Should We Translate? Evaluating Toxicity in Online Comments when Translating from Portuguese to English

JORDAN K. KOBELLARZ, Universidade Tecnológica Federal do Paraná, Brazil
THIAGO H. SILVA, Universidade Tecnológica Federal do Paraná, Brazil

Social media and online discussion platforms suffer from the prevalence of uncivil behavior, such as harassment and abuse, seeking to curb toxic comments. There are several approaches to classifying toxic comments automatically. Some of them have more resources and are more advanced in English, thus, stimulating the task of translating the text from a specific language to English. While researchers have shown evidence that this practice is indicated for certain tasks, such as sentiment analysis, little is known in the context of toxicity identification. In this research, we assess the performance of a freely available model for toxic language detection in online comments called Perspective API, widely adopted by some famous news media sites to identify different toxicity classes in online comments. For that, we obtained comments in Portuguese from two Brazilian news media websites during a politically polarized situation as a use case. Then, this dataset was translated to English and compared to four baseline datasets, two composed of highly toxic comments, one in Portuguese and other in English, and two composed of neutral comments, also one in Portuguese and other in English – all of them in its original language, not translated. Finally, human-annotated comments from the news comments dataset were analyzed to assess the scores provided by the Perspective API for the original and the translated versions. Results indicate that keeping the texts in their original language is preferable, even in comparing different languages. Nevertheless, if the translated version is strictly necessary, ways of dealing with the situation were suggested to preserve as much information as possible from the original version.

CCS Concepts: • **Applied computing** → **Document management and text processing**; • **General and reference** → *Evaluation*; • **Information systems** → *Web mining*; *Data mining*; • **Human-centered computing** → Empirical studies in collaborative and social computing.

Additional Key Words and Phrases: online comments, toxicity, translation, natural language, Perspective API

ACM Reference Format:
Jordan K. Kobellarz and Thiago H. Silva. 2022. Should We Translate? Evaluating Toxicity in Online Comments when Translating from Portuguese to English. In *Brazilian Symposium on Multimedia and Web (WebMedia '22), November 7–11, 2022, Curitiba, Brazil.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3539637.3556892

1 INTRODUCTION

Uncivil online behavior, such as harassment and abuse, discourages healthy interactions, leading to conflict and unpleasant experiences [16]. A special case is represented by rude, disrespectful, or irrational comments that can lead users to leave discussions [13]. Given the speed with which discussions are growing on the Web [16], different models for large-scale toxicity identification were created to solve specific challenges, such as for dealing with online comments [1, 27]. Among the challenges is the fact that this type of text contains varying degrees of subtlety inherent to the language, cultural aspects, specificities of context, presence of sarcasm, and use of figures of speech, which can mask the

Manuscript submitted to ACM

actual toxicity of a comment. Other challenges include the fact that online comments are usually short texts, containing spelling errors that occur sparsely in the dataset [27], making complex the creation of models with good generalization ability to identify toxic instances. Furthermore, machine learning-based models, such as those used for toxicity analysis, tend to be susceptible to adversarial attacks, in which a malicious user can adjust their comment, exchanging toxic words for a variant easily recognized by a human, but undetectable by the model, e.g., replacing the word "stupid" with "st.Up1d" [11]. Such a change in the input can cause a disturbance in the model's output, causing a toxic comment to be recognized as non-toxic [11].

These challenges are even greater when it is necessary to classify and compare the degree of toxicity between comments in different languages [27]. In this sense, a multilingual model was made available for free through a Google initiative called Perspective API[1]. This model can be accessed through a public API, which allows identifying different toxicity types in online comments. Given that this is a widely adopted tool by major news outlets for moderating comments on their portals, in addition to the fact that it is openly available, Perspective API is a potential candidate for applications in research involving the study of online abusive behavior in large-scale.

It is not uncommon for text analysis tools to work only for English or to have more resources available for this language. This is the case of Perspective API, where certain attributes returned by the API only work for English content. Furthermore, textual analysis in English is, in many cases, more developed, in terms of training instances and research, which can provide more accurate results [2, 17, 22]. In this sense, for example, the Perspective API presents different results for different languages, and demands that during the processing of a textual instance, the language of the comment is included as a parameter. This motivates translating content in a particular language into English before performing specific automated text analysis tasks.

While for some tasks there is evidence that the English translation process is recommended, for example, in sentiment analysis [2], little is known about the impact of this practice on the analysis of toxic comments. Therefore, this work aims to compare the model's performance in detecting toxicity provided by the Perspective API in comments in Brazilian Portuguese with its respective version automatically translated into English. For this, we used as a case study comments in Brazilian Portuguese on political content published during the 2018 Brazilian presidential elections by two news media sites identified in the literature for their high capacity to distribute diverse content among politically polarized groups [14]. This dataset was translated and compared, according to its degree of toxicity, with other baseline datasets containing highly toxic and non-toxic comments written, in Brazilian Portuguese and United States English, separately.

The results show that:

- the translation process artificially reduces the overall toxicity of the dataset by penalizing highly toxic comments in its original language. Thus, whenever possible, the recommendation is to keep the text in its original language, even in comparisons between different languages;
- the perception of toxicity is not consensual among volunteers and is challenging to assess objectively, thus, highlighting the challenges involved in analyzing toxicity in online comments;
- Perspective API delivered an acceptable performance in identifying toxicity in online comments written in Brazilian Portuguese.

The rest of this study is organized as follows. Section 2 presents the related work. Section 3 presents the methodological steps used in this study. Section 4 presents and discusses the results. Finally, Section 5 concludes the study and presents potential limitations and future work.

[1]https://perspectiveapi.com.

2 RELATED WORK

There is a considerable history of automated detection of uncivil behavior online. Regarding abusive language, it is possible to find from keyword-based proposals, in which text containing potentially abusive keywords are identified [10], to strategies that use machine learning [1, 8, 20, 27–29]. Specifically about toxicity in online comments, a model widely used in several academic works and by the industry is the Perspective API [9, 24, 25].

Online toxicity varies depending on target groups – e.g., terms used to express hatred against a community in Brazil are different from terms used against Latin Americans in the United States – and context – e.g., the text of hate about the indigenous community will probably be different in the context of a pro-Bolsonaro discussion and a discussion in the context of Funai[2] supporters. Furthermore, machine learning-based techniques require large amounts of high-quality annotated data, which can raise questions about model training [7].

Therefore, it is possible to find several works that focus on studying and criticizing toxicity classification approaches [7, 12]. However, to the best of our knowledge, there is no study focused on evaluating the impact of toxicity when translating comments, a task with several motivations to be performed, as we mentioned above.

Related to this issue, several studies have evaluated the impact of the English translation analysis in the context of sentiment analysis [26]. For example, Araujo *et al.* [2] found that simply translating the text of interest in a specific language into English and then using one of the best existing methods developed for English may be better than the existing language-specific approach evaluated. While there are many examples in favor of translation for sentiment analysis, some studies are pointing out negative aspects of this practice. In the context of language-specific knowledge, which includes abbreviations, slang, and emojis, Chen [4] argued that even for more resourceful languages, translation only captures generic patterns shared across languages and fails to gain language-specific sentiment knowledge. This could be a problem if the dataset has considerable specific knowledge. It is also possible to find proposals evaluating translation's impact in the context of topic extraction [5, 19]. For example, the authors of [5] found that results for translated content are quite similar to results for content in the original language.

In order to contribute to the gap identified in the literature, our study focuses on assessing toxicity in online comments when translated from Portuguese to United States English.

3 METHODOLOGY

The method applied in this research includes several steps to allow comparison between datasets. These steps are explained in the following sections.

3.1 Data collection and sampling

The base dataset used in this research was obtained by extracting comments from news articles regarding the 2018 Brazilian presidential election published in two news media sites: the G1 portal from Globo television network (g1.globo.com) and the UOL Notícias portal (noticias.uol.com.br). The news articles were selected from political messages containing links to these news media sites that were shared on the Twitter social network during the electoral period, starting 6 days before the first-round vote, from 2018-Oct-01, and ending 15 days after the second-round vote, on 2018-Dec-11 [14]. Both news media sites were identified in the literature as channels that could distribute content to individuals with different political orientations more efficiently than other similar sites during this polarized political event [15], potentially bypassing the filter bubbles [14]. For naming conventions, this dataset is referred to as **BRA_pt**. To be able

[2]Brazilian governmental protection agency for Amerindian interests and their culture.

to compare the toxicity between the original Brazilian Portuguese version of this dataset with its respective version translated into English, it was used the Microsoft Azure translation API[3] - the translated version of BRA_pt is referred to as **BRA_en** in this article.

As the objective of this research is to compare the toxicity identified in the BRA_pt and BRA_en datasets, four baseline datasets were obtained; two of them composed of highly toxic comments, one in English (TOXIC_en) and another in Brazilian Portuguese (TOXIC_pt), and the other two containing less toxic (neutral) comments, one in English (NEUTRAL_en) and another in Brazilian Portuguese (NEUTRAL_pt). These baseline datasets are presented below.

TOXIC_en: contains highly toxic United States English comments (not translated) taken from an open dataset with human-labeled Wikipedia comments according to different categories of toxic behavior[4], including "toxic", "severe_toxic", "obscene", "threat", "insult", and "identity_hate". The variable "toxic", representing the degree of toxicity of the comment, was used to select 5,000 comments with the highest value for this metric and compose the TOXIC_en dataset.

TOXIC_pt: contains highly toxic comments in Brazilian Portuguese (not translated) obtained from tweets manually annotated according to different toxicity categories in a publicly available dataset [18]. The available categories were "non-toxic", "LGBTQ+ phobia", "obscene", "insult", "racism", "misogyny", and "xenophobia" [18]. To select a representative sample, the number of toxicity categories linked to each comment was counted, except for the "non-toxic" category. This count was used to select 5,000 comments with the highest amount of toxic categories to compose the TOXIC_pt dataset.

NEUTRAL_en: contains non-toxic (neutral) comments in United States English (not translated) obtained through Reddit's public API, a network of communities where people with common interests interact in a forum-like system. To collect the data, the most famous 100 posts from the communities (*subreddits*) \AskHistorians, \changemyview, \COVID19, \everythingScience, and \science were selected. These communities were chosen because they contain potentially non-toxic discussions and the seriousness observed in the responses, given that stricter rules for posting were explicitly informed and seemed to be followed by their participants. Thus, this dataset was composed mostly of constructive comments. In addition to the text of the comments, other attributes were obtained, among them the "score" of the comment, which is a metric calculated by subtracting the negative votes from the positive votes that a given comment received. This metric was used to select 5,000 comments with the highest scoring to compose the NEUTRAL_en dataset.

NEUTRAL_pt: composed of non-toxic (neutral) comments in Brazilian Portuguese (not translated) obtained from a database of product reviews in a famous e-commerce business, B2W Digital, responsible for the *americanas.com* website, whose data were obtained between January and May 2018 [23]. This dataset was selected considering that among positive and lengthy product reviews, this dataset would have a less toxic vocabulary. Therefore, during data cleaning, evaluations with a score lower than 5 and that contained less than 20 unique characters were eliminated. This is important because the incidence of texts with repeated words was identified in cases where the evaluator only filled in the text field with no intention of making a careful assessment. After cleaning, the longest 5,000 evaluations were selected to compose the NEUTRAL_pt dataset.

For reproducibility purposes, all data used in this study are available on the research project website: https://sites.google.com/view/onlinepolarization.

[3]https://azure.microsoft.com/en-us/services/cognitive-services/translator.
[4]https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data.

3.2 Pre-processing

The pre-processing step was performed to maintain maximum integrity in the datasets presented in the previous section, only removing noisy instances that could negatively impact performance during toxicity identification. Escape sequences (such as line breaks), Markdown tags, and links have been removed. Sentences repeated in more than one comment, such as "Reply to @" and "REPORT COMMENT," which are clearly not part of the construction of a comment, were also identified and removed.

Since online comments tend to be noisy, usually containing spelling errors, which was possible to verify by analyzing a sample of random comments in each dataset, it was necessary to apply an autocorrection step of words. For this task, an algorithm called *SymSpell* [5] was applied, which depends on the creation of a dictionary with the correct number of words that will be used to replace those whose spelling does not conform to the dictionary standard. Thus, a dictionary was created for each dataset, including the words that appeared at least 5 times in the respective dataset, given that more frequent tokens tend to be the correct version. In contrast, rare tokens tend to be spelling errors. Then, into this initial dictionary, it was concatenated a standard dictionary of the language of each dataset, Brazilian Portuguese for the BRA_pt, TOXIC_pt, and NEUTRAL_pt datasets, and United States English for the BRA_en, TOXIC_en, and NEUTRAL_en datasets. These dictionaries were obtained from the OpenOffice repository[6]. The resulting dictionary is then processed by *SymSpell*, which generates permutations of the words through the character deletion procedure, resulting in a dictionary of permutations with words that would potentially be misspellings. After this procedure, the comments' autocorrection was performed. In this step, *SymSpell* was configured to autocorrect only words with a maximum edit distance of 2 characters (number of characters needed to turn a misspelled word into the correct word). It was observed that the increase to a maximum edit of 3 or more characters generated wrong corrections. Short words with wrong spelling were more likely to be replaced by words with correct spelling but unrelated to the corrected word. It is important to mention that the autocorrection step does not eliminate the possibility of interference from adversarial attacks [11], but it can reduce its incidence.

Considering that, even with the autocorrection procedure, the comments could still present noise, mainly caused by rare or unusual words, a pre-processing step was added to remove comments with many poorly recognized instances. For this, the comments were processed by a tool widely applied in the literature to recognize linguistic, psychological, and social characteristics called LIWC [21]. This tool was chosen because it has good coverage for several dictionaries of different languages and does not require complex steps in the pre-processing of the text to be analyzed. For datasets in English, the official internal LIWC dictionary was used, and an unofficial dictionary [3][7] for datasets in Brazilian Portuguese. After processing the datasets with LIWC, the "Dic" attribute was used to keep only comments with at least 50% of the words recognized by the respective LIWC dictionary – this attribute counts the percentage of words identified in the respective dictionary. Finally, considering that short comments can influence toxicity results, all comments containing less than 10 words were removed using LIWC's WC (Word Count) parameter, which counts the number of words identified in a text. Table 1 shows the count of the initial number of comments (Original size), the count after cleaning (Final size), and the percentage of comments that were removed after cleaning the data (% final). In this table, it is possible to verify that the impact of cleaning noisy instances was greater for the pairs BRA_pt and BRA_en; despite this, the dataset size is still large enough to consider in the analyses.

[5] https://symspellpy.readthedocs.io.
[6] https://www.openoffice.org/lingucomponent/dictionary.html.
[7] http://143.107.183.175:21380/portlex/index.php/pt/projects/liwc.

Table 1. Datasets size before and after cleaning.

Dataset	Original size	Final size	% final
BRA_pt	128,898	95,856	74.37(%)
BRA_en	128,898	96,827	75.11(%)
NEUTRAL_pt	5,000	4,989	99.78(%)
NEUTRAL_en	5,000	4,976	99.52(%)
TOXIC_pt	5,000	4,718	94.36(%)
TOXIC_en	5,000	4,927	98.54(%)

The next step after dataset cleaning is identifying toxicity in comments across all datasets. For this, all comments were pre-processed by the Perspective API. The results are presented in the next section.

3.3 Inferring toxicity: Perspective API

The Perspective API is a toxic comment classification model to help improve online conversations. This classifier assigns a continuous score between 0 and 1 to comments. A higher score indicates a greater likelihood that a reader will perceive the comment as containing the given attribute, e.g., toxicity [13]. For example, as presented in the API documentation, a comment like "You are an idiot" may receive a probability score of 0.8 for the TOXICITY attribute, indicating that 8 out of 10 people would perceive that comment as toxic [13]. With this, it is possible to use this score, for example, to remove toxic comments with a certain score [13]. Perspective architecture is composed by multilingual BERT-based models trained on data from online forums, that are distilled into single-language Convolutional Neural Networks (CNNs) for each language that they support – distillation ensures the models could be served and produce scores within a reasonable amount of time [13]. Perspective has so-called production attributes, tested in various domains and trained on significant amounts of human-annotated comments; these attributes are available for English, Portuguese, and many other languages. Also, it contains experimental attributes – English only – that are not recommended for professional use at this time [13]. Given the comment "" form example,

In this work, we focus on production attributes:

- TOXICITY - "A rude, disrespectful or irrational comment that is likely to cause people to leave a discussion".
- SEVERE_TOXICITY - "A comment that is too hateful, aggressive, disrespectful, or too likely to make a user leave a discussion or give up sharing their perspective. This attribute is much less sensitive to milder forms of toxicity, such as comments that include positive uses of profanity".
- IDENTITY_ATTACK - "Negative or hateful comments directed at someone because of their identity".
- INSULT - "Offensive, inflammatory or negative comment towards a person or a group of people".
- PROFANITY - "Swearing, profanity or other obscene or profane languages".
- THREAT - "Describes the intent to inflict pain, injury or violence against an individual or group".

4 RESULTS

Figure 1 shows a box plot to help understand the difference in the scores obtained for each perspective API attribute regarding the studied datasets. It is possible to see that the pairs of TOXIC_pt and TOXIC_en are similar to each other, as well as the pairs NEUTRAL_pt and NEUTRAL_en, which indicates that these pairs share characteristics in

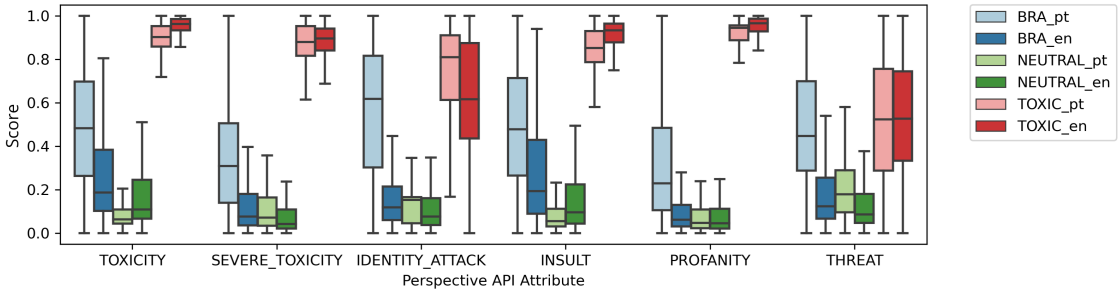

Fig. 1. Box plots showing the distribution of the scores of each Perspective API attribute for the considered datasets.

common, despite linguistic and data source differences. The pairs TOXIC_pt and TOXIC_en proved to be the most toxic, as expected, mainly concerning the attributes TOXICITY, SEVERE_TOXICITY, INSULT, and PROFANITY. The pairs NEUTRAL_pt and NEUTRAL_en proved to be the least toxic with all attributes, as also expected. These results indicate that the baseline pairs present desired characteristics for comparison with BRA_pt and BRA_en. In addition, the slight difference between the baseline pairs indicates, even if weakly, that the language may have a low influence on the toxicity presented by the Perspective API, considering these extreme cases.

When analyzing the pairs referring to the original version of the original dataset in Portuguese (BRA_pt) and its translation to English (BRA_en), it is possible to notice a big difference for any of the attributes of Perspective API, with toxicity being consistently lower in the version BRA_en. One possibility is that Perspective API may be inflating toxicity scores for Portuguese comments or doing the opposite with English comments, regardless of whether it is translated or not. However, this possibility is less likely to be the problem, given that a low difference in the comparisons between baseline pairs was identified between the Portuguese and English versions of such extreme cases. The other possibility is that machine translation decreases toxicity, either by eliminating toxic sentences or replacing them with less toxic ones. The objective assessment of these possibilities on large-scale is not a trivial task. Despite this, some analyzes were carried out to evaluate in which cases, original or translated, the Perspective API performed better. To perform these analyses, the absolute difference between the score of the TOXICITY attribute of the original version of each comment in BRA_pt and its respective translated version in BRA_en was calculated. This metric is referred to as "Diff" and indicates the disparity between the original and translated versions.

The first analysis was conducted to understand how TOXICITY and Diff values were distributed among the comments. For this, the histograms in Figure 2 show the number of comments according to the TOXICITY score for the datasets BRA_pt and BRA_en – the first two histograms, respectively –, and the number of comments according to the Diff value – on the third histogram. A notable characteristic in the first two histograms is that the toxicity scores of the original dataset (BRA_pt) are more evenly distributed among the comments. In contrast, they are concentrated close to zero in the translated dataset. This could indicate a loss of information during the translation or that the Perspective API may be returning low TOXICITY scores for texts in English. The last graph, in turn, shows that most comments had a low difference between the score of TOXICITY in the original and the translated version, between 0.0 and 0.5, with a higher concentration of comments with a Diff close to 0.0. This indicates that in most cases, the disparity between the scores was small; even so, the volume of cases in which there was greater disparity is considerable.

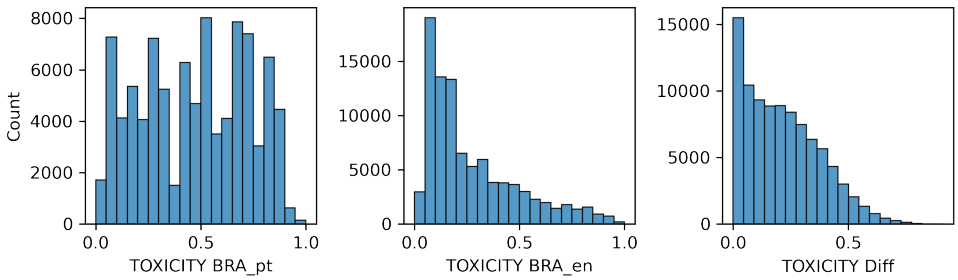

Fig. 2. Histograms containing the number of comments according to the TOXICITY score for the BRA_pt and BRA_en datasets (first two graphs, respectively) and the number of comments according to the Diff value (last graph).

To better understand in which cases there was high or low disparity, the relationship between TOXICITY scores was analyzed using scatter plots, shown in Figure 3. Each graph in this figure presents the TOXICITY score for the original and translated versions on the X and Y axes, respectively. The first graph, from left to right, presents these scores, including all comments. Following, the graphs in the sequence present the cuts made using the Diff value so that it was possible to capture only comments in which the difference between the original and translated versions was less than a certain threshold – displayed above each graph – being < 0.5, < 0.25, and < 0.1 in the second, third, and fourth graph, respectively. The red regression line indicates the trend in each case according to the cuts made.

An important characteristic observed in the first graph is the tendency of the TOXICITY score to be more inflated in the original version (BRA_pt). This result is in congruence with what was observed in the first two graphs of Figure 2, and makes it clear that the points where the TOXICITY score was reduced the most after translation were in cases where this score was high before translation – this indicates that more toxic cases tend to be penalized more after translation. This same characteristic does not apply to less toxic cases, which generally did not suffer a high impact after translation. Another notable characteristic, when cutting by the Diff value, is the reduction in the size of the dataset, given by the value of N presented in the title of each graph – the more aggressive the cut, the greater the loss of information. This analysis, in particular, is important to demonstrate the number of comments that preserved their toxicity according to different Diff values between the original and the translated versions. This result can be helpful in tasks where the translation step is crucial and significant differences caused by the translation step cannot be admitted. In this sense, it is important to consider that the greater the difference used in the cut, the greater the loss of highly toxic comments – which can also be a determining factor in the decision between translating or not.

To better understand the cases of high or low toxicity or with a high difference, manual analyzes of comments were performed according to the TOXICITY score. For this, four samples were extracted: (i) 100 random comments among the $10,000$ with the highest TOXICITY score for the Portuguese version (MOST_TOXIC); (ii) 100 random comments among the $10,000$ with the lowest TOXICITY score for the Portuguese version (LESS_TOXIC); (iii) 100 random comments out of the $10,000$ with the biggest difference between the TOXICITY score in the Portuguese and English versions (HIGHEST_DIFF); and (iv) 100 random comments across the entire dataset (RANDOM). After that, we invited three volunteers – women and men aged 25 to 40, who have a background in exact sciences or social sciences, all have a bachelor's degree, and two have a graduate degree. They received the original Portuguese version of the comment and both the TOXICITY score obtained for the original and translated version. Each volunteer was instructed to read the comment and choose the toxicity score that best applied to the text, without them knowing whether the score was

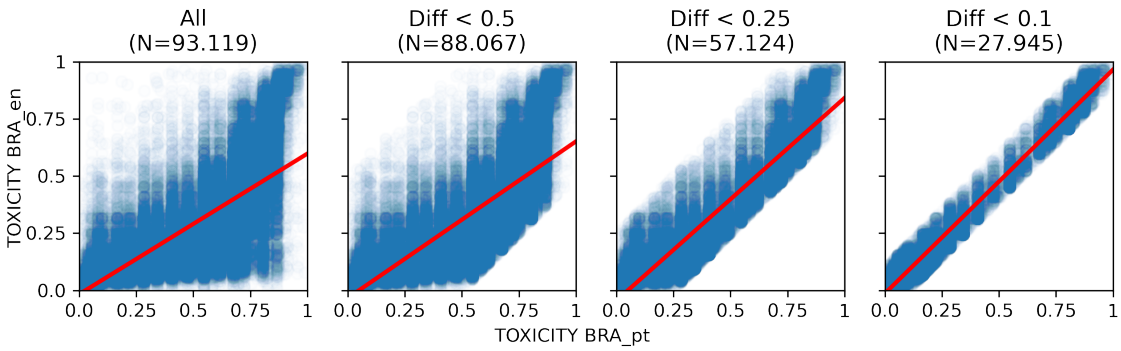

Fig. 3. Scatter plots depicting the ratio of the TOXICITY score of the original and translated version on the X and Y axes. Each plot is cropped according to different Diff values presented in its title, including the number of comments that were preserved in N. A red regression line shows the tendency of the TOXICITY score to favor the original (BRA_pt) or the translated version (BRA_en).

for the Portuguese or English version, so as not to bias their classification. For the same purpose, the comments were randomly organized, preventing possible influence on the task due to the order of presentation. After all comments were classified, the Fleiss' Kappa [6] metric was used to measure the degree of agreement between the volunteers.

Table 2 presents each one of the samples and the respective values of Fleiss' Kappa (κ), as well as the percentage of classifications made in favor of the Portuguese version (pt %) and in favor of the translated version (en %). In this table, cuts were also made by the Diff value to verify if the agreement between the volunteers was altered according to the difference in the toxicity scores. It is possible to observe that the agreement between the evaluators was always low, especially in the case of the less toxic sample – the low agreement, in this case, is because most of the comments in this sample have a tiny difference between the toxicity score presented for the original version and the one presented in the translated version; this is clear when analyzing the value of N in any case where Diff is greater than 0.1, so only the version without cut by the value of Diff should be considered for the case of LESS_TOXIC sample.

Another notable feature is that the original Portuguese version generally had the highest percentage of ratings in its favor, regardless of the sample. The only sample in which the toxicity score for the translated version obtained the highest percentage of ratings in its favor was the case of the MOST_TOXIC sample with Diff > 0.5. This case also had a low agreement between the volunteers, and the sample size is small ($N = 9$), so it can be disregarded. These characteristics indicate that, despite the low agreement among the volunteers, the comments in their original Portuguese version tended to receive a TOXICITY score more assertively by the Perspective API. It is also important to note that in the case of the sample with the greatest difference between the original and the translated versions (HIGHEST_DIFF), there was a greater agreement between the volunteers. However, the percentage favoring the original and translated versions was very close, indicating that the Perspective API can perform well even with translated texts. Still, the text in its original language is recommended to be used instead of the translation, if that is possible.

The graph shown in Figure 4 was created to understand the cases of disagreement between volunteers. It presents the count of ratings made individually by each volunteer ($V1$, $V2$, and $V3$), made by two volunteers ($V1 \cap V2$, $V1 \cap V3$, and $V2 \cap V3$), or unanimously by all volunteers ($V1 \cap V2 \cap V3$). A prominent characteristic is that in all samples where the three volunteers agreed ($V1 \cap V2 \cap V3$), the toxicity score with the most votes was BRA_pt, which is aligned with what was indicated in Table 2. For example, in the case of the LESS_TOXIC sample, there were 23 votes in

Table 2. Perspective API performance analysis for each sample according to manual classification performed by three volunteers. The table shows the cuts made in each sample by the Diff value and the respective Fleiss' Kappa values – indicating the agreement between the volunteers – and the percentage of favorable classifications for the TOXICITY score for the Portuguese version (BRA_pt %) or English (BRA_en %). It is essential to highlight that the low agreement between the volunteers in the LESS_TOXIC sample was caused by a large number of comments with a very low Diff, which made the volunteers' classification task difficult in this case – this is evident by looking at the sample size (N) for cases where Diff is greater than or equal to 0.1.

				TOXICITY	
Sample	**Diff**	**N**	κ	**BRA_pt %**	**BRA_en %**
LESS_TOXIC	–	100	0.06	59.3%	40.7%
	> 0.1	9	-0.16	74.1%	25.9%
	> 0.2	4	-0.33	75.0%	25.0%
	> 0.3	1	-0.50	66.7%	33.3%
	> 0.4	0	0.00	0.0%	0.0%
	> 0.5	0	0.00	0.0%	0.0%
MOST_TOXIC	–	100	0.23	52.0%	48.0%
	> 0.1	58	0.17	53.4%	46.5%
	> 0.2	31	0.08	55.9%	44.1%
	> 0.3	25	0.01	58.7%	41.3%
	> 0.4	14	-0.06	54.8%	45.2%
	> 0.5	9	-0.20	44.4%	55.6%
HIGHEST_DIFF	–	100	0.33	53.3%	46.7%
	> 0.1	100	0.33	53.3%	46.7%
	> 0.2	100	0.33	53.3%	46.7%
	> 0.3	100	0.33	53.3%	46.7%
	> 0.4	100	0.33	53.3%	46.7%
	> 0.5	52	0.35	56.4%	43.6%
RANDOM	–	100	0.18	57.7%	42.3%
	> 0.1	75	0.15	59.1%	40.9%
	> 0.2	55	0.13	57.0%	43.0%
	> 0.3	36	0.10	62.0%	38.0%
	> 0.4	19	0.03	68.4%	31.6%
	> 0.5	13	0.24	71.8%	28.2%

favor of the original version and 9 for the translated version, which is a high difference. This difference was smaller for the MOST_TOXIC, HIGHEST_DIFF, and RANDOM samples. Another characteristic presented in this figure is the considerable amount of cases in which the volunteers disagreed with the majority, represented in the cases $V1$, $V2$, and $V3$ in each graph. In this sense, there are some situations in which there was a clear imbalance between the individual votes; for example, in the LESS_TOXIC and RANDOM samples, it is possible to notice that $V1$ voted more than the double times in favor of the English score. In contrast, in the HIGHEST_DIFF sample, this same volunteer did the opposite, favoring the Portuguese score. This characteristic is important since it exposes the subjectivity in this type of analysis, making the task of detecting toxicity complex.

To illustrate some representative cases, Table 3 was built, presenting the original and translated comments and their respective TOXICITY scores. Values in bold and underlined represent the TOXICITY score, in which the three volunteers unanimously agreed that it would best fit the text of the comment. The first example comment in the

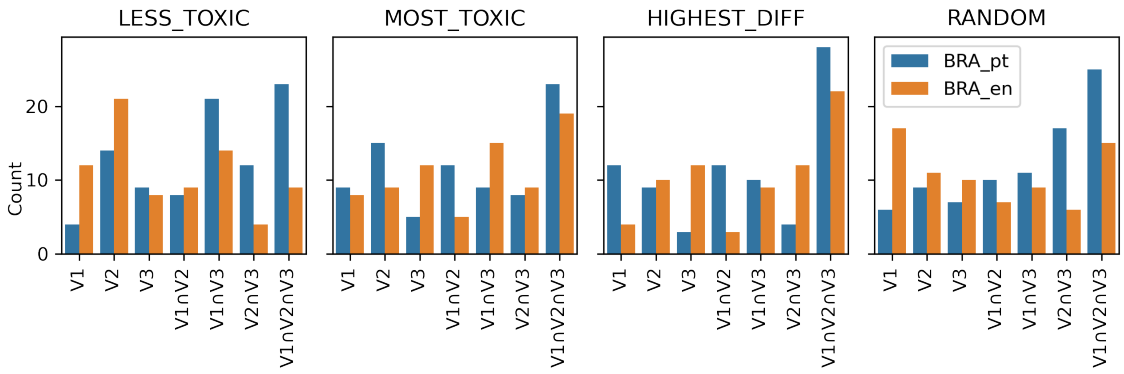

Fig. 4. Count of ratings made individually by each volunteer ($V1$, $V2$, and $V3$), made by two volunteers simultaneously ($V1 \cap V2$, $V1 \cap V3$, and $V2 \cap V3$), or unanimously by all volunteers ($V1 \cap V2 \cap V3$).

LESS_TOXIC sample was correctly translated into English; however, the translated version had a slightly higher toxicity score. A possible cause for this was translating the word "goleada" to "rout", which can also mean defeat, turmoil, or confusion. The second example in this sample had problems translating the word "viu", and the overall translation was slightly incoherent but understandable. Even so, the toxicity scores were similar between the original and the translated versions. Regarding the MOST_TOXIC sample, in the first example, the translation was slightly confusing but understandable, preserving the threat tone. Despite this, the toxicity was lower in the translated version, not reflecting the degree of toxicity that the volunteers judged to be more in line with the text of the comment. In the second example for this sample, the translation was done properly, in addition to generating a toxicity score that was more appropriate for the text, according to the volunteers.

In the case of the RANDOM sample, the first and second examples were adequately translated into English; even so, they showed a significant disparity between the toxicity scores. In this case, it was impossible to identify a clear pattern that would justify the value disparity. Finally, regarding the HIGHEST_DIFF sample, the first example demonstrates a case of an adversary attack, in which the words "LA-DRÃO" (robber), "V.ERDADE" (true), "C.EITA" (sect), and "M.4LD.Y.T.A." (damn) were not translated correctly, therefore, they may have been the cause of the lower toxicity score in the translated version. Despite that, the toxicity score for the original version appears to have been correctly inferred, even with the intentional change made by the author of this comment to mask the toxicity. This is a sign that the Perspective API may have been updated to try to resolve adversarial attack cases, a problem reported by Jain *et al.* [12] that can influence the API results. The second example of the HIGHEST_DIFF sample was also correctly translated into English, and the most acceptable toxicity score was from the translated version, according to the volunteers. This case seems to have inflated the toxicity score for the Portuguese version, so the volunteers ended up choosing the English toxicity score as the most appropriate.

5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

The results presented in this research show that automatic translation can cause comments considered to be very toxic in their original language - Brazilian Portuguese, in the case of this research - to have their toxicity score reduced. This does not happen with low-toxic comments, which, after translation, tend to keep their original characteristics,

Table 3. Examples of representative comments on each of the samples. TOXICITY scores in bold and underlined represent the scores at which the three volunteers unanimously agreed that it would best fit the text of the comment.

Sample	Comment		TOXICITY	
	BRA_pt	BRA_en	BRA_pt	BRA_en
LESS_TOXIC	No debate para governador do Rio realizado na Globo dia 25, o Eduardo Paes deu uma goleada, merece ser eleito.	In the debate for governor of Rio held at Globo on the 25th, Eduardo Paes gave a rout, deserves to be elected.	**0.082**	0.245
	O Bolsonaro não "alegou "recomendação médica, os próprios médicos deram entrevista recomendação que ele não fossea.a.todo mundo viua.a.	Bolsonaro did not "claim "medical recommendation, the doctors themselves gave interview recommendation that he was not.a.everyone viua.a.	0.100	**0.074**
MOST_TOXIC	tirso seu bebado, não sou macho só na net, venha aqui ou me passa seu endereço se tu é homem que vou ai te falar pessoalmente o meu é rua joão mania ,Santa Candida,Curitiba, (é bem perto da atual presidencia do teu patrão luladrao) venha aqui falar essas bobagens	tirso your drunk, I'm not male only on the net, come here or give me your address if you're man I'll tell you personally my is street john mania , Santa Candida, Curitiba, (it's very close to the current presidential boss luladrao) come here talk this nonsense	**0.861**	0.547
	KKKKKK agora o DataFAKE faz uma pesquisa minimamente verdadeira, porque já estava ficando mala.a.kkkkkkkkkk ridículos! Tentaram manipular o voto do brasileiro só que isso acabou!! só perderam o resto da credibilidade que tinham, se é que tinham alguma!	KKKKKK now DataFAKE does a minimally true search, because it was already getting mala.a.kkkkkkkkkk ridiculous! They tried to manipulate the brazilian vote only that this is over! they just lost the rest of their credibility, if they had any!	0.817	**0.676**
RANDOM	Pq voce não pergunta o seguinte: o PT estava 13 anos no Poder, o que fez além de AFUNDA O BRASIL, COM A MAIOR CORRUPÇÃO DO UNIVERSO ?	Why don't you ask the following question: the PT was 13 years in power, what did besides SINK BRAZIL, WITH THE GREATEST CORRUPTION OF THE UNIVERSE ?	**0.637**	0.097
	Encontrei funcionários descontentes por serem obrigados a mentir para o governo de Márcio França e que fez "tudo" de mentira, é óbvio, em 3 meses de governo, com dinheiro público. Estes dias ele culpa os governos anteriores quando cobrando por eventual erros , isto é Alckimin, Serra e Covas. Não sou do exército, mas vcs são do policiamento contrário. ¨Compramos 20 mil colete, milhares de carros, construímos dezenas de hospitais a.a.estamos tomando 10 ou mil obras públicas, está dizendo agora. Isto é proibido por lei. O candidato Dória deixou um jovem correto, o Bruno Covas na prefeitura. Ele deveria esclarecer isso. O PSB apoia Hadad ou é mentira?	I found officials unhappy about being forced to lie to the government of Márcio França and who did "everything" lie, of course, in 3 months of government, with public money. These days he blames previous governments when charging for eventual mistakes, i.e. Alckimin, Serra and Covas. I'm not in the army, but you're the opposite policing. "We bought 20,000 vests, thousands of cars, built dozens of hospitals a.a.we're taking 10 or a thousand public works, it's saying now. This is prohibited by law. The candidate Doria left a correct young man, Bruno Covas in the city hall. He should clear that up. Does the PSB support Hadad or is it a lie?	0.545	**0.154**
HIGHEST_DIFF	NÃO FOI SÓ O P.ALOFI QUE M.ENTIU NÃO, FOI O MUNDO TODO. SÓ O LU-LA-DRÃO FALA A V.ERDADE. ESSA É A C.EITA M.4LD.Y.T.A. CHAMADA P-TE.	IT WAS NOT ONLY P.ALOFI THAT M.ENTIU NOT, IT WAS THE WHOLE WORLD. ONLY LU-LA-DRÃO SPEAKS TO V.ERDADE. THIS IS C.EITA M.4LD.Y.T.A. CALL P-TE.	**0.676**	0.243
	Uma grande vergonha considerar assim as instituições brasileiras, seja PT ou PSL. E uma grande vergonha apadrinhar o erro crasso dos outros só por conta das relações familiares. Isso é o câncer do Brasil. Quando vai acabar?	It is a great shame to consider brazilian institutions, whether PT or PSL. And a great shame to patronize the crass mistake of others just because of family relationships. This is the cancer of Brazil. When will it end?	0.805	**0.357**

preserving the toxicity score. These characteristics are fundamental in deciding whether or not to translate a set of data in tasks that involve comparing texts in different languages. This finding contrasts with Araujo *et al.* [2], that found that simply translating the text of interest in a specific language into English and then using one of the best existing methods developed for English may be better than the existing language-specific approach evaluated. At the same time, this finding corroborates the argument of Chen [4] regarding language-specific knowledge, which includes abbreviations, slang, and emojis. In this sense, even for more resourceful languages, translation only captures generic patterns shared across languages and fails to gain language-specific sentiment knowledge, being a problem if the dataset has a considerable specific knowledge, as in the case of *online comments*. The translation process, as seen, implies artificially reducing the overall toxicity of the dataset because of the loss of information in more toxic comments. In this sense, the results presented in this work suggest that, if possible, the texts should be kept in their original language, even in comparisons between different languages. If this is not possible and translation is required, a dataset cut like the ones shown in Figure 3 can help eliminate instances where there is a very high disparity between the toxicity scores in the original and translated versions. When cutting, it is important to consider information loss: the more aggressive the cut, the more highly toxic instances are lost, while the loss of less toxic instances is proportionately lower.

In addition to practical guidelines, the results of manual assessments and examples of representative comments illustrate the challenges involved in analyzing toxicity in online comments. For example, volunteers had a low agreement in several cases, which indicates that, even for humans, the perception of toxicity is not consensual and challenging to assess objectively. Furthermore, there are non-trivial challenges for a generalist model, like the case of Perspective API, such as the inherent characteristics of the language that are lost in the machine translation process, the context of the comment (political context in the case of this research), the intentionality of the person who commented (using sarcasm or masking toxic words to circumvent automatic moderation systems). In this sense, considering the model's restrictions, it can be said that Perspective API delivered acceptable performance, especially for the untranslated version of the comments.

In this direction, some limitations need to be highlighted. The fact that the baseline datasets were obtained from different sources makes comparisons less robust, since the dynamics of interactions on each platform can influence the degree of toxicity. For example, on Twitter, comments need to be short and do not go through a moderation process – although they can be reported. On Reddit, however, comments have no size limit and, in general, go through a moderation process through participant voting and being moderated by community leaders designated especially for this task. We also recognize that the pre-processing step, although not aggressive, may have eliminated some representative comments in the BRA_pt and BRA_en datasets; despite this, the number of comments remaining to be processed by the Perspective API was high ($N > 95,000$) which reduces the chance that the removal influenced the results. Another limitation refers to the polarization characteristics in the BRA_pt and BRA_en datasets - it was found that several comments had subtle characteristics of toxicity, mainly with the presence of sarcasm related to the political context itself, which would be complex to be captured by a generic model for detection of toxicity and even by humans without context-awareness. Therefore, the political orientation of the volunteers may also have impacted the lack of agreement between them, given that the toxicity scores presented by Perspective API for the original or translated version could not indicate the same degree of toxicity perceived by the volunteers according to their political bias. Lastly, regarding the translation step, even with the manual evaluation step by the volunteers, it was not possible to identify exactly whether the Microsoft Azure Translation API has influenced the results and its accuracy – in this sense, further analysis should be carried applying other translation tools to understand its accuracy and compare its impact on toxicity results.

Since the comparisons made in this research are limited to Portuguese (original) and English (translated) languages, a significant contribution can be made by doing the reverse process: translating a dataset initially in English into Portuguese. This analysis would complement the findings of this work and make the guidance given on the decision to translate or not in toxicity analysis tasks more robust. Furthermore, analyzing other languages in other contexts is also essential, since the same behavior may not be replicated in different settings. In this sense, the method applied in this work is advantageous as a framework to assess the suitability of machine translation in any textual dataset.

ACKNOWLEDGMENTS

All stages of this study were financed in part by CAPES - Finance Code 001, project GoodWeb (Grant 2018/23011-1 from São Paulo Research Foundation - FAPESP), and CNPq (grant 310998/2020-4).

REFERENCES

[1] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. *Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions.* Association for Computing Machinery, New York, NY, USA, 3033–3040. https://doi.org/10.1145/3366423.3380074

[2] Matheus Araújo, Adriano Pereira, and Fabrício Benevenuto. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences* 512 (2020), 1078–1102.

[3] Pedro P Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluisio. 2013. An evaluation of the Brazilian Portuguese LIWC Dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. Sociedade Brasileira de Computação, Fortaleza, CE, Brazil, 215–219.

[4] Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-Powered Representation Learning for Cross-Lingual Sentiment Classification. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 251–262. https://doi.org/10.1145/3308558.3313600

[5] Erik De Vries, Martijn Schoonvelde, and Gijs Schumacher. 2018. No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis* 26, 4 (2018), 417–430.

[6] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions* 2, 212-236 (1981), 22–23.

[7] Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6786–6794. https://aclanthology.org/2020.lrec-1.838

[8] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (Patras, Greece) *(SETN '18)*. Association for Computing Machinery, New York, NY, USA, Article 35, 6 pages. https://doi.org/10.1145/3200947.3208069

[9] Samuel S. Guimarães, Julio C. S. Reis, Filipe N. Ribeiro, and Fabrício Benevenuto. 2020. Characterizing Toxicity on Facebook Comments in Brazil. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (São Luís, Brazil) *(WebMedia '20)*. Association for Computing Machinery, New York, NY, USA, 253–260. https://doi.org/10.1145/3428658.3430974

[10] Hatebase. 2022. Hatebase. https://hatebase.org Acessed May 31, 2022.

[11] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[12] Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. 2018. Adversarial Text Generation for Google's Perspective API. *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (2018), 1136–1141.

[13] Google Jigsaw. 2022. Perspective API. https://perspectiveapi.com Acessed May 31, 2022.

[14] Jordan K Kobellarz, Milos Brocic, Alexandre R Graeml, Daniel Silver, and Thiago H Silva. 2021. Popping the Bubble May Not be Enough: News Media Role in Online Political Polarization. https://doi.org/10.48550/ARXIV.2109.08906

[15] Jordan K. Kobellarz, Alexandre R. Graeml, Michelle Reddy, and Thiago H. Silva. 2019. Parrot Talk: Retweeting among Twitter Users during the 2018 Brazilian Presidential Election. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web* (Rio de Janeiro, Brazil) *(WebMedia '19)*. Association for Computing Machinery, New York, NY, USA, 221–228. https://doi.org/10.1145/3323503.3349559

[16] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 933–943. https://doi.org/10.1145/3178876.3186141

[17] Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint arXiv:2202.11176* (2022).

[18] João A. Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. https://doi.org/10.48550/ARXIV.2010.04543

[19] Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. *Political Analysis* 23, 2 (2015), 254–277.

[20] Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. (Aug. 2017), 41–45. https://doi.org/10.18653/v1/W17-3006

[21] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).

[22] Denilson Alves Pereira. 2021. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review* 54, 2 (2021), 1087–1115.

[23] Livy Real, Marcio Oshiro, and Alexandre Mafra. 2019. B2W-Reviews01-An open product reviews corpus. In *the Proceedings of the XII Symposium in Information and Human Language Technology*. 200–208.

[24] Bernhard Rieder and Yarden Skop. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society* 8, 2 (2021).

[25] Joni Salminen, Sercan Sengün, Juan Corporan, Soon-gyo Jung, and Bernard J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE* 15, 2 (02 2020), 1–24. https://doi.org/10.1371/journal.pone.0228723

[26] Gustavo Santos, Vinicius F S Mota, Fabrício Benevenuto, and Thiago H Silva. 2020. Neutrality may matter: sentiment analysis in reviews of Airbnb, Booking, and Couchsurfing in Brazil and USA. *Social Network Analysis and Mining* 10, 1 (2020), 45. https://doi.org/10.1007/s13278-020-00656-5

[27] Saurabh Srivastava, Prerna Khurana, and Vartika Tewari. 2018. Identifying Aggression and Toxicity in Comments using Capsule Network. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 98–105. https://aclanthology.org/W18-4412

[28] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, Montréal, Canada, 19–26. https://aclanthology.org/W12-2103

[29] Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2, 1–7.