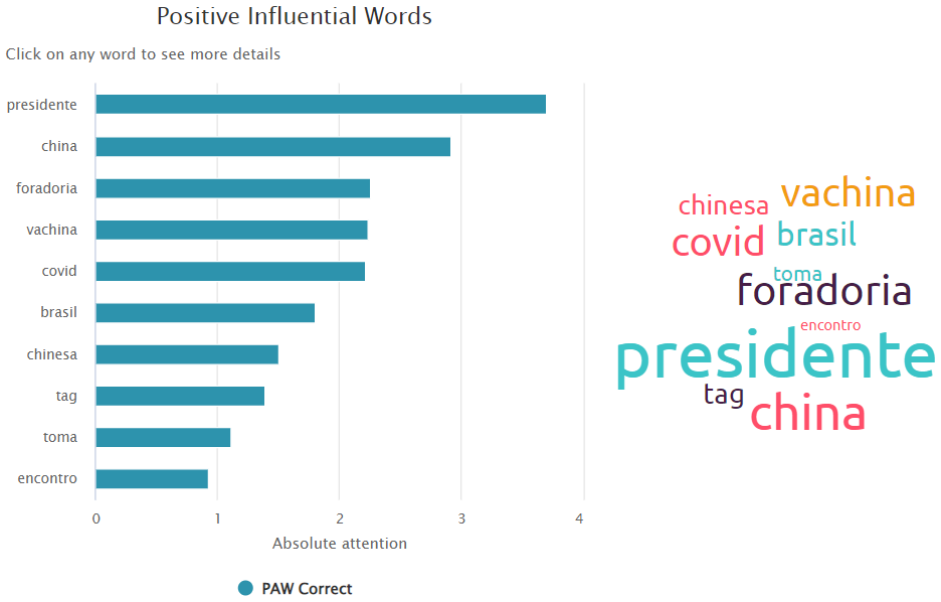


Figure 5.9 presents the top-10 PIWs in the case study. It shows two graphs, a horizontal bar chart containing each word with its corresponding Positive PAW and a word cloud of the PIWs presented in the graph. Both visualizations allow the user to compare PIWs' Positive PAW and identify the ones more related to the stances. In particular, the word cloud allows the user to focus on the words without dealing with the associated scores. In the example, "presidente" is the PIW with the highest Positive PAW with a great difference against the rest. The next graph will present a more in-detail view of this word's attention.

The user can inspect a word in more detail by selecting it from the horizontal bar chart. Figure 5.10 presents the details of the words "presidente", "eunaovoteinostf" ("I did not vote for the STF"), and "covid" in terms of the metrics AA and PAW. In all the figures, at the top, it can be seen the raw score and proportion of Positive/Negative PAW, which adds up to its AA. On the right, there is an icon and a message highlighting if the word is or is not a PIW. In all these cases, the icon is a checkmark as these words are, in fact, PIWs. Finally, a pie chart displays the AA distribution in the predicted stances at the bottom left of the figures. This distribution is calculated by adding the words' attention obtained on instances predicted as each stance and comparing the result with the total obtained. For the word, in Figure 5.10.(a), "presidente" it can be noticed that most of its AA, by far, is obtained in the instances predicted as *Anti-sinovaxxers* meaning that this word is closely

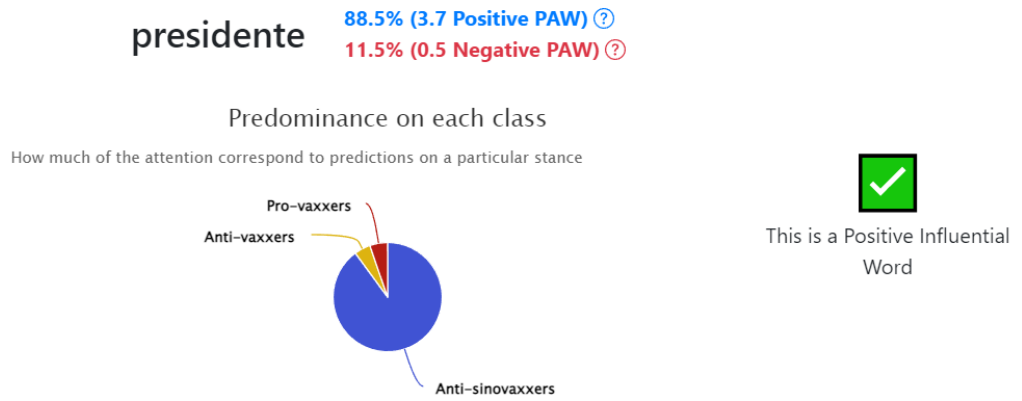
Figure 5.9: Horizontal bar chart and word cloud visualizations



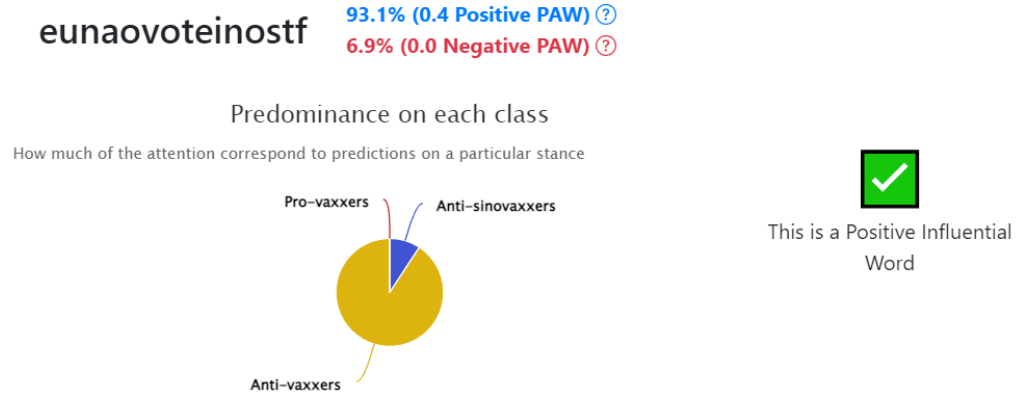
related to this stance. This is consistent with the case study, as *Anti-sinovaxxers* support and praise the president. For the word "eunaovoteinostf", in Figure 5.10.(b), most of its AA is obtained from the *Anti-vaxxers* stance, which also suits this stance's argument of rage against STF ruling (constitutionality). The word "covid", in Figure 5.10.(c), receives most of its AA from the *Pro-vaxxers* stance, which matches the *Pro-vaxxers* argument of praise for science by mentioning this word in their discussions.

The visualizations proposed in this section illustrate the use of the metrics as proof of their value for interpretability. Many other alternative visualizations and local interpretations can be derived from our framework to produce additional insights.

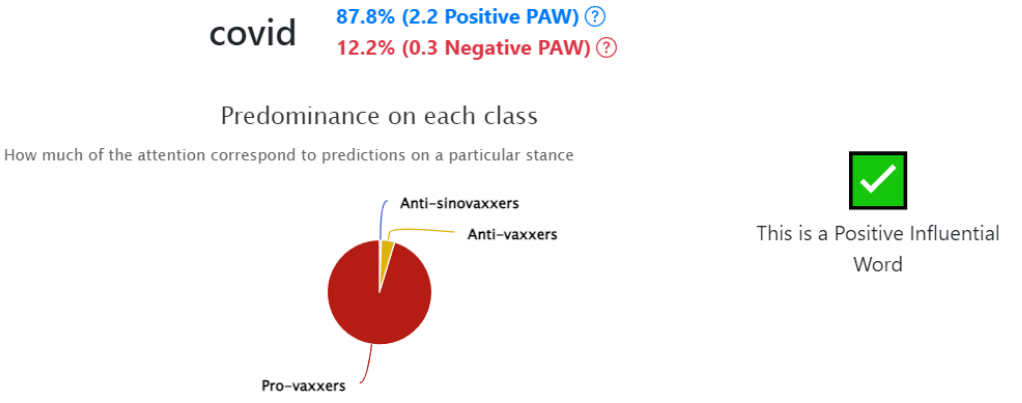
Figure 5.10: Word details visualization



(a) "presidente"



(b) "eunaovoteinostf"



(c) "covid"

6 EXPERIMENTS

This chapter describes in detail the experiments developed to assess our proposed framework. It enumerates the objectives defined for the proposed objectives, describes the datasets and models used, and explains the general method followed in all the experiments. We also detail how each research question was addressed and the derived results.

6.1 Objectives

We designed five experiments to evaluate our proposed interpretability framework for BERT-based stance classification. Each experiment answered a research question by addressing a specific aspect of the framework. The experiments, their related research question, and their objectives are:

- *Experiment #1: Does the BERT pre-trained model influence the results?* The use of attention weights as the basis for interpretability can only be reliable if the results generalize and are similar across multiple pre-trained models. We investigated this issue considering the multiple options available for the Portuguese language.

- *Experiment #2: Do the words with the highest absolute attentions (IWs) contribute to the correct predictions?* The interpretability framework can be useful only if it helps identifying and assessing the words that influence the correct predictions. We investigated the relationship between (high) attention scores and correct classification.

- *Experiment #3: Are the IWs representative of the domain and stances?* The IWs can only be used as the basis for interpretability if they are representative and meaningful concerning the stances expressed, contributing with insights for understanding the model's predictions.

- *Experiment #4: Does the vocabulary in the BERT pre-trained model affect the quality of the results?* BERT pre-trained models depend on the dictionaries, which dictate if a word needs to be broken down into smaller tokens. The interpretability framework traces tokens back to the original words whenever necessary so that the results should be comparable regardless of the completeness of the dictionary used.

- *Experiment #5: How does the proposed interpretability framework compare to Captum's Sequence Classification Explainer?* We believe that our interpretability framework using attention weights is valuable if its results are at least comparable to the results obtained using the *Sequence Classification Explainer* method of the Captum package.

6.2 Datasets and Models

We derived six datasets from the ones described in the case studies (Chapter 4). The datasets were constructed using a random sample of the tweets from each case study and were preprocessed by performing the actions listed in Section 5.2.1. Due to computational reasons and to maintain comparable results, each derived dataset has 6000 instances evenly distributed in the classes. By deriving more datasets, the results obtained can be further generalized.

For the Vaccination case study (Section 4.3), three variations of the original dataset were produced:

- **V-DS1**: contains a sample of the three stances presented in the original case study;
- **V-DS2**: contains a sample of two stances, namely the Pro-vax stance and the "anti-Chinese vaccine" stance (Anti-Sinovax);
- **V-DS3**: contains two stances, namely the Pro-vaxxers and the combination of Anti-Sinovax and Anti-Vax stances. This choice was made since there are similar arguments used in both of them (EBELING et al., 2022).

For the Social Isolation case study (Section 4.2), two variations from the original dataset were produced:

- **SI-DS1**: contains a sample of the three stances presented in the original case study;
- **SI-DS2**: contains a sample of two stances, namely the Chloroquinners and Quarentineers, discarding the neutral class, which was originally created for control purposes (EBELING et al., 2022).

The Hydroxychloroquine case study had only one derived dataset (**H-DS**) containing the three stances already described in Section 4.4.

Each derived dataset was randomly divided into three subsets: training set (72%), validation set (8%), and test set (20%). The former two were used for the training process of the stance classification models, and the test set served to address the research questions.

Then, we developed 16 stance classification models by fine-tuning pre-trained BERT models. Models were chosen for the English and Portuguese languages, as detailed in Section 2.1.1. Large models (e.g., BERT large) were not adopted as their comparison would not be fair with models created with less computational resources.

Table 6.1 details the datasets, stances, number of tweets, and the pre-trained models used for each case study. All the code related to the data preprocessing and each of these experiments is available in a public repository[1].

Table 6.1: Datasets and pre-trained models used

Datasets compositions				**BERT pre-trained Models**			
Case Study	**Dataset**	**Stance**	**# Tweets**	**BERTimbau**	**BERT M. Cased**	**BERT M. Uncased**	**BERT**
Vaccination	V-DS1	Anti-sinovaxxers	2,000	✓	✓	✓	
		Anti-vaxxers	2,000				
		Pro-vaxxers	2,000				
	V-DS2	Anti-vaxxers	3,000	✓	✓	✓	
		Pro-vaxxers	3,000				
	V-DS3	Anti-sinovaxxers ∪ Anti-vaxxers	3,000	✓	✓	✓	
		Pro-vaxxers	3,000				
Social Isolation	SI-DS1	Chloroquinners	2,000	✓	✓	✓	
		Neutrals	2,000				
		Quarentineers	2,000				
	SI-DS2	Chloroquinners	3,000	✓	✓	✓	
		Quarentineers	3,000				
H-DS	H-DS	Anti-Chloroquine	2,000				✓
		Neutrals	2,000				
		Pro-Chloroquine	2,000				

6.3 General Method

A key issue in the experiments is determining a minimum absolute attention score that could be considered for a token/word to be influential in the model's classification. By examining the distribution of the words' AA for all the datasets and models, we observed that, in general, higher scores (AA greater than or equal to 0.5) could be identified only among the top 70 words, at most; otherwise, the values are quite low.

We also observed that the IWs/PIWs attention scores tend to decrease for different ranges similarly, regardless of the model and dataset under analysis. For example, con-

[1] https://github.com/cacsaenz/attention-based-interpretability

sidering a BERTimbau model trained on the V-DS1 dataset, the AA for the 50th, 100th, 250th, and 500th IW were 0.310, 0.200, 0.100, 0.057, respectively. These values for each range are very similar to the other models trained in the different datasets. Therefore, all the experiments were performed considering different ranges of top-n IWs/PIWs, where n could take values [50, 100, 150, 200, 250, 300, 350, 400, 450, 500].

Each experiment assessed specific baseline metrics to compare the proposed metrics and framework results according to the research questions. Spearman correlation test was chosen to identify trends and verify the statistical significance of the results due to the non-normal distribution of attention scores considering the distinct cut-off ranges, with significance level $alpha = 0.05$. These statistical comparisons allow us to confirm or refute the hypotheses in the experiments.

The ranking metrics described in Section 2.4.2 focus on sets of (relevant) documents. However, in the context of our work, they are applied to sets of (influential) words. Those metrics will be used in the following experiments to assess our proposed method.

The stop words filtering step used the list of stop words available in the NLTK library for Python.

6.4 Experiment #1: Does the BERT pre-trained model influence the results?

The first experiment seeks to determine whether a specific BERT pre-trained model influences the results obtained in terms of performance and IWs. This comparison is relevant considering the different pre-trained models for the Portuguese language. It is assumed that if the performances of different models and their IWs in the same dataset are similar, the proposed interpretability method is generalized, despite the BERT model used, and it can be used to interpret the predictions.

6.4.1 Method

Models' performance is assessed using the previously described metrics: accuracy and weighted-average precision, recall, and F1. Those metrics are calculated for all the trained models evaluated using the test subset of all the Portuguese datasets (V-DS1, V-DS2, V-DS3, SI-DS1, and SI-DS2).

The difference in the IWs was analyzed by comparing the percentage of common

IWs between every pair of models in each Portuguese dataset for the pre-defined cut-offs.

6.4.2 Results

Figure 6.1 presents the performance metrics results of all the evaluated models. We observed that, for all metrics, the results are very similar among the models trained on the same dataset, and no BERT pre-trained model consistently outperformed the others. In addition, the performance of the classification models fined-tuned using BERT pre-trained models to address the Portuguese language is not very different from the one using the English BERT pre-trained model.

Figure 6.2 shows the intersections of the IWs between all pairs of combinations of the three models in the five Portuguese datasets, considering the different ranges for top-n IWs (i.e., top-50, top-100, ..., top-500). In general, we observed a significant intersection, with values ranging from 46% (BERTimbau $\cap$ BERT-M uncased in the V-DS2) up to 71% (BERTimbau $\cap$ BERT-M cased in the SI-DS1). The medians variate between 53% (BERTimbau $\cap$ BERT-M uncased in the SI-DS2) up to 64% (BERTimbau $\cap$ BERT-M cased in the SI-DS1). The intersection is higher between BERTimbau, and BERT-M cased because the former was trained in a cased corpus.

6.4.3 Discussion

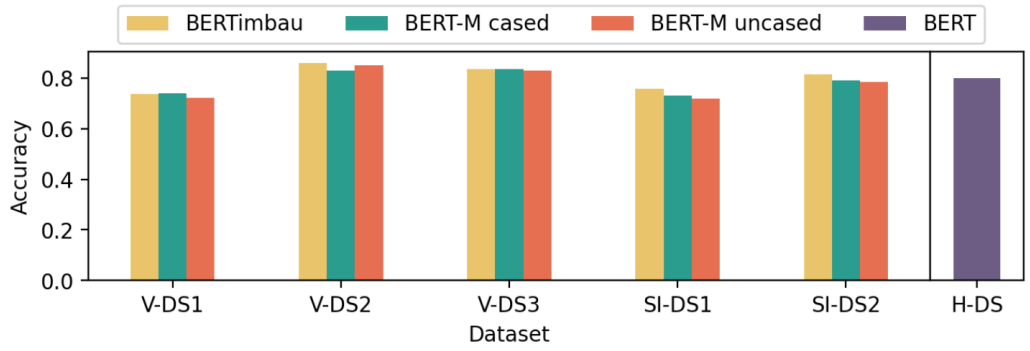
This experiment revealed that, despite the differences in the attention weights assigned by each model, they do not influence the models' performance. Also, there is a high proportion of common IWs between the different pairs of models. Thus, it is possible to state that the differences in pre-trained models tend not to create distortions.

6.5 Experiment #2: Do the words with the highest absolute attentions (IWs) contribute to the correct predictions?

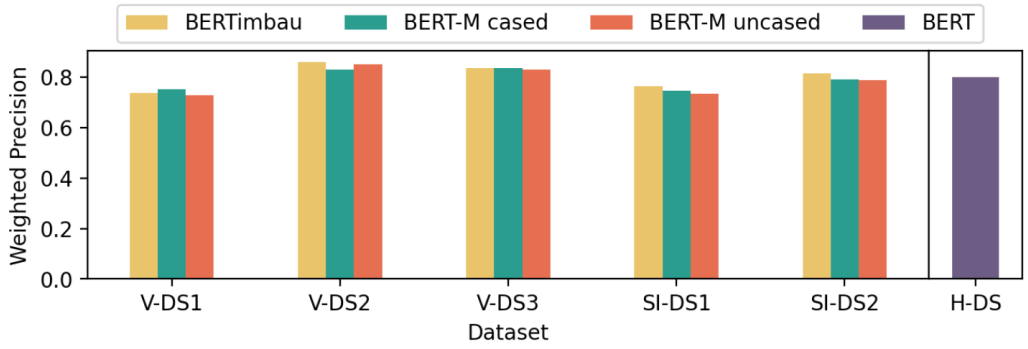
This second experiment determines if the IWs contribute to the correct predictions. The intuition is that there is an influence of the AA value of the words on correct predictions. We used two methods to make this assessment, detailed in the remaining of this section:

(a) the positive influence of IWs in the correct predictions made by the stance classification model (PIWs);

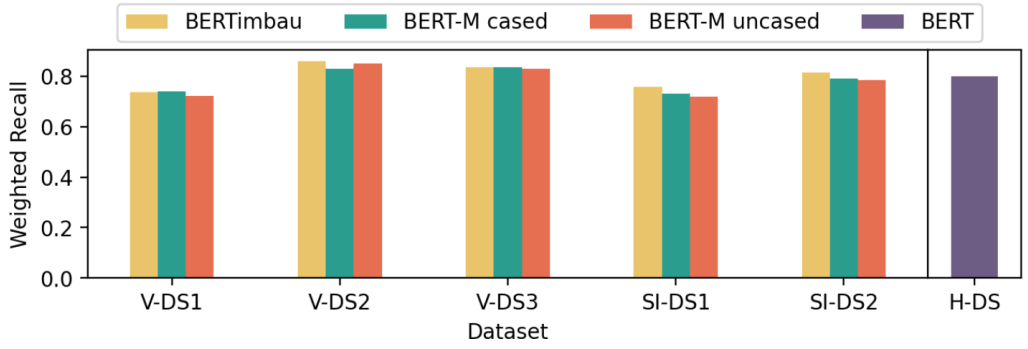
Figure 6.1: Performance metrics results for the models on each dataset



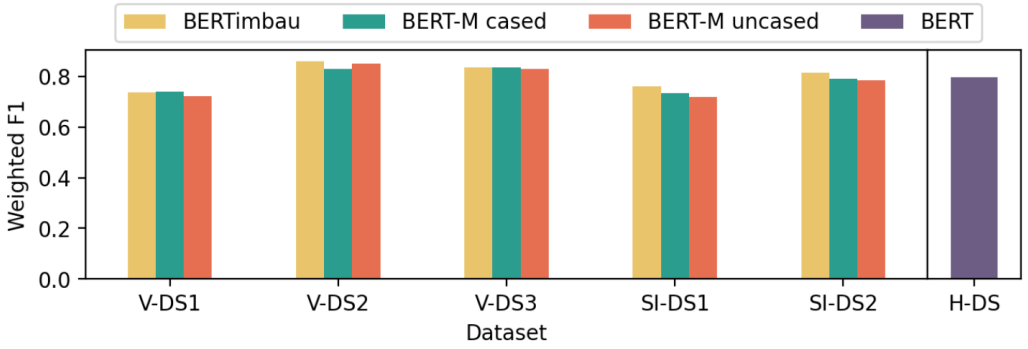
(a) Accuracy



(b) Weighted Precision

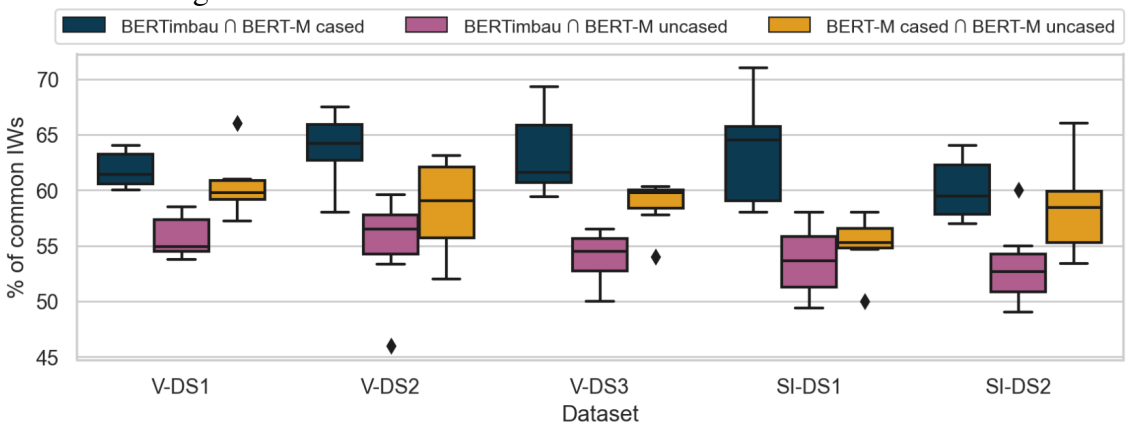


(c) Weighted Recall



(d) Weighted F1

Figure 6.2: IWs intersections for different cut-offs and models



(b) the influence of the IWs/PIWs on the prediction performance using the Leave-One-Out (LOO) technique.

6.5.1 Positive influential words (PIWs)

6.5.1.1 Method

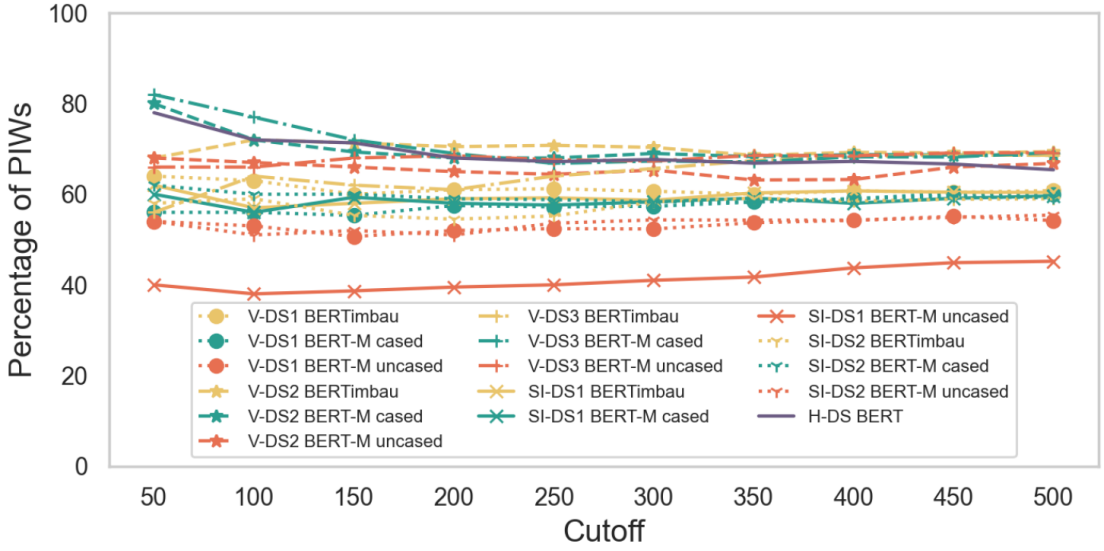
As mentioned in Section 5.3.3, defining a minimum Positive PAW of the AA to consider an IW also a PIW is a challenge. Thus, we proposed the concept of MPA to allow us to find PIWs. This assessment seeks to determine how predominant are the PIWs in the different cut-offs of IWs and if high AA scores can be associated with a higher contribution to the correct classification. For that purpose, for each IW, we calculated the PAW in correctly predicted instances. Then, we verified if it exceeds the MPA threshold based on the respective model's performance so that the IW can also be considered a PIW. Considering the similar values for the performance metrics due to the balanced datasets in our experiments, the *accuracy* of each model was the metric used to establish the MPA threshold.

We performed this analysis considering the cut-off ranges in each model to verify a minimum attention value for such a relation and generalize the observed behavior. Finally, the different cut-offs for each model were used to calculate the Spearman correlation between the percentage of PIWs and the respective average word's AA.

6.5.1.2 Results

Figure 6.3 shows the percentage of PIWs for all models across the different cut-offs. We can observe that the percentage is above 50% in all models with a single exception (SI-DS1 Bert-M uncased). Otherwise, the percentage ranges from 51% (top-150 V-DS1 BERT-M uncased) up to 82% (top-50 V-DS3 BERT-M cased). Also, we can notice that the percentage variation is stable by the top-150 cut-off. Table 6.2 shows the correlation analysis per model, where statistically significant correlations are highlighted in gray. We observe that there is not a pattern of statistically significant correlation between the average AA score of a cut-off and the percentage of PIWs.

Figure 6.3: Percentages of PIWs in different cut-offs, models and datasets



Hence, despite the significant proportion of IWs that are also PIWs, it is necessary to refute the hypothesis that the higher the absolute attention, the more a word contributes to correct classifications. However, PIWs are the most important words for correctly classifying instances and the most valuable ones to understand what differentiates stances. Finding high percentages of these words in sets of IWs implies that, at least proportionally, IWs tend to be PIWs. More insights were obtained in the subsequent assessments.

6.5.2 Positive Leave-One-Out words (PLOO)

6.5.2.1 Method

The second assessment involved the LOO technique, used to assess feature influence in classifications in related works (JAIN; WALLACE, 2019; WIEGREFFE; PIN-

Table 6.2: Spearman correlation results between IWs Avg. AA and Percentage of PIWs

Dataset	Model	Correlations in the different cut-offs
		IWs Avg. AA vs. Percentage of PIWs
V-DS1	BERTimbau	0.4924 (p=0.1482)
	BERT-M cased	-0.9119 (p=0.0002)
	BERT-M uncased	-0.5758 (p=0.0816)
V-DS2	BERTimbau	0.3333 (p=0.3466)
	BERT-M cased	0.4985 (p=0.1425)
	BERT-M uncased	0.3951 (p=0.2584)
V-DS3	BERTimbau	-0.9240 (p=0.0001)
	BERT-M cased	0.5273 (p=0.1173)
	BERT-M uncased	-0.8537 (p=0.0017)
SI-DS1	BERTimbau	-0.3697 (p=0.2931)
	BERT-M cased	-0.1033 (p=0.7763)
	BERT-M uncased	-0.9058 (p=0.0003)
SI-DS2	BERTimbau	-0.4559 (p=0.1854)
	BERT-M cased	0.4756 (p=0.1647)
	BERT-M uncased	-0.8085 (p=0.0046)
H-DS	BERT	0.9394 (p=0.0001)

TER, 2019). In short, it consists in performing multiple executions, each time removing a given feature (word) from all the instances of the test dataset and executing the classification to verify the performance variation on the results. If the performance degrades, the word contributes positively to the classification. In our work, we refer to these words as *Positive Leave-One-Out* (PLOO) words. Otherwise, if the performance improves or is maintained, the word has a negative contribution or is irrelevant for correct predictions.

We use the accuracy metric as the performance measure in our analysis. The hypothesis is that a predominance of PLOO words within the IWs/PIWs subsets would reveal that the interpretability framework finds influential words for the model's performance.

This technique was applied to all models, considering the IWs/PIWs for all the cut-offs, obtaining the respective percentage of PLOO words in each case. Then, we analyze the relationship between the percentage of PLOO words and the average word attention within each cut-off using the Spearman correlation test.

6.5.2.2 Results

Figure 6.4 shows the percentages of PLOO words obtained in different cut-offs for different datasets and models. It is possible to observe that the percentage of PLOO words is consistently higher for all the models in the top-50 cut-off, ranging from 68% (V-DS1 BERT-M cased) down to 40% (V-DS2 BERT-M uncased). These values tend to

Figure 6.4: Percentage of PLOO words in different cut-offs, models and datasets

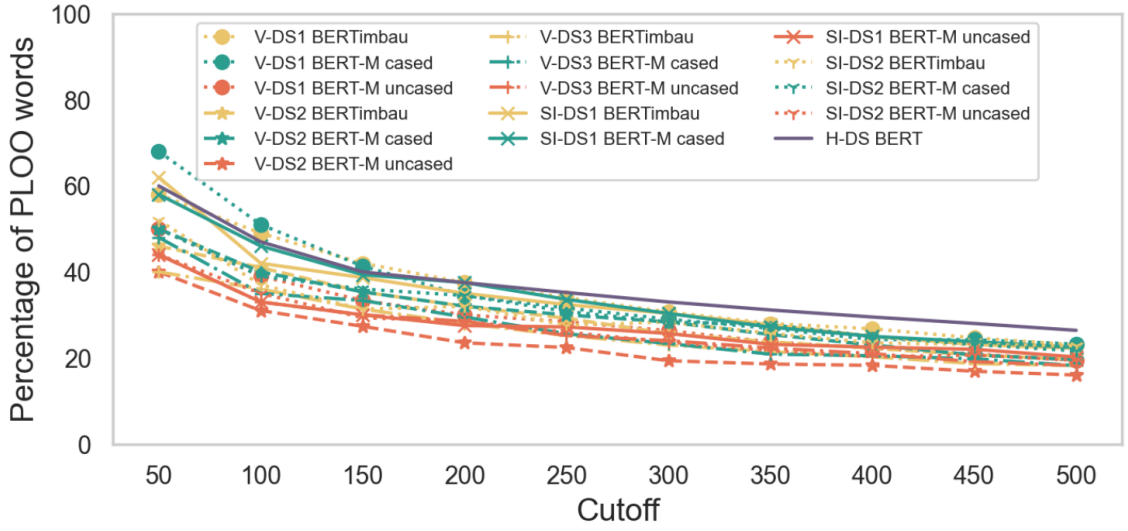


Table 6.3: Spearman correlation results between IWs/PIWs Avg. AA and Percentage of PIWs

Dataset	**Model**	**Correlations in the different cut-offs**	
		IWs Avg. AA vs. Percentage of PLOO	**PIWs Avg. AA vs. Percentage of PLOO**
V-DS1	BERTimbau	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M cased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M uncased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
V-DS2	BERTimbau	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M cased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M uncased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
V-DS3	BERTimbau	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M cased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M uncased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
SI-DS1	BERTimbau	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M cased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M uncased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
SI-DS2	BERTimbau	0.9758 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M cased	1.0000 (p=0.0000)	1.0000 (p=0.0000)
	BERT-M uncased	0.9879 (p=0.0000)	1.0000 (p=0.0000)
H-DS	BERT	1.0000 (p=0.0000)	1.0000 (p=0.0000)

decrease as the number of words in the cut-off increases, revealing a relationship between absolute attention and the percentage of PLOO words.

Table 6.3 shows the results of the Spearman correlation test between IWs/PIWs AA scores and the percentage of PLOO words found. We found a strong positive correlation in both cases. This correlation is even stronger when considering only the average word attention of PIWs.

This assessment allows us to conclude that high word AA scores improve the probability of a positive influence on the correct classification, as well as the probability of a PIW being also a PLOO word.

6.5.3 Discussion

These assessments provided evidence that the IWs and, more specifically, PIWs, contribute to the correct classification of instances and, therefore, can help interpret the model's predictions. Although the hypothesis that high absolute scores are related to correct predictions (PIW) was refuted, the analyses revealed that the higher the score, the higher the probability of a word being a PLOO.

6.6 Experiment #3: Are the IWs representative in the domain and stances?

The third experiment assesses the representativeness of the IWs in the domain. If the IWs are domain-representative and meaningful in terms of the arguments used by the different stances, the interpretability framework can help understand the model's decisions and stances arguments. For this assessment, we used the TF-IDF index (term frequency-inverse document frequency), which measures the relevance of words in documents. We also used c-TFIDF, an adaptation of TF-IDF proposed in BERTtopic, to quantify the relevance of words per topic.

6.6.1 TF-IDF Influential Words

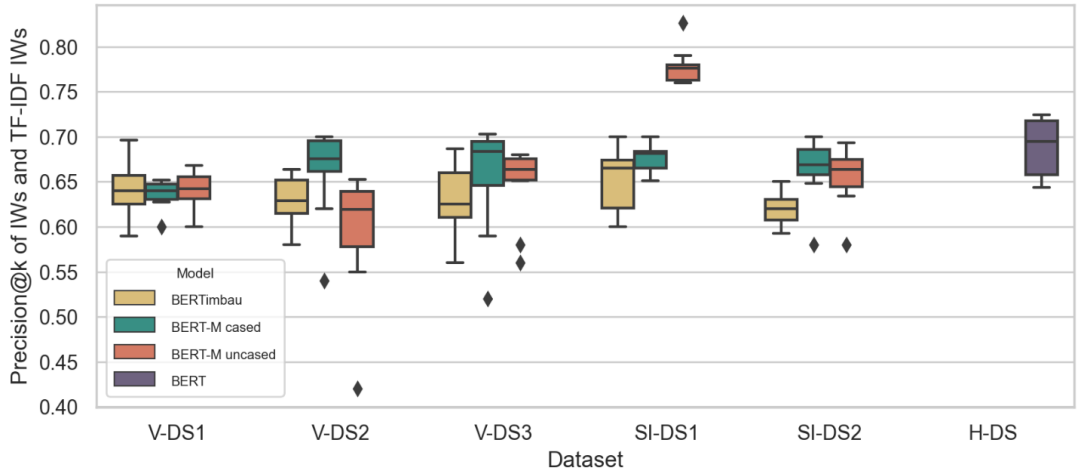
6.6.1.1 Method

The TF-IDF index is used to obtain the most influential words considering the whole dataset (TF-IDF IWs) and identify their alignment with the IWs. The hypothesis is that if the proposed method identifies domain-representative words, a significant alignment should exist.

TF-IDF assigns a score to each word in a document, considering a corpus that in our case are the sampled datasets. Hence, we averaged the TF-IDF index for each word considering all the instances where the word appeared in order to create TF-IDF IWs, and be able to compare them with the IWs.

We used the Precision@k and NDCG@k measures to compare the alignment of IWs with TF-IDF IWs. The proportion of IWs within TF-IDF IWs was assessed with Precision@k, where k represents different cut-off values. Also, we performed Spearman correlation tests to determine the relationship between the words' AA and TF-IDF IWs

Figure 6.5: Precision@k of IWs and TF-IDF IWs distributions for different ranges, models and datasets



scores using the average AA within each cut-off.

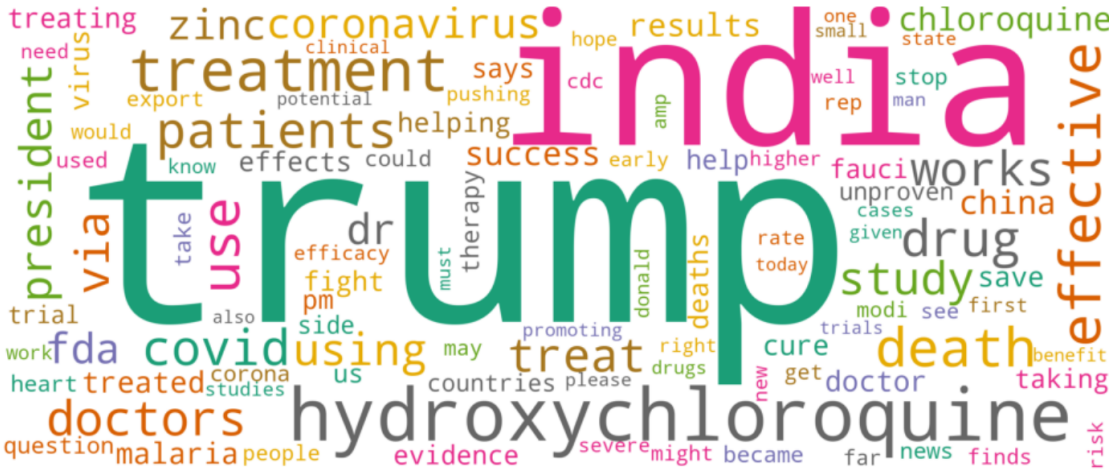
NDCG@k measures rankings of k elements. We used it in this experiment to compare the alignment of the AA scores in IWs with regard to TF-IDF IWs scores. We assessed a smaller range of rankings [10, ..., 100] with small steps of 10 (i.e., 10, 20, 30, ..., 100) since we wanted to evaluate the differences for meaningful/higher AA scores. Notice that NDCG@k does not penalize the absence of words in the rankings, possibly leading to bias. Therefore, we introduced a penalty by assigning a *score* of 0 to words that were not found within the TF-IDF IWs sets. The more the Values closer to 100%, the more similar the rankings are.

6.6.1.2 Results

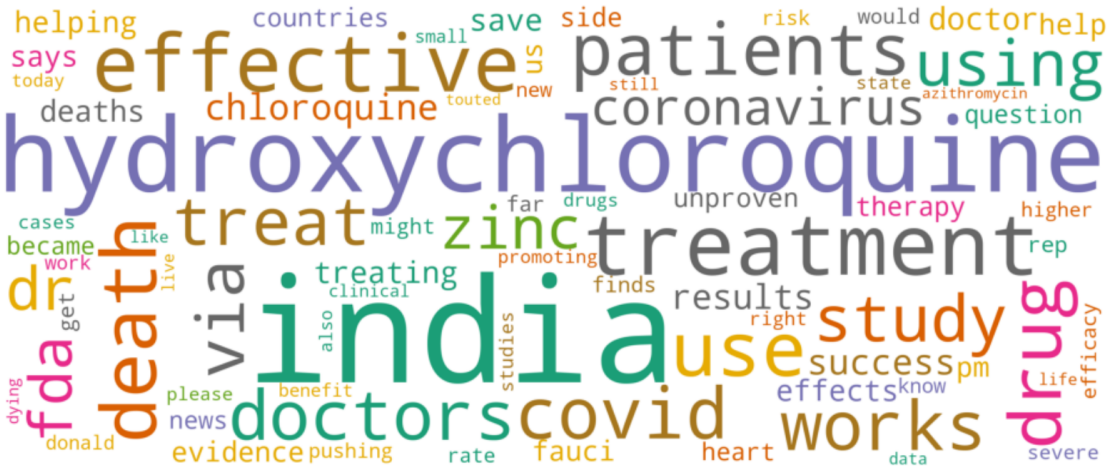
The boxplots in Figure 6.5 show the distribution of the Precision@k between IWs and TF-IDF IWs, per model. It is possible to observe that the medians are high, fluctuating between 0.63 (BERTimbau in SI-DS2) and 0.78 (BERT-M uncased in SI-DS1).

We illustrate the results using the Hydroxychloroquine case study (Section 4.4) using Figure 6.6. Figure 6.6.(a) presents a word cloud of the IWs that are common to the top-150 TF-IDF IWs for the stance classification model with the greatest Precision@k results (i.e., BERT model trained with the H-DS dataset). In this figure, we observe words aligned with the arguments summarized in Table 4.3. However, we identify that some words can be related to multiple stances at the same time. That is the case of "trump" (the biggest word in the figure), as well as "president" (in green, center left); both are terms used by Pro/Anti Chloroquine stances. There are also several words common to the three stances at the same time, such as "hydroxychloroquine" (in gray, bottom center),

Figure 6.6: Word cloud of top-150 IWs/PIWs intersected with TF-IDF IWs in the H-DS dataset using BERT



(a) Top-150 IWs and TF-IDF IWs

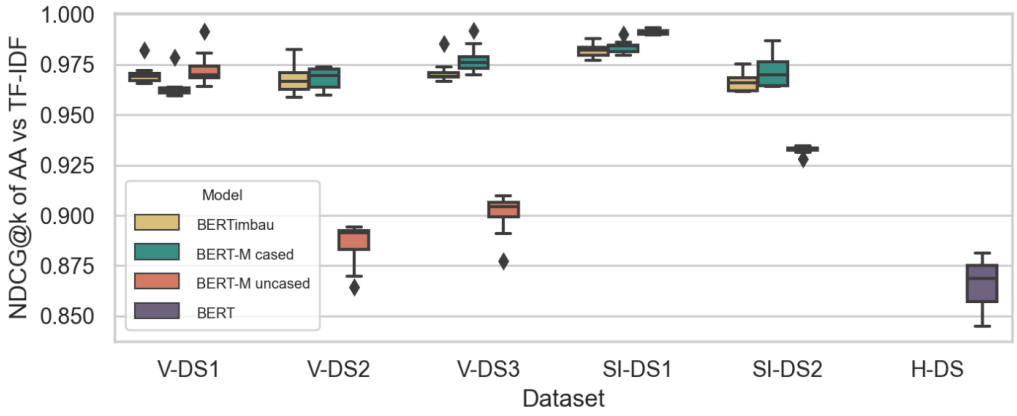


(b) Top-150 PIWs and TF-IDF IWs

"treatment" (in brown, top left), "covid" (in green, bottom left), "chloroquine" (in green, top right) and "drug" (in gray, center right). There are also words related to specific stances and particular events during the collection of the dataset, such as:

- "india" (in pink, top): Used by the Neutral stance, referencing the country from which the (hydroxy)chloroquine was imported;
- "death" (in yellow, bottom right), "unproven" (in gray, center right): Used by the Anti-Chloroquine stance when claiming that (hydroxy)chloroquine is dangerous for COVID-19 patients and has unproven value;
- "success" (in orange, top center), "clinical" (in orange, top left), "effective" (in orange, center right): Used by the Pro-Chloroquine stance claiming that clinical trials results have proven the effectiveness of the drugs;

Figure 6.7: NDCG@k of rankings based on AA vs. TF-IDF



- "fda" (in purple, bottom left), "fauci" (in pink, center right): Used by the Pro-Chloroquine stance when referring to the FDA and Dr. Anthony Fauci, who rejected the use of these drugs.

Figure 6.6.(b) presents the PIWs aligned with the top-150 TF-IDF IWs. Compared to the one in Figure 6.6.(a), this word cloud reveals that the alignment in this case is more restricted (i.e., there are fewer common words). However, we can identify that many terms used by different stances simultaneously do not appear anymore. That is the case of "trump", "president", "gov", "vaccine" or "virus", among others. These words may confuse the model by being used to develop pro, against, and neutral arguments, and they disappeared since PIWs are words that positively contributed to the correct classification. We also highlight that the previously described words related to specific instances in Figure 6.6.(a) are present in this word cloud also, as their contribution is positive to the correct classification.

Finally, Figure 6.7 shows the distribution of NDCG@k for each model, dataset, and range. The medians range from 0.86 (H-DS with BERT) to 0.99 (SI-DS1 with BERT-M uncased). This strong alignment between both scores and the previous Precision@k results allows us to conclude that although a few words are not influential according to both metrics, the way those metrics rank the words is very similar.

6.6.2 BERTopic words

The works used as case studies have applied BERTopic to understand the pro/against stances related to the COVID-19 context, which allowed us to obtain a deep familiarity with the arguments and terms used by the distinct groups. Thus, this experiment uses BERTopic's c-TFIDF metric to expand the analysis on the representativeness of

the IWs with regard to the obtained topics. c-TFID is a class-based version of TF-IDF, in which it considers the number of classes (or topics) rather than the number of documents. We assume that if the IWs are aligned to the words that represent the topics the most in each stance, the proposed interpretability framework could identify words related to the arguments expressed by the different stances.

6.6.2.1 Method

We gathered a list of relevant Topic Words (TWs) from the complete original datasets described in Chapter 4. We used the original datasets since using the sampled ones could introduce a bias in the results by limiting the domain to what was expressed in the sample and reducing the number of topics represented. Hence, for each original dataset, the following steps were performed:

a) for each stance, we applied BERTopic to obtain the 10 most important topics;

b) from each topic, we selected the top-10 most relevant words according to c-TFIDF, excluding NLTK stopwords;

c) for each stance in each dataset, we combined the top-10 words of each topic, obtaining a single list per dataset.

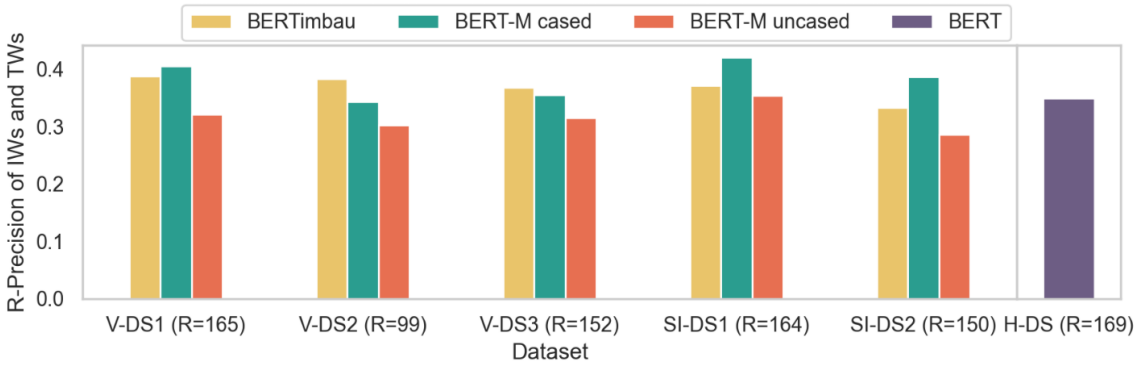
The number of topics was defined experimentally. We found out that 10 topics best represented the datasets. We chose 10 words per topic because this is the maximum number of words per topic recommended by BERTopic authors to gather relevant words.

We obtained a total of 165 TWs for V-DS1, 99 TWs for V-DS2, 152 TWs for V-DS3, 164 TWs for SI-DS1, 150 TWs for SI-DS2, and 169 TWs for H-DS using this method. Since there were different counts of TWs for each dataset, we chose the R-Precision metric to analyze how many of the IWs were also TWs, considering R equal to the number of TWs in the dataset. As previously described, high values of R-Precision indicate strong alignment with the arguments expressed by the stances.

6.6.2.2 Results

Figure 6.8 displays the distribution of R-Precision results for the different datasets and models. We can observe that BERT-M uncased was consistently outperformed by the other models. We find reasonable R-Precision values for the different models and

Figure 6.8: R-Precision of IWs vs. TWs for each model and dataset

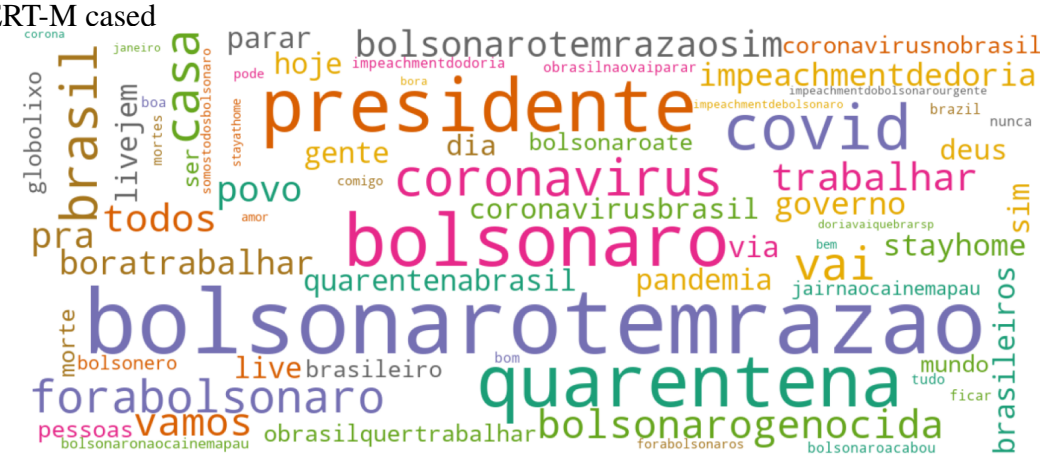


datasets, ranging from 0.29 (BERT-M uncased in SI-DS2) to 0.42 (BERT-M cased in SI-DS1). Although the R-Precision values are not high, particularly if compared to the ones obtained using TF-IDF IWs, we regard them as acceptable results since the TWs are words associated exclusively with each polarized class's top-10 most relevant topics in the original dataset. Hence, many other arguments used to express stances may not be well represented in the sample datasets from where the IWs were extracted, preventing a more accurate assessment of the IWs that are also TWs.

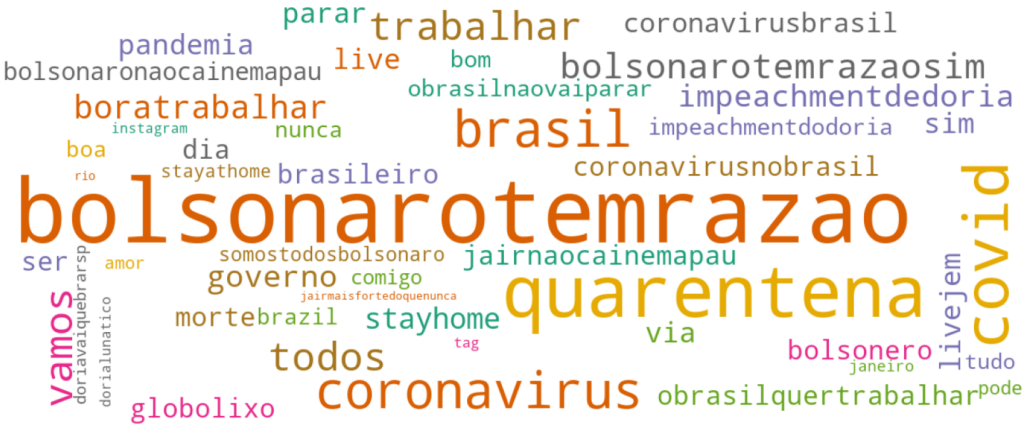
We adopted the Social Isolation case study (Section 4.2) to interpret the encountered results. The word cloud in Figure 6.9.(a) presents the IWs that are also TWs for the stance classification model with the greatest R-Precision (i.e., BERT-M cased in the SI-DS1 dataset). Similar to the previous experiment, we identify several words aligned with the arguments of the stances of this case study, detailed in Table 4.2:

- "presidente" (in orange, top center) and "bolsonaro" (in pink, center) are words referencing the president Jair Bolsonaro, who was a central part of the discussion about the social isolation measure;
- "bolsonarotemrazao" (in purple, bottom center), "bolsonarotemrazaosim" (in grey, top center) and "boratrabalhar" (in brown, center left) are hashtags used by the *Chloroquinners* to manifest support to the president;
- "forabolsonaro" (in purple, bottom left) and "bolsonarogenocida" (in green, bottom center) are hashtags used by the *Quarenteners* as a sign of rejection of the Brazilian president;
- "impeachmentdodoria" (in orange, top right) and "impeachmentdeedoria" (in pink, top center) are words rejecting the prospective presidential candidate João Dória, used mostly by the *Chloroquinners*;

Figure 6.9: Word cloud of the IWs/PIWs that are also TWs in the SI-DS1 dataset using BERT-M cased



(a) IWs and TWs



(b) PIWs and TWs

- "coronavirusnobrasil" (in orange, top right), "quarentena" (in green, bottom right) and "covid" (in purple, top right) are words mostly used by the *Neutrals* as hashtags to receive more attention and discuss other measures and implications of the social isolation.

We also assessed the distribution of the PIWs that are also TWs in the SI-DS1 dataset using BERT-M cased, in the word cloud presented in Figure 6.9.(b). We can notice that this visualization lacks words related to many stances simultaneously, such as "presidente" and "bolsonaro" as they may confuse the model and contribute to misclassification. We also identify the words closely associated with the arguments of the case study, such as "bolsonarotemrazao" or "quarentena", among others.

6.6.3 Discussion

These experiments proved that the IWs and PIWs are also distinctive and representative in the domain. We found alignments between them and our proposed metrics despite the distinct premises used to quantify baseline domain representativeness metrics, such as TF-IDF and c-TF-IDF. The alignment is strong between our AA and TF-IDF importance. Although the low percentages of IWs that are also TWs, these are encouraging values due to the limited number of TWs we were able to collect (i.e., 10 words for each 10 topics of each stance) and the fact that they are obtained from the whole dataset and not only the sampled one. Hence, our interpretability framework can be used to highlight words influential for the stance classification and related to the domain and be able to understand the arguments expressed by the stances.

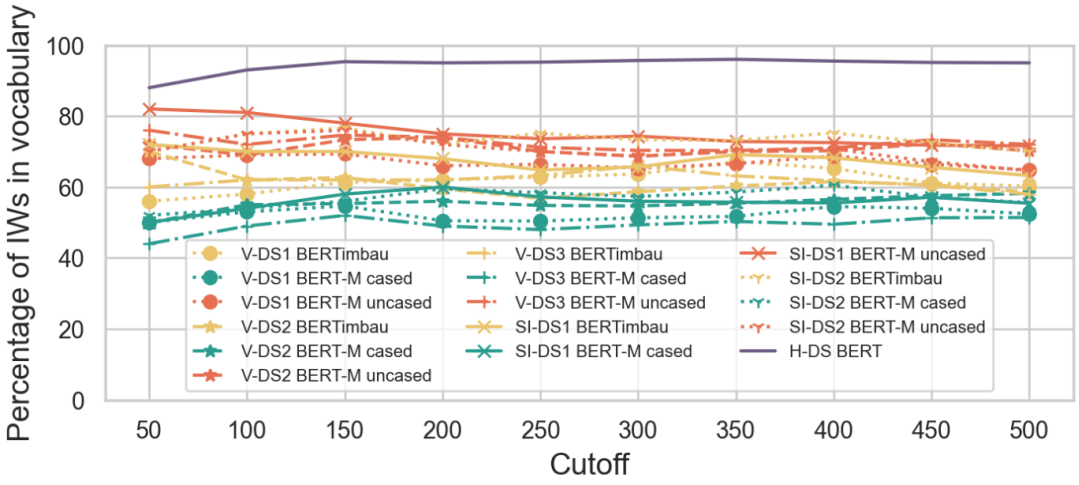
6.7 Experiment #4: Does the vocabulary in the BERT pre-trained model affect the quality of the results?

The proposed method in this research compensates for the lack of words in a BERT pre-trained models' vocabulary by tracing back tokens to the original words and consolidating the tokens' weights into word-related attention. This experiment assesses if the completeness of the vocabulary (i.e., the contents and size) of a BERT pre-trained model affect the quality of the results obtained. The hypothesis is that if the vocabulary does not influence the quality of the results, then the method for producing word-related attention scores is correct.

6.7.1 Method

We analyzed the proportion of IWs found in the vocabularies of the different BERT models, using the different cut-offs. Using correlation tests, we compared the average AA vs. the percentage of IWs in the model's vocabulary to determine if the greater the AA, the more probable an IW is part of the vocabulary. Then, we analyzed the correlation between the average AA of the IWs with the percentage of PIWs found among them to verify if the presence of a word in the vocabulary can influence the fact of it being a PIW. In both tests, the lack of correlation would indicate that the framework to

Figure 6.10: Percentage of IWs found in models vocabularies



recompose the words' attention is correct.

6.7.2 Results

Figure 6.10 shows the percentage of IWs that can be found in each model's vocabulary for different cut-offs, datasets, and models. It is possible to see that at least 40% of the IWs are present in the model's vocabulary as complete words and not as subtokens. We also observe that the English BERT pre-trained model has the highest percentages, while the BERT-M uncased model provided the most complete dictionary for the Portuguese language datasets. A clear upward/downward trend in the percentages cannot be identified as the cut-off size increases.

Table 6.4 presents the statistical correlation test results between the average AA and the percentage of IWs in vocabulary. It shows both positive and negative weak correlations, with no consistent pattern. Also, it is possible to see that there is not a consistent (positive or negative) statistically significant correlation between the number of PIWs and the number of IWs in vocabulary.

The fact that there is no correlation between these pairs of variables becomes clear when considering the performance of BERT-M uncased: despite it being the Portuguese model that contains the largest number of IWs in its vocabulary, models derived from it in general performed poorly in all the previous experiments (i.e., PIWs, PLOO words, IWs that are TF-IDF IWs or TWs). Likewise, although the dictionary of the English model is superior to the others, the corresponding stance classification model does not consistently outperform the other models in these assessed aspects.

Table 6.4: Spearman correlation results for the Percentage of IWs in vocabulary

Dataset	Model	Correlations in the different cut-offs	
		Avg. AA vs. Percentage of IWs in vocabulary	Percentage of PIWs vs. Percentage of IWs in vocabulary
V-DS1	BERTimbau	-0.3939 (p=0.2600)	-0.5714 (p=0.0844)
	BERT-M cased	-0.3333 (p=0.3466)	0.1763 (p=0.6261)
	BERT-M uncased	0.5152 (p=0.1276)	-0.0545 (p=0.8810)
V-DS2	BERTimbau	0.5515 (p=0.0984)	-0.1394 (p=0.7009)
	BERT-M cased	-0.7697 (p=0.0092)	-0.4438 (p=0.1989)
	BERT-M uncased	0.0790 (p=0.8282)	0.0854 (p=0.8146)
V-DS3	BERTimbau	0.2025 (p=0.5748)	-0.1969 (p=0.5855)
	BERT-M cased	-0.5897 (p=0.0728)	-0.0486 (p=0.8939)
	BERT-M uncased	0.5046 (p=0.1369)	-0.1804 (p=0.6179)
SI-DS1	BERTimbau	0.7538 (p=0.0118)	-0.0973 (p=0.7892)
	BERT-M cased	-0.0667 (p=0.8548)	-0.2371 (p=0.5096)
	BERT-M uncased	0.9879 (p=0.0000)	-0.8875 (p=0.0006)
SI-DS2	BERTimbau	0.1337 (p=0.7126)	-0.1616 (p=0.6556)
	BERT-M cased	-0.3939 (p=0.2600)	-0.7561 (p=0.0114)
	BERT-M uncased	0.6261 (p=0.0528)	-0.8598 (p=0.0014)
H-DS	BERT	-0.4316 (p=0.2129)	-0.3951 (p=0.2584)

6.7.3 Discussion

Given that the degree to which a word could be considered an IW/PIW is not related to its presence as a complete word in the BERT model's dictionary, we conclude that the dictionary size/content does not affect the results. Thus, the proposed framework to recompose words' attention scores is correct and compensates for the lack of words in the models' dictionaries.

6.8 Experiment #5: How does the proposed interpretability framework compare to Captum's Sequence Classification Explainer?

This last experiment aims to establish if the results obtained by the proposed method are comparable to an alternative for interpretation of BERT models present in the Captum package (i.e., the *Sequence Classification Explainer*). Such an alignment would further indicate the value of attention weights for interpretability purposes.

6.8.1 Method

As described in Section 2.3.3, the Sequence Classification Explainer is designed to value the words' influence in single predictions using attribution scores. Those scores range from -1 to 1.

Since this experiment intends to compare our IWs against the words determined as influential by this baseline method, only positive values were considered. In this way, we evaluate only the positive relation between words and the predicted class. Thus, we replaced all negative Captum values with zero. To be able to compare our aggregated attention weight (i.e., the AA), we also created an aggregated attribution score for each word considering the whole test set, referred to as *Absolute attribution score*.

First, we traced the tokens back into the original words using the same method proposed to calculate the AA in Section 5.3.2. Then, we aggregated the attribution scores by averaging each word's score considering the set of instances in order to calculate the *Absolute attribution score*. Words with the highest *Absolute attribution score* are considered the most influential ones for the classification (Captum IWs).

Precision@k and NDCG@k were used for the analysis using the same ranges and steps as described in subsection 6.6.1.1, comparing IWs/PIWs with Captum IWs.

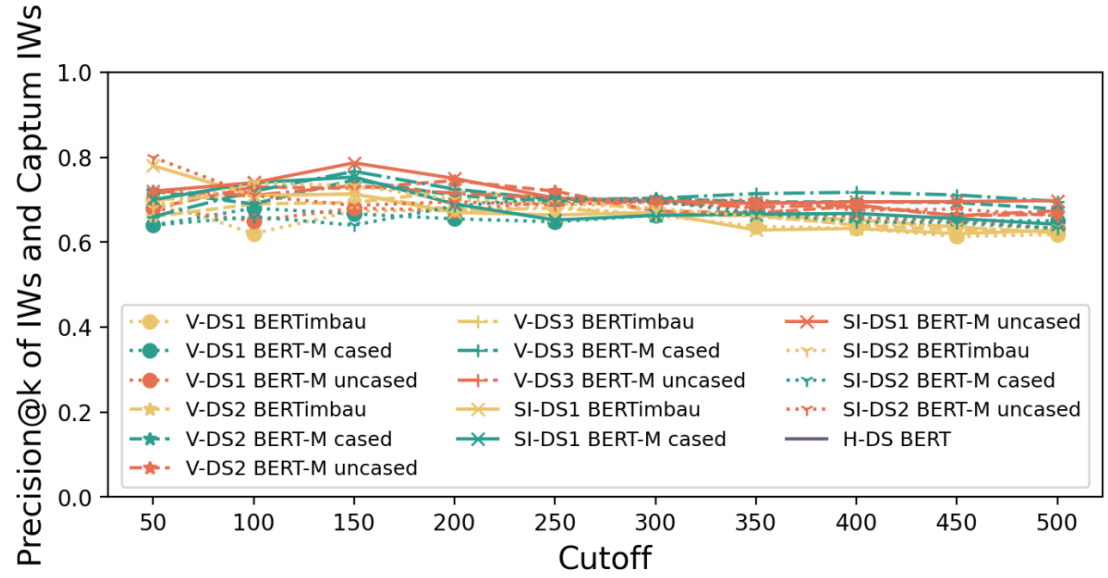
6.8.2 Results

Figure 6.11.(a) presents the Precision@k of IWs that are also Captum IWs. We observe that the Precision@k value is always superior to 0.6, regardless of the k. Notice that, in general, the Precision@k tends to slightly decrease in the same way that the average AA does. Similar behavior is identified in Figure 6.11.(b), which shows the Precision@k of PIWs that are also Captum IWs. The line chart in this figure also shows correlations, although the values are slightly lower for PIWs (between 0.4 and 0.65).

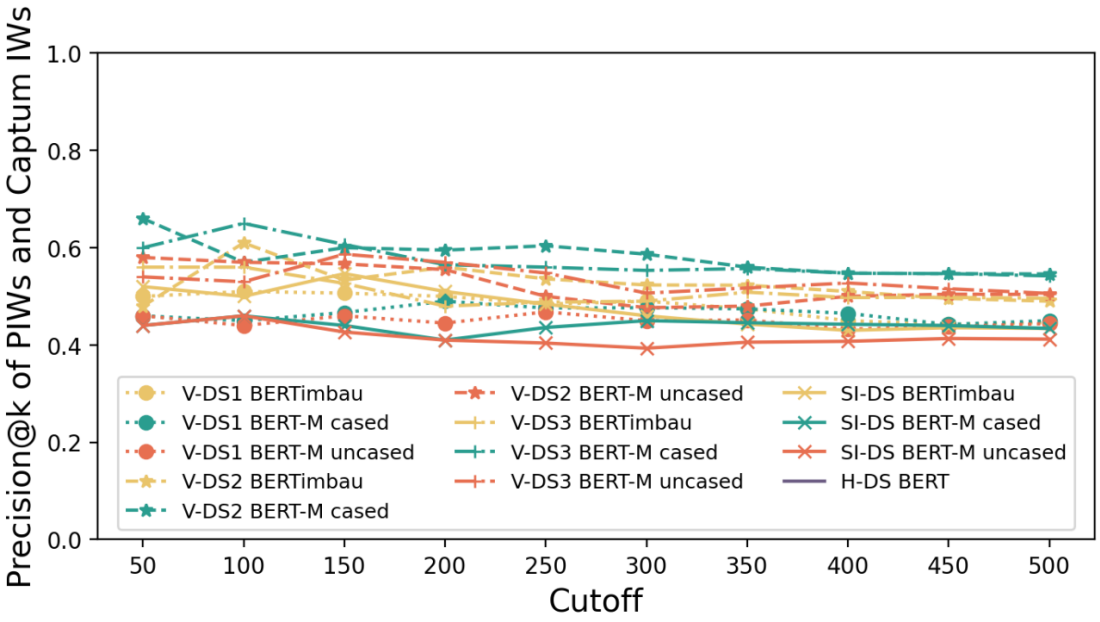
The results of the Spearman tests presented in Table 6.5 reveal the existence of a statistically significant positive correlation between the average AA of IWs and the Precision@k (highlighted in gray), with a few exceptions (BERT-M cased in V-DS1, V-DS2, V-DS3 and SI-DS2, BERT-M uncased in V-DS1, and BERT in H-DS). Similar behavior can be found in this table when comparing PIWs and Captum IWs, with a statistically significant positive correlation in half of the cases and positive correlation in all the cases except BERT-M uncased in SI-DS2.

The boxplots in Figure 6.12.(a) outline the distributions of the NDCG@k for each model. The medians range from 0.92 (V-DS3 with BERT-M cased) to 0.99 (V-DS2 with BERTimbau), revealing a significant alignment of these two scores. It complements the previous Precision@N result in that, although some words are not at the intersection of both methods, the relevance of the common words is ranked very similarly

Figure 6.11: Precision@k of IWs/PIWs that are also Captum IWs



(a) Precision@k of IWs and Captum IWs



(b) Precision@k of IWs and Captum PIWs

by both methods. The NDCG@k scores were also high for the comparison with PIWs (Figure 6.12.(b)), with medians between 0.81 (BERT in H-DS) up to 0.99 (V-DS2 with BERT-M cased).

6.8.3 Discussion

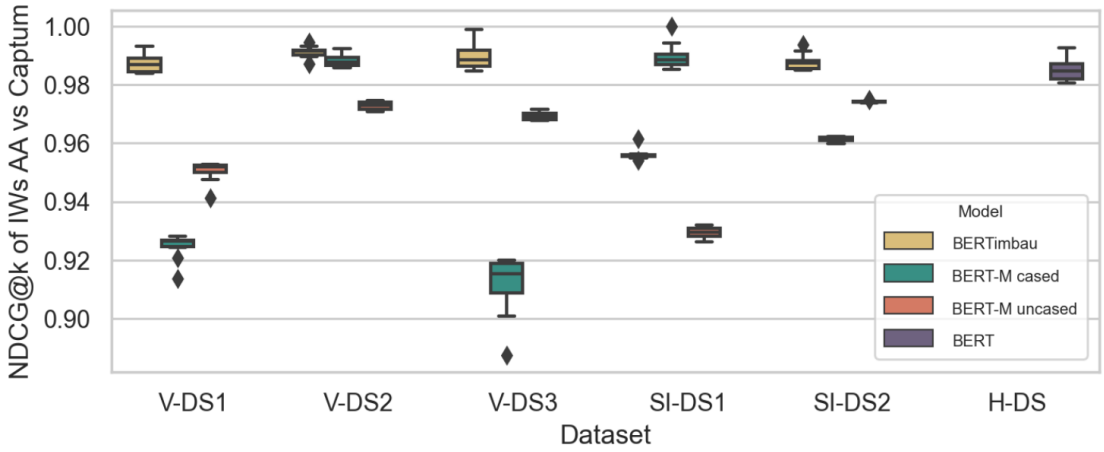
The strong significant alignment between the proposed interpretability framework and Captum's Sequence Classification Explainer proved that our framework could find

Table 6.5: Spearman correlation results between Precision@k of IWs/PIWs and Captum IWs

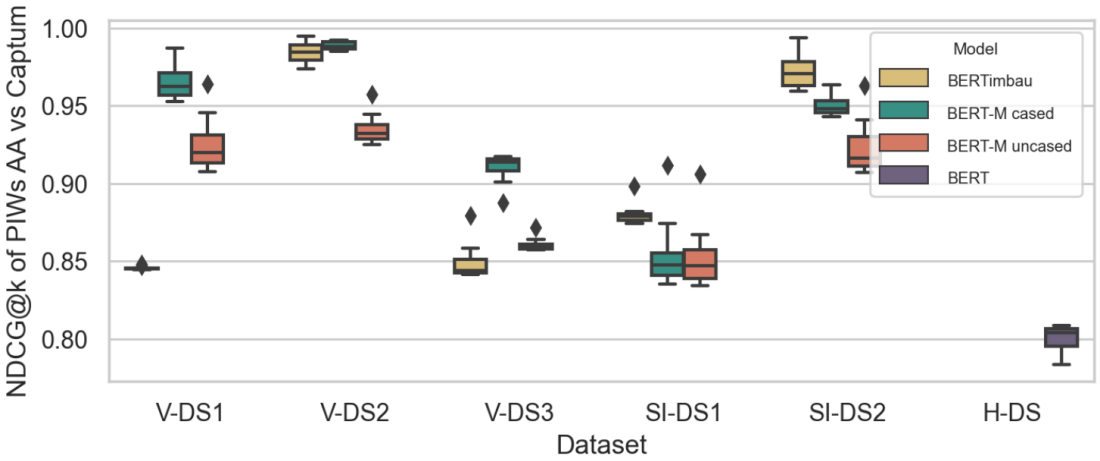
Dataset	Model	Correlations in the different cut-offs	
		Precision@k of IWs and Captum IWs	**Precision@k of PIWs and Captum IWs**
V-DS1	BERTimbau	0.6848 (p=0.0289)	0.9240 (p=0.0001)
	BERT-M cased	0.1030 (p=0.7770)	0.2249 (p=0.5321)
	BERT-M uncased	0.3497 (p=0.3219)	0.5046 (p=0.1369)
V-DS2	BERTimbau	0.6364 (p=0.0479)	0.4182 (p=0.2291)
	BERT-M cased	0.6121 (p=0.0600)	0.8424 (p=0.0022)
	BERT-M uncased	0.7091 (p=0.0217)	0.6444 (p=0.0443)
V-DS3	BERTimbau	0.9394 (p=0.0001)	0.4255 (p=0.2202)
	BERT-M cased	0.2121 (p=0.5563)	0.9515 (p=0.0000)
	BERT-M uncased	0.9152 (p=0.0002)	0.7333 (p=0.0158)
SI-DS1	BERTimbau	0.9423 (p=0.0000)	0.9030 (p=0.0003)
	BERT-M cased	0.7818 (p=0.0075)	0.1534 (p=0.6723)
	BERT-M uncased	0.7212 (p=0.0186)	0.4545 (p=0.1869)
SI-DS2	BERTimbau	0.8511 (p=0.0018)	0.5061 (p=0.1355)
	BERT-M cased	0.1155 (p=0.7507)	0.7212 (p=0.0186)
	BERT-M uncased	0.7576 (p=0.0111)	-0.4182 (p=0.2291)
H-DS	BERT	-0.0909 (p=0.8028)	0.9152 (p=0.0002)

the words that contributed the most to the classification, leveraging only BERT's attention weights. Despite a slight difference among the words considered as IW/PIW and Captum IW, high AA scores increase the probability of a word being influential based on both models, revealing the potential of AA word scores to interpret BERT models for stance classification. Moreover, considering that the Sequence Classification Explainer relies on complex computations that take more computational and time resources, and focuses only on instance-level interpretability, we can affirm that our proposed is a more useful interpretability solution as it can perform model-level interpretability relying solely in BERT's attention weights to obtain insights from the more influential words in the predicted instances.

Figure 6.12: NDCG@k of IWs/PIWs rankings based on AA vs. Captum score



(a) NDCG@k of IWs AA and Captum



(b) NDCG@k of PIWs AA and Captum

7 CONCLUSIONS AND FUTURE WORK

This work proposed a framework for interpretability of BERT-based stance classification that finds the most (positive) influential words for the correct classification of stances and uses them to understand the model decisions. It targets users who do not necessarily have a strong knowledge of the internal workings of BERT but who want to understand the reasons behind the model's predictions.

Compared to related work, we developed a broader level of interpretability focused on the overall model behavior and influential words by using BERT's attention weights. The proposed concepts of *Absolute Attention*, *Proportional Attention Weight*, and *Minimum Positive Attention* help identifying the words that most influenced on the correct classifications.

We set up a broad experimental scenario involving three COVID-19 related case studies. We derived six different datasets that, combined with four BERT pre-trained models, allowed us to analyze the results of sixteen fine-tuned stance classification models. We obtained encouraging answers to our research questions and generalization evidence for our findings. The main insights are:

(a) the choice of a specific BERT pre-trained model do not influence the results, which confirms the generalization of our framework;

(b) high attention scores do not correlate with correct classification but improve the probability of finding words that positively affect the model performance (PLOOs) and influence the correct classification (PIWs);

(c) the IWs are representative of the domain, and the PIWs can be used to identify how the model leverage the arguments expressed by the stances to perform a prediction;

(d) the vocabulary of a BERT model does not influence the results obtained using our interpretability framework, and thus, our method to recompose tokens' attentions into words' attention weights is correct;

(e) the results obtained using our attention-based framework for model-level interpretability are comparable to baseline methods such as Captum's *Sequence Classification Explainer* which relies on other features, additional complex computations and is oriented only to instance-level interpretability.

This research work resulted in two publications explicitly about the interpretability framework:

- **SÁENZ, C. A. C.**; BECKER, K. Interpreting bert-based stance classification: a case study about the brazilian covid vaccination. In: **SBC (Ed.). XXXVI Simpósio Brasileiro de Banco de Dados, 2021**. [S.l.: s.n.], 2021. p. 12p.

- **SÁENZ, C. A. C.**; BECKER, K. Assessing the use of attention weights to interpret bert-based stance classification. In: **Proc. of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI/IAT 2021**. [S.l.: s.n.], 2021

The framework emerged as part of our research project on the political polarization on COVID stances, used in this dissertation as the case studies, which resulted in four publications:

- EBELING, R.; **SÁENZ, C.A.C.**; et al. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario. In: **Proc. of the 15th Intl. Conference on Web and Social Media (ICWSM) - To appear**. [s.n.], 2022.

- EBELING, R.; **SÁENZ, C.A.C.**; et al. Quarenteners vs. chloroquiners: A framework to analyze how political polarization affects the behavior of groups. In: **Proc. of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI/IAT 2020**. [S.l.]: IEEE, 2020. p. 203–210.

- EBELING, R.; **SÁENZ, C.A.C.**; et al. Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In: **SBC. Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. [S.l.], 2020. p. 89–96.

- EBELING, R.; **SÁENZ, C.A.C.**; et al. The effect of political polarization on social distance stances in the brazilian covid-19 scenario. **Journal of Information and Data Management**, v. 12, n. 1, p. 86–108, Aug. 2021.

We received two awards from our research work:

- Best Student Paper Award - Runner up (WI-IAT21) for the paper (SáENZ; BECKER, 2021).

- Best Paper Award (KDMiLe 2020) for the paper (EBELING et al., 2020b);

Finally, our earlier studies on the use of BERT and its applications on fake news also resulted in publications:

- **SÁENZ, C. A. C.**; DIAS, M.; BECKER, K. Combining compact news representations generated using distilbert and topological features to classify fake news. In: **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2020. p. 209–216. ISSN 2763-8944.
- **SÁENZ, C. A. C.**; DIAS, M.; BECKER, K. Assessing the combination of distilbert news representations and difusion topological features to classify fake news. **Journal of Information and Data Management**, v. 12, n. 1, Aug. 2021.

This research still can be improved, and current limitations can be tackled by future work involving:

- investigating additional metrics to identify the relevant and influential words in correct classifications to mitigate limitations;
- using our frameworks to other sets of tokens apart from words (uni-grams) such as bi-grams or tri-grams;
- extending the comparison of IWs/PIWs with alternative interpretability methods (e.g., TranShap);
- exploring other ways to define the MPA threshold to determine the IWs that can be considered PIWs;
- examining local stance-related interpretability analysis using our framework;
- developing a tool for identifying influential classification words based on the proposed framework and assessing its contribution in other interpretation scenarios.

REFERENCES

ABNAR, S.; ZUIDEMA, W. Quantifying attention flow in transformers. In: **Proc. of the 58th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2020. p. 4190–4197.

ALAMMAR, J. **The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)**. 2018. Available from Internet: <http://jalammar.github.io/illustrated-bert/>.

ALDAYEL, A.; MAGDY, W. Stance detection on social media: State of the art and trends. **Information Processing & Management**, v. 58, n. 4, p. 102597, 2021. ISSN 0306-4573.

ALVAREZ-MELIS, D.; JAAKKOLA, T. S. On the robustness of interpretability methods. In: **WHI 2018**. [s.n.], 2018. Available from Internet: <https://www.microsoft.com/en-us/research/publication/on-the-robustness-of-interpretability-methods/>.

AYOUB, J.; YANG, X. J.; ZHOU, F. Combat covid-19 infodemic using explainable natural language processing models. **Information Processing Management**, v. 58, n. 4, p. 102569, 2021. ISSN 0306-4573. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0306457321000704>.

BAI, B. et al. Why attentions may not be interpretable? In: **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining**. New York, NY, USA: Association for Computing Machinery, 2021. (KDD '21), p. 25–34. ISBN 9781450383325.

CAPTUM. **Interpreting BERT Models (Part 1)**. 2022. Available from Internet: <https://captum.ai/tutorials/Bert_SQUAD_Interpret>.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. **Electronics**, v. 8, n. 8, 2019. ISSN 2079-9292. Available from Internet: <https://www.mdpi.com/2079-9292/8/8/832>.

CHEFER, H.; GUR, S.; WOLF, L. Transformer interpretability beyond attention visualization. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2021. p. 782–791.

DAUDERT, T. Exploiting textual and relationship information for fine-grained financial sentiment analysis. **Knowledge-Based Systems**, v. 230, p. 107389, 2021. ISSN 0950-7051.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: **Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)**. [S.l.: s.n.], 2019. p. 4171–4186.

EBELING, R. et al. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario. In: **Proc. of the 15th Intl. Conference on Web and Social Media (ICWSM) - To appear**. [s.n.], 2022. Available from Internet: <arXiv:2110.03382>.

EBELING, R. et al. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario. **arXiv preprint arXiv:2110.03382**, 2021.

EBELING, R. et al. The effect of political polarization on social distance stances in the brazilian covid-19 scenario. **Journal of Information and Data Management**, v. 12, n. 1, p. 86–108, Aug. 2021.

EBELING, R. et al. Quarenteners vs. chloroquiners: A framework to analyze how political polarization affects the behavior of groups. In: **Proc. of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI/IAT 2020**. [S.l.]: IEEE, 2020. p. 203–210.

EBELING, R. et al. Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In: SBC. **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. [S.l.], 2020. p. 89–96.

FILHO, J. A. W. et al. The brWaC corpus: A new open resource for Brazilian Portuguese. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Available from Internet: <https://aclanthology.org/L18-1686>.

GHOSH, S. et al. Stance detection in web and social media: A comparative study. In: CRESTANI, F. et al. (Ed.). **Experimental IR Meets Multilinguality, Multimodality, and Interaction**. Cham: Springer International Publishing, 2019. p. 75–87. ISBN 978-3-030-28577-7.

GIORGIONI, S. et al. Unitor @ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In: **Proc. of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)**. [S.l.]: CEUR-WS.org, 2020. (CEUR Workshop Proceedings, v. 2765).

JAIN, S.; WALLACE, B. C. Attention is not Explanation. In: **Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1**. [S.l.: s.n.], 2019. p. 3543–3556.

KAWINTIRANON, K.; SINGH, L. Knowledge enhanced masked language model for stance detection. In: **Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [s.n.], 2021. p. 4725–4735. Available from Internet: <https://www.aclweb.org/anthology/2021.naacl-main.376>.

KOKALJ, E. et al. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In: **Proc. of the EACL Hackashop on News Media Content Analysis and Automated Report Generation**. [s.n.], 2021. p. 16–21. Available from Internet: <https://www.aclweb.org/anthology/2021.hackashop-1.3>.

KUCHER, K. et al. StanceVis prime: visual analysis of sentiment and stance in social media texts. **Journal of Visualization**, v. 23, n. 6, p. 1015–1034, dec. 2020.

LAI, M. et al. Multilingual stance detection in social media political debates. **Computer Speech Language**, v. 63, p. 101075, 2020. ISSN 0885-2308. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0885230820300085>.

LI, S. **Explain NLP models with LIME amp; SHAP**. 2019. Available from Internet: <https://towardsdatascience.com/explain-nlp-models-with-lime-shap-5c5a9f84d59b>.

LIPTON, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. **Queue**, Association for Computing Machinery, New York, NY, USA, v. 16, n. 3, p. 31–57, jun 2018. ISSN 1542-7730. Available from Internet: <https://doi.org/10.1145/3236386.3241340>.

LUKASIK, M. et al. Gaussian processes for rumour stance classification in social media. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 37, n. 2, feb 2019. ISSN 1046-8188. Available from Internet: <https://doi.org/10.1145/3295823>.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. p. 4765–4774.

MADSEN, A. et al. **Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining**. 2021.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. USA: Cambridge University Press, 2008. ISBN 0521865719.

MARCINKEVIČS, R.; VOGT, J. E. Interpretability and explainability: A machine learning zoo mini-tour. **arXiv preprint arXiv:2012.01805**, 2020.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, v. 267, p. 1–38, 2019. ISSN 0004-3702. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0004370218305988>.

MOLNAR, C. **Interpretable Machine Learning**: A guide for making black box models explainable. [S.l.: s.n.], 2019. <https://christophm.github.io/interpretable-ml-book/>.

MUTLU, E. C. et al. A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. **Data in Brief**, v. 33, p. 106401, 2020. ISSN 2352-3409.

POPAT, K. et al. STANCY: Stance classification based on consistency cues. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 6413–6418. Available from Internet: <https://aclanthology.org/D19-1675>.

RIBEIRO, M.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations**. San Diego, California: Association for Computational Linguistics, 2016. p. 97–101. Available from Internet: <https://aclanthology.org/N16-3020>.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in bertology: What we know about how bert works. **Transactions of the Association for Computational Linguistics**, v. 8, p. 842–866, Dec 2020. ISSN 2307-387X.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature Machine Intelligence**, Nature Publishing Group, v. 1, n. 5, p. 206–215, 2019.

SáENZ, C. A. C.; BECKER, K. Assessing the use of attention weights to interpret bert-based stance classification. In: **IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology**. New York, NY, USA: Association for Computing Machinery, 2021. (WI-IAT '21), p. 194–201. ISBN 9781450391153. Available from Internet: <https://doi.org/10.1145/3486622.3493966>.

SÁENZ, C. A. C.; BECKER, K. Interpreting bert-based stance classification: a case study about the brazilian covid vaccination. In: SBC (Ed.). **XXXVI Simpósio Brasileiro de Banco de Dados, 2021**. [S.l.: s.n.], 2021. p. 12p.

SÁENZ, C. A. C.; DIAS, M.; BECKER, K. Combining compact news representations generated using distilbert and topological features to classify fake news. In: **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2020. p. 209–216. ISSN 2763-8944. Available from Internet: <https://sol.sbc.org.br/index.php/kdmile/article/view/11978>.

SÁENZ, C. A. C.; DIAS, M.; BECKER, K. Assessing the combination of distilbert news representations and difusion topological features to classify fake news. **Journal of Information and Data Management**, v. 12, n. 1, Aug. 2021. Available from Internet: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/1895>.

SERRANO, S.; SMITH, N. A. Is attention interpretable? In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 2931–2951.

SHAPLEY, L. S. A value for n-person games. In: ____. **The Shapley Value: Essays in Honor of Lloyd S. Shapley**. [S.l.]: Cambridge University Press, 1988. p. 31–40.

SLACK, D. et al. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: **AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)**. [s.n.], 2020. Available from Internet: <https://arxiv.org/pdf/1911.02508.pdf>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: **Proceedings of the 34th International Conference on Machine Learning - Volume 70**. [S.l.]: JMLR.org, 2017. (ICML'17), p. 3319–3328.

TENNEY, I.; DAS, D.; PAVLICK, E. BERT rediscovers the classical NLP pipeline. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. Florence, Italy: Association for Computational Linguistics, 2019. p. 4593–4601.

TSAKALIDIS, A. et al. Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. In: **Proceedings of the 27th ACM International Conference on Information and Knowledge Management**. [S.l.: s.n.], 2018. p. 367–376.

VANTA, T.; AONO, M. Stance classification and rumor analysis: Using new dialog-act features and augmenting input tweets. In: **2020 7th International Conference on Advance Informatics: Concepts, Theory and Applications (ICAICTA)**. [S.l.: s.n.], 2020. p. 1–6.

VASHISHTH, S. et al. **Attention Interpretability Across NLP Tasks**. 2019.

VASWANI, A. et al. **Attention Is All You Need**. 2017.

VIG, J. A multiscale visualization of attention in the transformer model. In: **Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. [S.l.: s.n.], 2019. p. 37–42.

WANG, X. et al. A novel network with multiple attention mechanisms for aspect-level sentiment analysis. **Knowledge-Based Systems**, v. 227, p. 107196, 2021. ISSN 0950-7051.

WEI, P.; LIN, J.; MAO, W. Multi-target stance detection via a dynamic memory-augmented network. In: **The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2018. (SIGIR '18), p. 1229–1232. ISBN 9781450356572. Available from Internet: <https://doi.org/10.1145/3209978.3210145>.

WIEGREFFE, S.; PINTER, Y. Attention is not not explanation. In: **Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)**. [S.l.: s.n.], 2019. p. 11–20.

WU, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. **CoRR**, abs/1609.08144, 2016. Available from Internet: <http://arxiv.org/abs/1609.08144>.

YILMAZ, Z. A. et al. Applying BERT to document retrieval with birch. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 19–24. Available from Internet: <https://aclanthology.org/D19-3004>.

ZHANG, Q. et al. From stances' imbalance to their hierarchicalrepresentation and detection. In: **The World Wide Web Conference**. New York, NY, USA: Association for Computing Machinery, 2019. (WWW '19), p. 2323–2332. ISBN 9781450366748. Available from Internet: <https://doi.org/10.1145/3308558.3313724>.

APPENDIX A — RESUMO EXPANDIDO

Modelos baseados na arquitetura Transformer, particularmente BERT (DEVLIN et al., 2019), têm obtido resultados dentro do estado da arte em diferentes tarefas de Procesamento de Linguagem Natural (PLN), tais como classificação de textos, *question answering* ou tradução (YILMAZ et al., 2019; GHOSH et al., 2019; GIORGIONI et al., 2020; KAWINTIRANON; SINGH, 2021; DAUDERT, 2021; WANG et al., 2021). A grande variedade de modelos pré-treinados com *corpora* em vários idiomas (e.g., português, inglês), de diferentes tamanhos (e.g., *base*, *large*) e formatos (e.g., *cased*, *uncased*), permitiu o desenvolvimento de modelos BERT que, por meio do ajuste fino usando conjuntos de dados de domínio específicos, podem obter resultados com desempenho superior.

No entanto, os benefícios de desempenho do BERT vieram à custa da interpretabilidade (TENNEY; DAS; PAVLICK, 2019; ROGERS; KOVALEVA; RUMSHISKY, 2020). De acordo com Molnar (2019), interpretabilidade é o grau em que uma pessoa pode entender os motivos de uma predicão produzida por um modelo de aprendizado de máquina (Machine Learning - ML). A interpretabilidade tem por objetivo fornecer aos usuários *insights* sobre os resultados obtidos por um modelo, o que pode auxiliar ainda mais na realização de modificações. Existem tentativas de adaptar técnicas de interpretabilidade de ML existentes ao BERT, como TranSHAP (KOKALJ et al., 2021), com base em SHAP (LUNDBERG; LEE, 2017); e Captum, que é baseado em *Integrated Gradients* (SUNDARARAJAN; TALY; YAN, 2017).

Outra tendência tem sido utilizar pesos de atenção do BERT para fins de interpretabilidade. Os pesos de atenção são fornecidos pelos mecanismos internos de atenção que são centrais à arquitetura Transformer subjacente do BERT. Estudos manifestaram opiniões contraditórias, destacando os prós e contras de usar esses valores para interpretabilidade (JAIN; WALLACE, 2019; WIEGREFFE; PINTER, 2019; SERRANO; SMITH, 2019; VASHISHTH et al., 2019; BAI et al., 2021). Além disso, o BERT possui internamente um grande número de pesos de atenção, o que dificulta sua interpretação. Propostas existentes tentaram consolidar esses valores (ABNAR; ZUIDEMA, 2020; CHEFER; GUR; WOLF, 2021) ou fornecer uma forma de visualizá-los intuitivamente (VIG, 2019). No entanto, estes trabalhos têm duas limitações principais. Primeiro, eles são direcionados à interpretabilidade em nível da instância, dificultando a identificação de padrões nas previsões gerais feitas pelo modelo. Em segundo lugar, essas técnicas se concentram em *tokens*, que geralmente são partes de palavras com pouco sen-

tido em si, dificultando a compreensão dos pesos de atenção em termos de semântica no mundo real.

Este trabalho aborda a classificação de posicionamentos, ou seja, a tarefa de identificar a posição (e.g., a favor, em contra) expressa por uma pessoa sobre um tema em avaliação (ALDAYEL; MAGDY, 2021), na qual os modelos baseados em BERT alcançaram resultados estado da arte (GIORGIONI et al., 2020; KAWINTIRANON; SINGH, 2021). Em particular, nosso grupo de pesquisa investigou posicionamentos sobre questões subjacentes à pandemia de COVID-19 (e.g., vacinação, isolamento social) e como elas são influenciadas pela polarização política (EBELING et al., 2020a; EBELING et al., 2021b; EBELING et al., 2022). O objetivo principal do presente trabalho é investigar como os mecanismos de atenção podem ser aproveitados para entender as predicões de polarização feitas pelos modelos BERT.

O presente trabalho propõe um *framework* de interpretabilidade para identificar as palavras mais influentes nas predicões. O framework é focado no modelo como um todo e, assim, relaciona um escore de atenção (*Atenção Absoluta*) a palavras que são significativas dentro de um conjunto de documentos para identificar aquelas mais importantes para a classificação (*Palavras Influentes*). Nós propomos uma métrica (*Peso de Atenção Proporcional*) para identificar as palavras influentes que mais contribuem para a correta classificação das instâncias (*Palavras Positivamente Influentes*). O *framework* parte dos pesos dos *tokens* coletados para cada instância de acordo com o método descrito em (CHEFER; GUR; WOLF, 2021), e desenvolve um nível mais amplo de interpretabilidade:

- relacionando os pesos de atenção dos *tokens* às suas palavras originais para cada instância;
- agregando os escores de atenção das palavras em instâncias individuais, criando uma medida geral de influência das palavras em relação às previsões feitas pelo modelo.

A.1 Contribuições da Dissertação

Com resultados preliminares apresentados em (SÁENZ; BECKER, 2021) e (SáENZ; BECKER, 2021), as principais contribuições desta dissertação são:

- Um *framework* de interpretabilidade para identificar as palavras mais influentes na classificacao de posicionamento usando modelos BERT, baseadado em pesos

internos de atenção. Ao contrário dos trabalhos relacionados (CHEFER; GUR; WOLF, 2021; ABNAR; ZUIDEMA, 2020; VIG, 2019), esta proposta fornece um nível mais amplo de interpretabilidade focado no comportamento geral do modelo em relação a um conjunto de dados de teste. Também agrega *tokens* em palavras que podem ser semanticamente relacionadas ao domínio. Também propomos métricas para medir a relevância delas em predições (corretas);

- Um amplo conjunto de experimentos quantitativos e estatísticos envolvendo diferentes estudos de caso, conjuntos de dados, modelos pré-treinados de BERT e métricas para avaliar o framework proposto. Os resultados fornecem *insights* e padrões valiosos que nos permitem generalizar nossos resultados, contribuindo com mais evidências sobre o valor dos pesos de atenção para a interpretabilidade dos modelos BERT para a classificação de posicionamentos.

A.2 Principais Resultados Alcançados

Montamos um amplo cenário experimental envolvendo três estudos de caso relacionados ao COVID-19 (EBELING et al., 2020b; EBELING et al., 2022; MUTLU et al., 2020). Derivamos seis conjuntos de dados diferentes que, combinados com quatro modelos pré-treinados de BERT, nos permitiram analisar os resultados de dezesseis modelos treinados para a classificação de polarização. Obtivemos respostas encorajadoras para nossas perguntas de pesquisa e evidências de generalização para nossas descobertas. Os principais *insights* foram:

(a) a escolha de um determinado modelo pré-treinado de BERT, bem como o tamanho/conteúdo do dicionário relacionado, não influenciam nos resultados;

(b) altas pontuações de atenção não se correlacionam com a classificação correta, mas melhoram a probabilidade de encontrar palavras que afetam positivamente o desempenho do modelo e influenciam a classificação correta ;

(c) as palavras influentes são representativas do domínio, e as palavra positivamente influentes podem ser usados para identificar como o modelo utiliza os argumentos expressos para predizer uma posição;

(d) o vocabulário de um modelo BERT não influencia os resultados obtidos usando nosso framework de interpretabilidade;

(e) os resultados obtidos por nosso framework usando pesos de atenção internos do modelo para interpretabilidade são comparáveis aos métodos de linha de base, como o *Sequence Classification Explainer* de Captum, que utiliza diferentes métricas de contribuição.