Interpreting BERT-based stance classification: a case study about the Brazilian COVID vaccination

Carlos Abel Córdova Sáenz, Karin Becker

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{cacsaenz, karin.becker}@inf.ufrgs.br

Abstract. *The actions to control the COVID-19 pandemics should be based on scientific facts. However, Brazil is facing a politically polarized scenario that has influenced the population's behavior regarding social distance or vaccination issues. This paper addresses this subject by proposing a BERT-based stance classification model and an attention-based mechanism to identify the influential words for stance classification. The interpretation mechanism traces tokens' attentions back to words, assigning word attention scores (absolute and relative). We use these metrics to assess if words with high attention weights correspond to domain intrinsic properties and contribute to the correct classification of stances. Our experiments revealed good results for stance classification (F1=0.752), and that 74% of the top-100 words with the highest absolute attention are representative of the arguments that support the investigated stances.*

1. Introduction

The COVID-19 pandemic has caused a high degree of polarization worldwide. People's behavior towards social isolation, scientifically unproven treatments (e.g., chloroquine, ivermectin), or vaccination has been greatly affected by political beliefs. In [Ebeling et al. 2020], we proposed a multi-dimensional model to understand the influence of political orientation in the behavior regarding social distance expressed on Twitter. We used this model to analyze the COVID-19 vaccination scenario [Ebeling et al. 2022], identifying three stances: (a) users eager to get vaccinated (i.e. Pro-vaxxers), (b) users against (mandatory) vaccination (i.e. Anti-vaxxers), and (c) users specifically against Coronavac, developed in partnership with a Chinese laboratory (i.e. Anti-sinovaxxers). As part of this ongoing research, we also seek to determine if it is possible to develop a predictive model of these stances based on the arguments expressed in tweets.

Stance classification is the machine learning task of identifying the position (e.g., in favor, against) expressed by a person on an issue being evaluated [ALDayel and Magdy 2021]. Different techniques have been applied for this task, where transfer-learning using BERT [Devlin et al. 2019] has achieved state-of-the-art results [Giorgioni et al. 2020, Kawintiranon and Singh 2021].

Interpretability and explainability have become extremely important in predictive models. In this work, we adopt the definition of interpretability as the degree to which a person can understand the reasons for a prediction made by a Machine Learning (ML) model, as proposed in [Molnar 2019]. Interpretability mechanisms such as Shapley values and LIME contribute to providing insights on the influential features for classification [Kokalj et al. 2021]. There has been a significant amount of research to understand whether attention weights, which are key to BERT's Transformer architecture, can be leveraged for interpretation [Jain and Wallace 2019, Wiegreffe and Pinter 2019]. Some

techniques have been proposed [Chefer et al. 2021, Vig 2019]. However, they only enable to analyze influential tokens from the input text rather than words, or they display multiple attention weights relative to the different layers and attention mechanisms inside the model. Thus, in practice, one can have difficulty making sense of these weights in terms of the semantics in the real world.

In this work, we develop a case study in the context of COVID-19 vaccination to assess the contribution of attention weights in understanding the stances automatically classified using BERT. We developed a BERT-based classification model to classify tweets into the three stances identified in [Ebeling et al. 2022] and proposed an attention-based interpretability mechanism to understand influential features. To address the Portuguese language, we used the BERTimbau language model [Souza et al. 2020]. The interpretability mechanism proposed expands the attention consolidation proposed in [Chefer et al. 2021] by relating tokens to the original words and associating them to two word attention metrics: absolute and relative. Then, we use these metrics to assess if words with high attention weights correspond to properties intrinsic to the domain and contribute to the correct classification of stances.

Our results were encouraging. We achieved an F-measure of 0.752 for the stance classification task. The absolute word attention metric revealed to be a useful tool for interpreting the model, as 74% of the top-100 words with the highest absolute attention are also relevant in the arguments used to support stances, and 68% are contributing to correct classifications. We also gained insights into the tokenization process of BERTimbau, and how it influences the proposed metrics.

The rest of this paper is structured as follows. Section 2 describes related work. Section 3 describes the case study about the COVID-19 vaccination in Brazil. Section 4 presents the proposed framework. Section 5 describes the experiments performed. Section 6 draws conclusions and points to future work.

2. Related Work

Supervised stance classification has been addressed using traditional ML and Deep Learning (DL) algorithms, where state-of-the-art results were achieved using BERT [Giorgioni et al. 2020, Kawintiranon and Singh 2021]. BERT is first used to create vector representations of the input text by transfer-learning of a language representation model, which can be used either as features of ML algorithms or adding a layer to fine-tune the model for the classification task (e.g., *BertForSequenceClassification* in HuggingFace *transformers* package[1]). Pre-trained models exist for different languages. For the Portuguese language, the current options are multilingual BERT and BERTimbau [Souza et al. 2020].

Despite the excellent results in many natural-language processing applications, BERT-based models are black boxes, and thus it is not easy to identify the influential features for classification. Attention mechanisms, the main components of the Transformer architecture, are central to the good performance of BERT-based models. However, the architecture relies on various sets of attention mechanisms, called "heads", distributed throughout the network layers. Thus the relationship between the input, the attention weights, and the outcomes of the model is not straightforward [Rogers et al. 2020]. Another difficulty is that BERT processes the input text as a set of tokens extracted (rather

[1]https://huggingface.co/transformers

than words). Thus the attention weights are assigned over these items, which are more difficult to understand outside the model.

In the current effort to make ML and DL models interpretable, to achieve a better and faster reception of users to these techniques in everyday life, works have discussed the value of attention weights to provide some level of interpretability to BERT-based models [Jain and Wallace 2019, Wiegreffe and Pinter 2019], with arguments in favor and against their contribution. However, this is still a subject of debate that deserves further investigation [Rogers et al. 2020]. Some works propose visualization tools, such as *bertviz* [Vig 2019], that allows visualizing the attention of the tokens in a text under different perspectives, but always considering each attention weight relative to a given layer and head, which is difficult to be understood by a non-expert user. To summarize the attention mechanisms inside BERT, [Abnar and Zuidema 2020] proposed a strategy to condense these values obtained by each token in the whole network from the weights in each part of the network. However, it assumes attentions be linearly combined, a condition that cannot be guaranteed. The technique proposed in [Chefer et al. 2021] leverages LRP (Layer-wise Relevance Propagation) to overcome this limitation and summarizes the attention weights using information related to both the relevance and gradient. This technique highlights the tokens that most contributed to classifying a text, which may not have clear semantics and can compose many different words.

We contribute to the field by building on the solution proposed in [Chefer et al. 2021] to associate consolidated attention weights to words containing the tokens so that one can leverage this knowledge to assess the behavior of a stance classifier in terms of the domain characteristics and the contribution of this words to the correct classification of the stances.

3. Case study: stances on Twitter about COVID-19 vaccination

Despite Brazil's successful history of disease eradication thanks to large-scale National Immunization Programs (NIPs), political polarization has been a significant obstacle to planning and implementing a COVID-19 NIP. Many argue that, throughout 2020, the Federal Government neither supported the national research centers (e.g., Fiocruz, Butantan) to produce vaccines nor undertook substantial efforts to secure contracts with the international pharmaceutical industries to buy vaccines. In addition, Bolsonaro has undermined many initiatives of the governor of São Paulo, João Dória, regarding Coronavac, developed by the Butantan Institute in association with the Chinese pharmaceutical company Sinovac. Bolsonaro and Dória are potential candidates for the 2022 presidential elections, and thus COVID-19 vaccination has been discussed under a strong political bias.

In [Ebeling et al. 2022], we analyzed the political influence in stances expressed in Twitter about COVID-19 vaccination. We crawled tweets containing the terms "vaccine" or "vaccination" from Jan. 1st to Dec. 21st 2020, covering the discussions before the beginning of the vaccination in Brazil (January 2021). The tweets in this period encompass the start of the pandemic, the development of vaccines, and the start of their availability in some high-developed countries worldwide. Based on the most frequent hashtags, we identified three different stances described below. The specific hashtags used to filter out these groups are detailed in [Ebeling et al. 2022].

- *Pro-vaxxers*: the stance in favor of vaccines is expressed using hashtags to support vaccination programs (e.g., VaccinesForLife) and raise awareness about its urgency (e.g., VaccineNow). The hashtags in Portuguese are: #EuVouTomarVacina, #VacinaBrasil,

Table 1. Central arguments used to express each stance

Pro-Vaxxers	Anti-vaxxers	Anti-Sinovaxxers
joy and gratitude for vaccines	individual choice	opposition to mandatory vaccination
expectation of getting vaccinated ASAP	opposition to mandatory vaccination	distrust and rejection of Coronavac
praise for science, national research institutes and the Brazilian Public Health System	criticisms towards governors' "dictatorship"	mistrust/prejudice against the "Chinese" origin
strong criticisms to Bolsonaro	rage against STF ruling (constitutionality)	opposition to Dória
criticisms to government's actions regarding an NIP	support to the President and Federal Government	praise to Bolsonaro

#VacinaÉAmorAoPróximo, #VacinaJá, #VacinaNoBrasil, #VacinaParaTodos, #VacinasPelaVida, #VacinaUrgenteParaTodos, #VemVacina. This stance is represented by 17,290 tweets;

- *Anti-vaxxers*: the stance against COVID-19 vaccination is expressed using hashtags that express no intentions to get vaccinated (IWontTakeVaccine) or against mandatory vaccination to reach community immunization. The hashtags are: #EuNãoVouTomarVacina, #NãoVouTomarVacina, #VacinaNão, #VacinaObrigatóriaNão. This stance is represented of 26,760 tweets;
- *Anti-sinovaxxers*: we identified a stance specifically against Coronavac, using references to the "Chinese vaccine" or the derogatory expression "vacchina". Although it could be regarded as a anti-vaxx stance, we found that the arguments were different and more politically motivated. Thus we decided to maintain it as a separate stance. The hashtags in Portuguese are: #VachinaNão, #VachinaNãoPresidente, #VachinaObrigatóriaNão and #VacinaChinesaNão. This stance is represented by 14,510 tweets.

We deployed two topic modeling techniques (LDA and BERTopic) to understand these stances, as summarized in Table 3. The Pro-vaxxers are anxious to get immunized, praise science advances, celebrate Brazilian research centers in the development of a COVID vaccine, and criticize Bolsonaro and the Federal Government's (lack of) actions towards a NIP. The Anti-vaxxers are concerned about the individual choice regarding vaccines that they do not trust. They are against mandatory vaccination, perceive governors' requests as dictatorship and the ruling of the Supreme Court as unconstitutional, and support the Federal Government position. The Anti-Sinnovaxers stance goes beyond (mandatory) vaccination, also focusing on the rivalry between Bolsonaro and Dória, and risks considering the "Chinese" origin of Coronavac.

We use our earlier work as a case study to investigate the automatic classification of stances using BERT-base models, with a particular focus on assessing if attention weights can be leveraged to identify and interpret the features that are influential to the correct prediction of stances.

4. Framework for Stance Classification and Interpretability

We propose a framework to address tweets written in the Portuguese language expressing stances regarding COVID-19 vaccination. We seek to determine if it is possible to develop a predictive model of these stances based on the arguments expressed in tweets and understand the words that influence the (correct) classification. As depicted in Figure 1, the framework is divided in two parts: a) stance-classification by fine-tuning a BERT model pre-trained with a corpus in Brazilian Portuguese (BERTimbau), and b) an interpretability mechanism to analyze the influential words for the stance classification in terms of attention weights. We propose the concept of *word attention*, in which attention weights assigned to tokens by BERT are transferred back to the original words of the input.

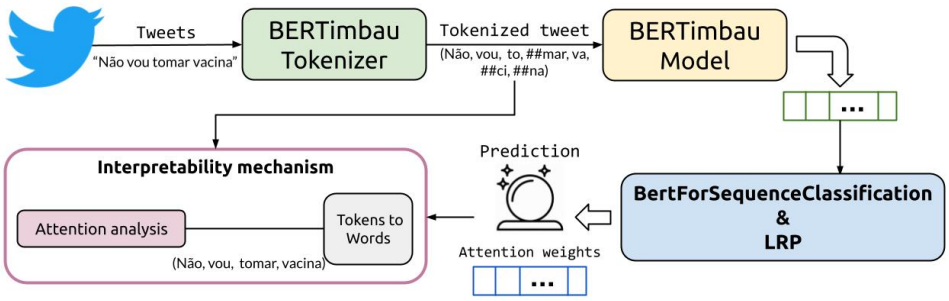

Figure 1. Framework for stance classification and interpretability

The interpretability mechanism is targeted at users who do not necessarily have a strong knowledge of the internal workings of BERT, and use it as a tool to understand why the model is making such predictions and why it is hitting and failing in some cases for each polarized position. It extends the technique of [Chefer et al. 2021], which consolidates the attention weights of the relevant tokens throughout the whole network by relating these tokens and the respective weights to the original words of the input text. As an interpretability mechanism, we expect the proposed word attentions to be useful in identifying important words for classification in general and specific groups and their relevance for correct or wrong predictions.

4.1. Stance classification using BERTimbau

This step aims at the creation of a stance classification model based on the fine-tuning of BERT. The input is a set of labeled tweets, and the outputs are the stance classification predictions and the attention weights of all tokens. We applied typical pre-processing actions over the original tweet set, such as removing mentions/URLs/special characters and eliminating short tweets. Finally, we also took out all hashtags used to define the classes as they could introduce bias to the model. For reproducibility purposes, further details on the pre-processing, classification, and attention calculation can be found in a public repository[2].

Since we are dealing with tweets written in Portuguese, we chose BERTimbau as the language representation model ("bert-large-portuguese-cased"). The textual input is first tokenized and then transformed into a vector using BERTimbau. Then, we fine-tune this model using *BertForSequenceClassification*, a classification algorithm provided by the *transformers* library[3]. The attention weights are obtained using the adapted LRP technique proposed in [Chefer et al. 2021] from the results obtained by the classifier. We save the predictions, the tokenized tweets and the attention weights, as they work as inputs for the interpretability mechanism we propose.

4.2. Interpretability mechanism for the stance classification

Recall that BERT and its variations split the texts into pieces of words called tokens instead of words they receive as inputs for the model. BERT has a fixed vocabulary that contains words identified in the pre-training corpus that it uses for the tokenization process. If a word in the input text received by BERT is also contained in the vocabulary, it is considered a single token. If not, the word is divided into several tokens using the *WordPiece* algorithm. The problem with this strategy is that tokens do not have meaning

[2]https://github.com/cacsaenz/sbbd2021-stance-explanations
[3]https://huggingface.co/transformers/

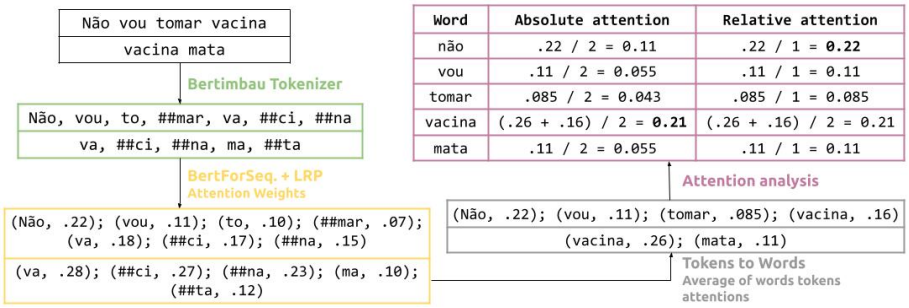

Figure 2. Transformations from tweet to attention weights of words

by themselves. Their relevance with regard to the domain can only be analyzed if the context of the word within which they were contained is identified. In addition, the same token can be part of more than one word. For instance, in the BERTimbau model, the word "vacina" is not in the vocabulary, and thus it is decomposed and explored subsequently by BERT as three tokens: "va", "##ci" and "##na".

Internally, BERT defines and adjusts attention weights in each layer and head, giving higher weight to the tokens that it considers most important. To get the attention weights for each token, we use the modified LRP technique proposed in [Chefer et al. 2021]. Then we propose the concept of *word attention*, in which attention weights assigned to tokens are transferred back to the original words in each tweet, as illustrated in Figure 2. For each tweet, first, we calculate the attention of its words based on the average of the tokens that compose them. Then we calculate the words' attention with regard to the dataset, according to two metrics: *absolute* and *relative*. While the former is calculated for the whole test set, the latter considers only the tweets in which the word appears. Both proposed metrics complement each other and give different perspectives for analyzing the results obtained. They can be used as a tool to analyze words, which are semantically easier to understand within the domain while relying on tokens weights leveraged by BERT models.

a) Absolute attention: the absolute attention of a word in a collection of tweets is the average of the individual word attentions in the total of tweets, multiplied by one hundred. The words with the highest absolute attention are those that in percentage contributed the most in the total attention of the collection of tweets. High values for absolute attention are due to the frequency of the word in the dataset rather than the original high weights assigned to the words. We assume that representativeness in terms of frequency is important for classification pattern identification.

b) Relative attention: the relative attention of a word in a collection of tweets is the average of the word's attention, multiplied by one hundred, considering only the tweets where it appears. The words with relative attention close to 100 are considered very relevant by the BERT model in the tweets that appear within the group. High values for relative attention are due to the original high weights assigned to the words. However, these words may appear in only a few tweets, and for some reason, have received a high attention weight from the classifier. In this sense, this metric allows us to identify words that in some way were considered relevant by the BERT model, regardless of their prevalence.

5. Experiments

The goals of our experiments were: a) to evaluate the performance of the stance classifier in our case study, b) to determine if the words with the highest attentions are representative

	Accuracy	Precision	Recall	F1
LR	0.642	0.644	0.642	0.642
RF	0.585	0.589	0.585	0.586
SVM	0.635	0.638	0.635	0.635
KNN	0.545	0.554	0.545	0.545
XGB	0.630	0.632	0.630	0.631
BertForSeq.	**0.752**	**0.753**	**0.752**	**0.752**

Table 2. Weighted-averaged performance metrics for different algorithms

	Accuracy	Precision	Recall	F1	Tokens with the highest attention
Anti-sinovaxxers	0.735	0.766	0.735	0.750	a, Presidente, ##ia, Fora, Dor, ##Do, ##a, o, ##ria, de
Anti-vaxxers	0.735	0.693	0.735	0.714	a, que, de, vac, ##o, o, não, ##ina, é, e
Pro-vaxxers	**0.785**	**0.801**	**0.785**	**0.793**	vac, de, a, ##ina, o, que, da, e, ##ro, é

Table 3. Individual performance results using *BertForSequenceClassification*

in the domain, and c) to assess if the words with the highest attention are influential in the prediction results.

Due to computational limitations to train BERT models, we selected a random sample of 3,000 instances, evenly distributed for each class (Pro-vaxxers, Anti-vaxxers, and Anti-sinovaxxers), resulting in a balanced dataset.

5.1. Stance classification performance

In this experiment, we compare the performance of the model fine-tuned using *BertForSequenceClassification*, and five other ML algorithms (e.g., Logistic Regression, Random Forests, SVM, KNN, and XGBoost). In both cases, we used the pre-trained BERTimbau "bert-large-portuguese-cased" model to obtain vectorial representations of the tweets. We used 80% of the dataset for training-validation of the model and the remaining 20% for testing. To evaluate the results, we used Accuracy, Precision, Recall, and F1. The results are aggregated using a weighted average.

As it can be seen in Table 2, the *BertForSequenceClassification* model clearly outperforms the others in all metrics. The individual results for each class using *BertForSequenceClassification* are presented in Table 3, where it is possible to see that the group *Pro-vaxxers* is the one with the best results. The lowest results are obtained by *Anti-vaxxers* and *Anti-sinovaxxers*. A possible explanation is the similarity in some of the arguments used by users of both groups (e.g., against mandatory vaccination and endorsement to President Bolsonaro).

Table 3 displays the top-10 tokens with the highest attention for each class. We can notice how hard it is to trace most of them (e.g., ##o, ##ia, da, a) back to the arguments related to each stance. The only exceptions are "Presidente" or "Fora", which are tokens that have a meaning for the domain and are present in BERTimbau's vocabulary. The remaining experiments aim to confirm if our interpretability mechanism helps to understand these results.

5.2. Words attentions and domain representativeness

This experiment aims to assess if the words obtained through the proposed attention metrics are representative and had meaning within the domain of tweets. Thus, we propose mapping tokens back to the words they compose in each token and summarizing their attention in terms of the proposed absolute and relative word attentions. To assess their representativeness, we compare them with the arguments identified in [Ebeling et al. 2022] using BERTopic, and TF-IDF.

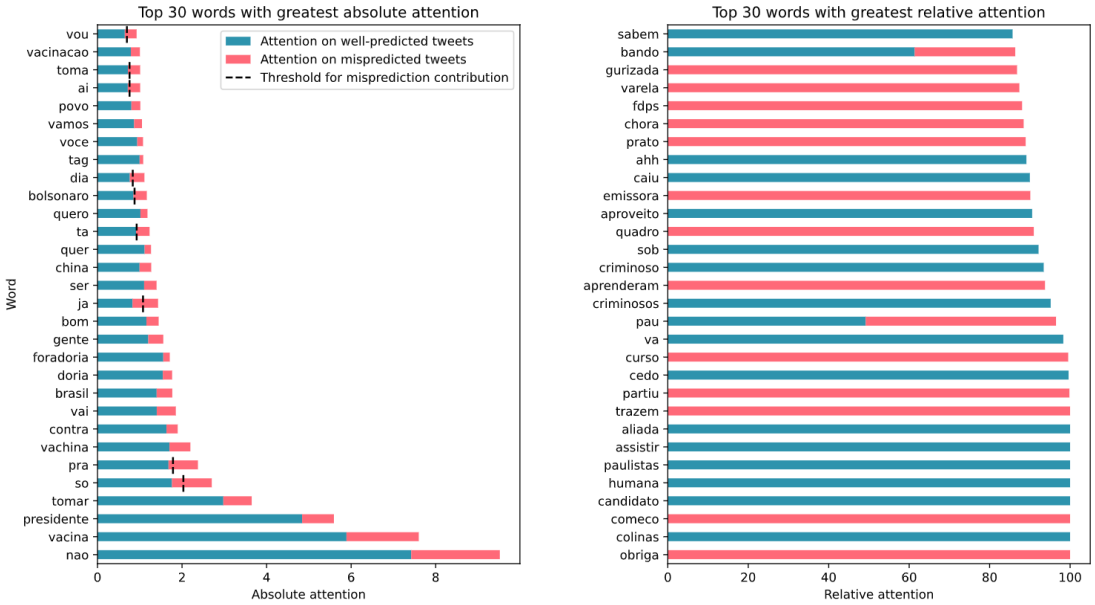

Figure 3. Words with greatest absolute and relative attentions

After converting the tokens and their attentions to words with attention weights, we excluded from our analysis stop words, as they do not contribute to the interpretation analysis. To this end, we used the list in the NLTK[4] library. Next, we calculate each word's absolute and relative attention and choose the top-100 words with the highest results. Figure 3 illustrates the top-30 highest attention words, regarding absolute attention (left-hand side) and relative attention (right-hand side). The words are in the y axis, where the x-axes contain the word attention weight according to each metric.

Regarding the words with the highest *absolute attention*, we can notice several words representative of the case study (e.g., "vacina", "presidente", "tomar", "vachina", etc), as they can be traced back to the topic modeling analysis performed in [Ebeling et al. 2022], as summarized in Table 3. However, we also identify words that could be considered stop words since they are very generic (e.g., "nao", "pra", "so").

The words with the highest *relative attention* solve this issue by considering only the tweets in which the tokens are present, and thus the attention values are much higher. We did a frequency analysis on these words within the test set, and we identified that they are not frequent (about 1% on average). Thus, the model assigns a high attention value that cannot be adjusted due to the few times it finds those words during training.

Finally, we performed a relevance analysis using the TF-IDF index. Since this measure is relative to every word within every tweet, we calculated the TF-IDF average to summarize the word's relevance in the test set. We selected the top-100 words based on the mean TF-IDF index and calculated the intersection with the absolute and relative attention words sets. We identified that 78% of words with the highest absolute attention is also in the TF-IDF set of words. On the other hand, the intersection between the relative attention words set and the TF-IDF set is empty.

These findings allow us to conclude that absolute word attention is the best metric for interpretability purposes since most words are representative of the domain according to the two assessment criteria used and their frequency on the dataset. Nevertheless, it

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

also includes general words. This is consistent with the reasons for which BERTimbau assigns the underlying tokens a higher attention weight. Despite the higher attention values, words with high relative attention cannot be used for understanding the model predictions since they are scarce and not representative.

5.3. Words contribution to model's classification results

This last experiment aims to determine if words with high absolute attention have a positive or negative influence on classification, thus assessing their potential as a tool to interpret the influential words in the classification. In the graphs of Figure 3, the bar associated with each word depicts the weight proportion with regard to the correct (blue) and incorrect (red) classifications.

Recall the absolute attention of a given word is calculated as the average of that word's attention considering all training set tweets. Thus, the proportional attention weight for correct predictions considers the sum of the weights only in the correctly classified instances divided by the total number of predictions. Likewise, the proportional weight for incorrect predictions considers the sum of weights only in the misclassified instances. The sum of these two proportional weights is equal to the word's absolute attention. For instance, the absolute attention of the word "president" is 5.59, where 86.6% of this weight corresponds to correctly classified instances (4.84), and the remaining 13.5% to the misclassified instances (0.75). Given that the model's accuracy is 75.2%, we use this threshold to define the minimum proportion weight to state that the word contributes to correct classification. In Figure 3, dashed lines indicate when the proportion for correct classification is smaller than expected. We can observe in Figure 3 that 70% of the words among the top-30 highest absolute attention contribute to the correct classification. Most words contributing to misclassification are generic words (e.g. "so", "pra", "ja") that could be regarded as stop words. The exceptions are the words "bolsonaro" and "toma", which are representative of the domain. However, they can be found in polarized tweets of different stances as manifestations either of support and detraction.

The proportional weights calculated for the relative attention (right-hand side of Figure 3) reinforces that this metric is not appropriate for interpretation. Given that they are calculated only considering the tweets in which they appear, we can see that, with two exceptions, they are concentrated in either incorrect or correct predictions.

To complement this analysis, we considered the words with the highest absolute attention within each class, depicted in Figure 4. Notice that the thresholds for maximum proportion for incorrect classification (dashed lines) calculated for each class are different due to the distinct number of correctly/incorrectly classified instances of each class (Table 3). Between 60 and 66% of the words positively influence the classification of the respective class. We can observe that the words with highest attention are aligned with the representative arguments expressed in each stance [Ebeling et al. 2022], and most of them are also relevant with respect to their TF-IDF (intersections range between 60 to 69%). For the *Anti-sinovaxxers*, terms like "ditadoria", "ditador", "doria" or "presidente" are frequently used to express the stance and received high attention of the model. However, not all of them contribute to correct classification. In addition to words that are used in the other groups ("nao", "tomar"), surprisingly, we found very specific terms representative of this stance ("doriaditador","foramaia"). For the *Anti-vaxxers*, we found more generic words (e.g., "nao, "pra", "so") than words used in their central arguments (e.g., "stf", "obrigatoria", "vacina", "tomar"). Recall that the worst performance was associated with this class, and these terms can explain this behavior. For the *Pro-vaxxers*, the majority

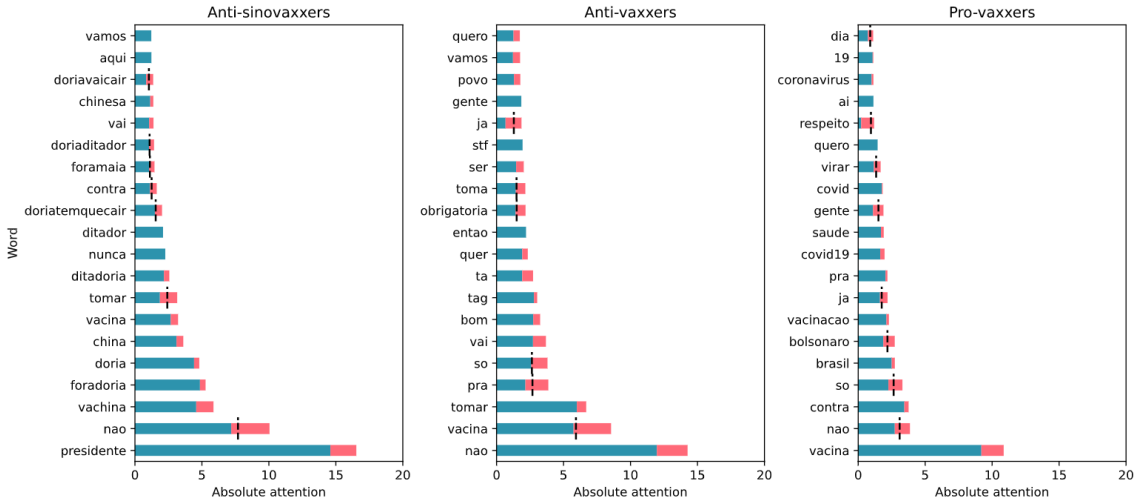

Figure 4. Words with greatest absolute attention per group

of words found were related to the topics discussed by the users of this group. The word "bolsonaro", although representative, contributes to misclassification, possibly because it is used to express support and rejection to him and the policies that he imparts.

Notice that some words contribute to the correct classification of some classes and misclassification of others. This is the case of "nao" that contributes to the classification of *Anti-vaxxers* but harms the other classes, and "pra", which contributes to the hits in the *Pro-vaxxers* and the errors in the *Anti-vaxxers*. We analyzed the top-100 highest absolute attention words to understand this behavior and observed that they contain tokens in common with representative words of other groups. Examples are "foramaia" (*Anti-sinovaxxers*) and "forabolsonaro" (*Pro-vaxxers*) that share the token "fora", or "doriaditador" (*Anti-sinovaxxers*) and "obrigatoria" (*Anti-vaxxers*) that share the token "##ia". We believe the model might be getting confused when receiving tokens shared by influential words of different groups.

The fact that the BERTimbau vocabulary lacks many relevant words in Brazilian Portuguese, or meaningful words for this domain (e.g., "vacina"), makes them be split during the tokenization step into tokens that end up being very common for different words in the polarized groups at the same time. To investigate if this issue is significant, we selected the top-100 tokens with the highest attention weights and manually inspected them. We observed that most of them are tokens composing words (prefix ##) and identified only 13 words that are not pronouns, articles, or conjunctions (e.g., "você", "a", "e"). This issue deserves further investigation, as it impacts not only the interpretation but also the classification itself.

Based on these observations, we conclude that the attention weights can be leveraged to identify words closely related to the tweets' domain and influential to the classification. The absolute word attention metric yielded the best result, enabling the identification of representative words in the domain, as confirmed by the TF-IDF relevance index and the topic modeling analysis done in [Ebeling et al. 2022]. Nevertheless, the tokenization step performed by BERTimbau using its vocabulary seems to cause some side effects. As many domain-related words are not present in BERTimbau's vocabulary, they are divided into tokens that could be present simultaneously in words with different meanings and contexts. This issue may be having a negative impact during the training

of the model. In that sense, we could explore other BERT models (e.g., Multilingual) or ways other than the mean to aggregate the individual weights in words.

6. Conclusions and Future Work

In this article, we analyzed the influence of the internal attention weights of the BERT model on the classifications made in a case study on the polarization of people regarding the Brazilian vaccination against COVID-19. We proposed a framework to classify stances expressed in tweets in Portuguese using BERTimbau and an interpretation mechanism that obtains the most relevant words in terms of attention weights for model decision-making. The interpretability mechanism is targeted at users who do not necessarily have a strong knowledge of the internal workings of BERT and may use it to gain insights on influential words used in the predictions.

With regard to the case study, our assessment was positive. We got promising results in stance classification using BERTimbau for the Portuguese language, where the BERT fine-tuned model yielded the best results (F1 = 75.2%). The *Anti-vaxxers* was the most difficult class to predict, and using the interpretability mechanisms proposed, we hypothesize two reasons: a) errors are due to similar arguments used by *Anti-sinovaxxers*, and b) BERT is considering various generic words used in expressions by the other groups as influential for a particular one.

The absolute word attention metric provided the more relevant insights, compared to the relative one. We find the results promising since we could trace their semantics back to the domain and the representative stance arguments. In addition, the proportional attention weight enabled the identification of the words that contribute to hits in the classification (about 70%). However, we also identified words with high absolute attention contributing to the classifier's wrong predictions. Our findings make us hypothesize that this is due to the common tokens that exist between representative words of different classes. The experiments also provided insights on the tokens resulting from BERTimbau and respective vocabulary, providing directions for future aggregation of individual tokens weights, and means to consolidate word attentions regarding the dataset.

Future work includes, among others, (a) improving the stance classification model, (b) comparison with other BERT models for the Portuguese language (e.g., multilingual), (c) exploring the use of token-free alternatives [Clark et al. 2021], (d) increasing the size of the BERTimbau vocabulary, (e) alternative to aggregate tokens' attention into words and new word attention metrics, and (f) exploring the behavior of our framework in case studies from other domains.

Acknowledgments: This research is partially supported by CNPq (131178/2020-2), CAPES (Código de Financiamento 001) and FAPERGS (19/2551-0001862-2).

References

Abnar, S. and Zuidema, W. (2020). Quantifying attention flow in transformers. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.

ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing Management*, 58(4):102597.

Chefer, H., Gur, S., and Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Clark, J. H., Garrette, D., Turc, I., and Wieting, J. (2021). Canine: Pre-training an efficient tokenization-free encoder for language representation. *arXiv preprint arXiv:2103.06874*.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186.

Ebeling, R., Régis, C. C., Nobre, J. C., and Becker, K. (2022). 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.*

Ebeling, R., Sáenz, C. C., Nobre, J. C., and Becker, K. (2020). Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 89–96. SBC.

Giorgioni, S., Politi, M., Salman, S., 0001, R. B., and Croce, D. (2020). 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)*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jain, S. and Wallace, B. C. (2019). 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*, pages 3543–3556.

Kawintiranon, K. and Singh, L. (2021). 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*, pages 4725–4735.

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., and Robnik-Šikonja, M. (2021). 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*, pages 16–21.

Molnar, C. (2019). *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Wiegreffe, S. and Pinter, Y. (2019). 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)*, pages 11–20.