The Challenges of Creating a Parallel Multilingual Hate Speech Corpus: An Exploration

Katerina Korre, Arianna Muti, Alberto Barrón-Cedeño
DIT, Università di Bologna
Forlì, Italy
{aikaterini.korre2, arianna.muti2, a.barron}@unibo.it

Abstract

Hate speech is infamously one of the most demanding topics in Natural Language Processing, as its multifacetedness is accompanied by a handful of challenges, such as multilinguality and cross-linguality. Hate speech has a subjective aspect that intensifies when referring to different cultures and different languages. In this respect, we design a pipeline that will help us explore the possibility of the creation of a parallel multilingual hate speech dataset, using machine translation. In this paper, we evaluate how/whether this is feasible by assessing the quality of the translations, calculating the toxicity levels of original and target texts, and calculating correlations between the newly obtained scores. Finally, we perform a qualitative analysis to gain further semantic and grammatical insights. With this pipeline we aim at exploring ways of filtering hate speech texts in order to parallelize sentences in multiple languages, examining the challenges of the task.

Keywords: hate speech, parallel data, cross-linguality, single-trip translation

Warning: *This paper potentially contains hate speech.*

1. Introduction

Online hate speech is a popular research topic in Natural Language Processing (NLP), as it threatens the civility of the online communities targeting vulnerable groups and minorities (Flick, 2020; Nobata et al., 2016), rendering the task of hate speech detection of utmost importance. However, the multifacedness of hate speech poses a great challenge from both a linguistic and a computer science perspective. For instance, the constant evolution and the implicitness of language make it hard for current models to keep up (Yin and Zubiaga, 2021). Additionally, the presence of cultural nuances further complicates matters, hindering the development of resources that could be universally applicable across different contexts and languages.

Accommodating linguistic and cultural diversity is quite often neglected in NLP. While some multilingual and cross-lingual approaches exist (Lee et al., 2023b,a; Arora et al., 2023), there is a need for multicultural and cross-cultural approaches, as well (Hershcovich et al., 2022). NLP systems need to be sensitive to cultural factors to avoid biases while analyzing language data, acknowledging the differences among cultural norms, beliefs, and values that can greatly influence language use. Language- and culture-sensitive approaches should begin at the level of the annotation and corpus creation, as biases in supervised models are integrated from the very first step. However, employing annotators that are proficient in different languages or from diverse cultural backgrounds can be a difficult and expensive task.

In this paper, we focus on cross-linguality, while also touching upon cross-culturality. In an attempt to enrich the parallel resources that could be used for hate speech detection purposes, we design a pipeline that filters online hate speech instances that can be translated with a minimum effect on meaning and toxicity levels of the original text. Specifically, we extract examples from an already existing hate speech dataset (Davidson et al., 2017), and we automatically generate translations into Greek and Italian. We use back-translation (Ueffing et al., 2007) in order to perform a quality check of the translation (Moon et al., 2020). We apply a toxicity classifier (Devlin et al., 2019) in the original, translated, and backtranslated sentences to produce toxicity scores. We calculate correlations between the scores and examine which sentences can possibly be integrated into a parallel cross-lingual dataset. Finally, we perform a qualitative analysis on the translations to identify any patterns that could help in the optimisation of our pipeline.

Our contributions are the following:

- We design a pipeline which allows a semi-automatic filtering of parallel data that controls also the quality of the translation.
- We perform a qualitative analysis on the translated examples providing semantic and grammatical insights on the parallelisation of trans-

LREC-COLING 2024, pages 15842–15853

lated texts.

The rest of the paper is structured as follows. In Section 2, we present the related work with regard to approaches on multilinguality and crosslinguality. In Section 3 , we present our method, including the data and the model we used. In Section 4 we show our results, followed by a qualitative analysis in Section 5. Finally, we close with a discussion in Section 6 and our conclusions and future steps in Section 7.

2. Background

In this work, we differentiate between the definitions of two frequently interchangable terms: 'cross-lingual' and 'multilingual approaches'. A cross-lingual (or cross-language) approach refers to transferring knowledge or capabilities from one language to another. It is more about adapting NLP models trained in one language to work with other languages, often with limited data. Cross-lingual models typically start with a model trained in a specific language, and then they are adapted or fine-tuned for other languages using transfer learning techniques. This adaptation can be done using parallel data, machine translation, or other methods (Lample and Conneau, 2019; Spohr et al., 2011). The primary purpose of cross-lingual datasets is to facilitate tasks like machine translation or cross-lingual information retrieval by providing data in multiple languages (Lee et al., 2023a). Multilingual approaches, on the other hand, focus on a single, unified system that can work with many languages simultaneously (Sato et al., 2018). Multilingual models are trained on a diverse corpus of text from various languages. These models are designed to be *language-agnostic* and handle multiple languages without language-specific components (Yang et al., 2020; Spohr et al., 2011).

In this study, we focus on the parallelisation of translated data that can be used in both multilingual and cross-lingual approaches. As defined by Barrón-Cedeño et al. (2015), parallel texts are essentially precise translations or close approximations with only slight language-specific differences when compared to a comparable corpus, which should ideally consist of texts in multiple languages that are similar in both structure and content. Therefore, we use the term 'parallel multilingual data' to refer to our dataset [1].

[1]You can find the dataset at https://github.com/katkorre/Parallel-Multilingual-Hate-Speech-Corpus.git

2.1. Multilingual Approaches on Hate Speech Detection

A certain amount of research has been conducted on the front of multilingual approaches in hate speech detection. One of the most popular ways to explore hate speech detection techniques in multilingual settings is shared tasks, such as SemEval 2019 Task 5, which was organized into two subtasks: a main binary subtask in which models had to detect the presence of hate speech in English and Spanish, and a more fine-grained, explainable one, which focused on identifying further features in hate speech such as the aggressive attitude, the target harassed, and if the incitement is against an individual rather than a group (Basile et al., 2019).

Examining more languages simultaneously, Deshpande et al. (2022) compiled and released a multilingual dataset including 11 languages on which they evaluated deep learning models that are effective in a multilingual setting and generalize reasonably well to languages not present in the dataset, while also highlighting problems such as class imbalance (hate speech vs non-hate speech) and semantic biases within the different languages. Ousidhoum et al. (2019) presented a multilingual hate speech dataset of Arabic, English and French tweets and found that deep learning models perform better than BOW-based models in most of the multilabel classification tasks that they tested (i.e., directness of speech, hostility type of the tweet, target attribute, target group, annotator's sentiment).

There is also work conducted on the evaluation of multilingual models which takes into account language, bias, as well as data imbalance (Röttger et al., 2022; Ousidhoum et al., 2020). In addition, there are attempts to enrich datasets with more information such as author and annotator demographics (Hilte and et al., 2023; Huang et al., 2020)

2.2. Cross-lingual and cross-cultural Approaches

Recently, there has been an emergence of cross-lingual approaches in NLP hate speech detection. Given the limited resources in several languages, one-shot and few-shot are two of the most preferred approaches (Mozafari et al., 2022; Zia et al., 2022; Tita and Zubiaga, 2021; Stappen et al., 2020). Another preferred approach is knowledge transfer that can be enhanced with augmentation methods. For example, Pamungkas and Patti (2019) implemented a hybrid approach with deep learning and a multilingual lexicon to cross-domain and cross-lingual detection of abusive content. Bigoulaeva et al. (2022) used cross-lingual word embeddings to train neural network systems on a source language and apply it to a target language to make up for the lack of labeled examples. They also incorporate

Name		Language	Instances	Tokens	Chars
Hate Speech/Offensiveness (Davidson et al., 2017)		EN	1,000	18	83
Offensive Greek Tweet (Pitenis et al., 2020)	(FT)	EL	4,779	19	119
EVALITA (Sanguinetti et al., 2020)	(FT)	IT	6,837	23	148
Measuring Hate Speech (Sachdeva et al., 2022)	(FT)	EN	5,966	31	155

Table 1: Statistics for the data used in the translation experiments, as well as the data used for fine-tuning Mbert (FT). This table includes the language, the total number of used instances, as well as the average number of tokens and characters.

unlabeled target language data for further model improvements by bootstrapping labels using an ensemble of different model architectures. Arango et al. (2021) propose a hate specific data representation (i.e. hate speech word embeddings) and evaluate its effectiveness against general-purpose universal representations most of which, unlike their proposed model, have been trained on massive amounts of data. They focus on a cross-lingual setting, in which one needs to classify hate speech in one language without having access to any labeled data for that language. Bigoulaeva et al. (2021) used bilingual word embeddings-based classifiers and they achieve good performance on the target language by training only on the source dataset. Using their transferred system, they bootstrap on unlabeled target language data, improving the performance of standard cross-lingual transfer approaches. They use English as a high-resource language and German as the target language for which only a small amount of annotated corpora are available. Their results indicate that cross-lingual transfer learning together with their approach to leverage additional unlabeled data is an effective way of achieving good performance on low-resource target languages without the need for any target-language annotations.

When considering cross-cultural approaches, it is essential to acknowledge and account for the impact of cultural differences on annotation and translation. Lee et al. (2023a) delve into how individuals from different countries perceive hate speech, introducing CReHate, a cross-cultural re-annotation of the sampled SBIC dataset (Sap et al., 2020). This dataset includes annotations from five distinct countries: Australia, Singapore, South Africa, the United Kingdom, and the United States. Their statistical analysis highlights significant differences based on nationality, with only 59.4% of the samples achieving consensus among all countries. In a separate study, Lee et al. (2023b) attempt to quantify the cultural insensitivity of three monolingual (Arabic, English and Korean) hate speech classifiers by evaluating their performance on translated datasets from the other two languages, showing that hate speech classifiers evaluated on datasets from other cultures yield significantly lower F1 scores, up to almost 50%. Compared to their study, we focus on the initial machine translation step. Hershcovich et al. (2022) highlight the necessity of addressing the lack of cross-culturality in NLP and explore existing strategies to pave the way for a solution. They pinpoint three key areas for mitigating cross-cultural disparities: data collection, model training, and translation. They emphasize the importance of diverse annotation, understanding the trade-off between generalization and adaptation in model usage, and the limitations of reference-based evaluation methods, advocating for culture-sensitive human evaluation. Our approach is based on the three areas outlined by Hershcovich et al. (2022). However, we diverge in our method as we seek to automate certain aspects by minimizing reliance on human annotators through the use of translation.

3. Methodology

Our proposed methodology can serve as a means of identifying high-quality translation instances, which could potentially be incorporated into a parallel corpus, without requiring human experts. In this way, not only will the cost be reduced as we rely less on human evaluation, but also the process will be streamlined, allowing for the creation of additional parallel corpora not only for addressing hate speech detection, but also for other NLP tasks.

Data The data used for our translation experiments are instances extracted from the Davidson et al. (2017) dataset, which originally contained instances of both hate speech and offensiveness. For the purposes of our study, we extract only those that are labeled as hate speech, as we focus on this linguistic phenomenon, touching upon cultural implications. This is also due to the fact that the primary objective of this study is to examine the challenges of the parallelisation of hate speech, and not hate speech detection per se.

For fine-tuning our model, we use the Offensive Greek Tweet dataset (Pitenis et al., 2020) consisting of offensive and non-offensive text samples, and an Italian hate speech dataset, EVALITA (Sanguinetti et al., 2020), which consists of hate speech and non-hate speech instances. For English, we use a sample of the Measuring Hate Speech dataset (Sachdeva et al., 2022), which

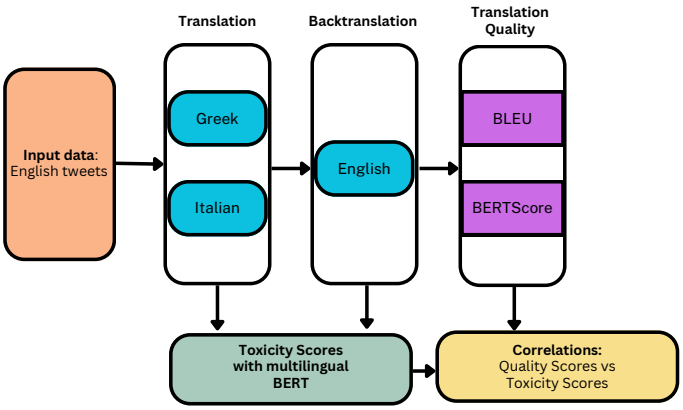

Figure 1: Visualisation of translation and evaluation pipeline.

also contains hate speech and non-hate speech instances. Table 1 shows basic statistics of the data.

Models Our experiments predominantly revolve around machine translation, therefore we use the out-of-the-box translation engine of ModernMT (Bertoldi et al., 2018) through their official API[2]. It is a context-aware, incremental and distributed general purpose Neural Machine Translation system based on the Fairseq Transformer model (Ott et al., 2019). To evaluate the translation quality, we employ two metrics; the first one is BLEU (Papineni et al., 2002), in order to see the similarity of n-grams and evaluate the translations on a structural level. The second is BERTscore (Zhang et al., 2020), which allows us to examine the translations on a semantic level. BERTscore measures the similarity between embeddings, thus we use it as a proxy to assess the quality of the translation, while also taking into account semantic and contextual information.

In terms of toxicity scoring, we use multilingual BERT (mBERT) (Devlin et al., 2019) by training three separate monolingual models on the English, Greek and Italian data, while also training one single multilingual model with all the data simultaneously. More specifically, the models are initialized with the 'bert-base-multilingual-cased' checkpoint and are further trained on the Offensive Greek Tweet dataset (Pitenis et al., 2020), on the EVALITA dataset (Sanguinetti et al., 2020), and the Measuring Hate Speech dataset (Sachdeva et al., 2022). During fine-tuning, we use the AdamW optimizer with a learning rate of 2e-5 and trained the model for three epochs. The input text is tokenized with a maximum sequence length of 128, which was the default.

Pipeline The research method revolves around producing translations with a minimum change in the original meaning and toxicity levels with respect to the source text. To achieve this while avoiding the need for human experts or manual translators, and to automate the process as much as possible, we opt for a round-trip translation, that is producing translations and backtranslations (Lee et al., 2023b; Beddiar et al., 2021). Figure 1 presents the pipeline.

First, we conduct a pilot experiment with a limited dataset of 100 instances, manually checking the quality of the translation, as well as generating quality and toxicity scores, as intended in our main experiments. Since we observe good translation quality from ModernMT, we proceed to the larger-scale experiment by translating and backtranslating the whole set of 1000 hate speech instances. Once obtained all the translations and the backtranslations, we apply BLEU and BERTScore on the backtranslations as the hypothesis and with the source text as the reference. We opt for using these two metrics because they cover two aspects of translation that we examined, the overlap of the translation on a structural level and the overlap on a semantic level.

Then, we apply our fine-tuned mBERT on the original English text, Greek and Italian translations, and the backtranslations to produce toxicity scores for all versions. We do not use more traditional metrics, such as F1 or accuracy to evaluate these models, as we do not look into whether a sentence is toxic or not (for which we would also need the gold standard labels) but at the actual degree of toxicity. In addition, accounting for the possibility that toxicity scoring might not capture deeper semantic nuances, we resort to a qualitative analysis, to see

[2] https://www.modernmt.com; translations were carried out between September and October 2023

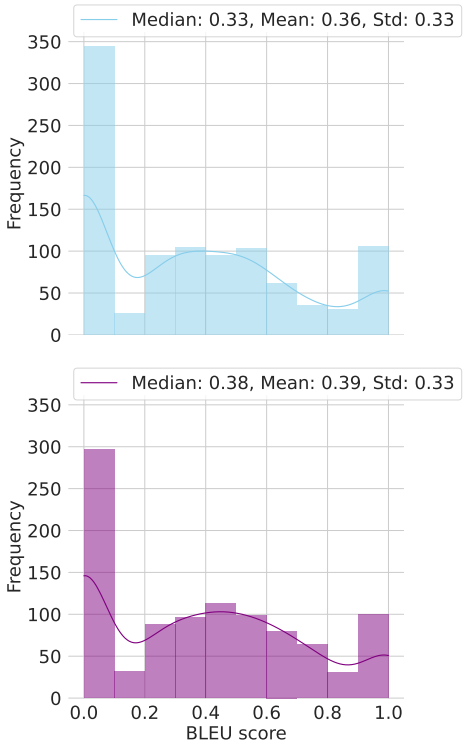

Figure 2: Distribution of BLEU scores for backtranslations from Greek (top) and from Italian (bottom).

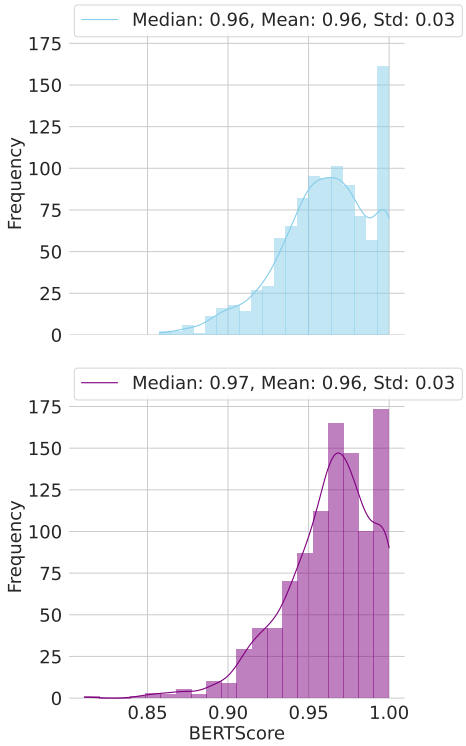

Figure 3: Distribution of BERTScores for backtranslations from Greek (top) and from Italian (bottom).

on which level we can compromise by using only computational methods. We also calculate correlations between translation scores and the toxicity scores as we wanted to look at the relationship between translation quality and the presence of toxic content. More specifically, we examine:

- The correlation between BLEU and BERTScore and Toxicity Score for Greek and Italian Translations.
- The correlations between BLEU and BERTScore and Toxicity Score for the backtranslated English data.

A positive correlation might indicate that the better the translations are, the lower toxicity could be, while a negative correlation might suggest the opposite. Correlating the scores of the translated data helps assess whether the quality of the data is related to the toxicity level in the data. Finally, we set a toxicity threshold and we extract all the sentences that could be used in a parallel multilingual dataset and perform a qualitative analysis.

4. Experimental Results

4.1. Translation Quality

In order to asses the translation quality, we compare the backtranslation into the source language with the original source texts. Figure 2 shows the distribution of BLEU scores on instances backtranslated both from Greek and Italian. The results can be described as moderately good. A substantial portion of the translations reached scores ranging from 0.6 to 1.0, which indicates a level of translation quality that spans from good to perfect. The backtranslations from Italian slightly outperform those from Greek, with more instances surpassing the threshold of 0.6.

Figure 3 shows the results in terms of BERTScore. Both Greek and Italian scores were quite high, with the values falling within the range of approximately 0.8 to 1.0, which indicates an adequate match between candidate translation and reference.

This indicates that some translations may not be very accurate when evaluated with the BLEU metric, suggesting a limited n-gram overlap. In contrast, the BERTScore values for both Greek and Italian translations look more promising. Considering that in this work we value semantic similarity more than the structural sentence matching, the translations appear to perform well in terms of capturing semantic content.

4.2. Toxicity Evaluation

The primary objective of toxicity evaluation is to assess whether the toxicity is retained throughout the entire pipeline procedure. The averaged toxicity

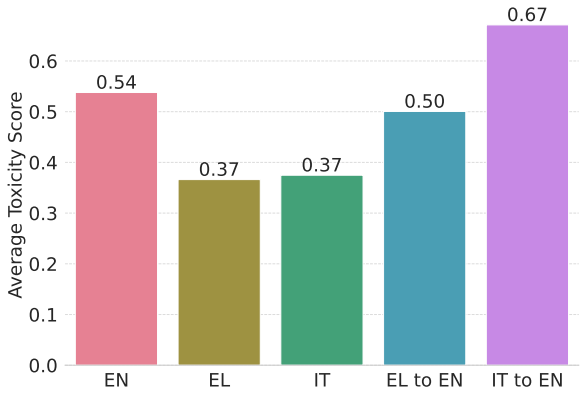

Figure 4: Average toxicity scores from monolignual models.

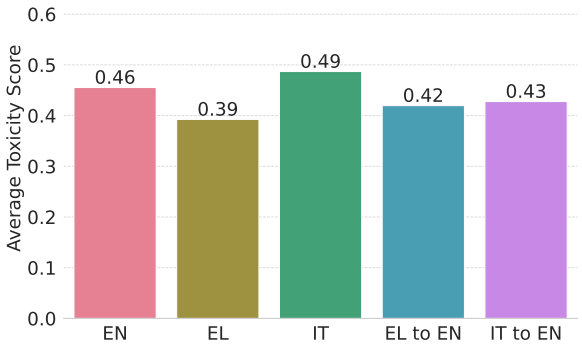

Figure 5: Average toxicity scores from multilingual models.

scores are presented in Figures 4 for individually trained monolingual and 5 for multilingual models.

When scoring with the mBERT models that are trained separately in the three languages, we observe some discrepancies. The English texts, including both the source and the backtranslations, have the higher averaged toxicity score, ranging from 0.54 to 0.67, while the score is lower for the Greek and Italian translations to 0.37. However, we would have expected more pronounced discrepancies in the toxicity levels of the English texts, indicating that the reason might be that backtranslations restore the toxicity of the text. In fact, the toxicity when backtranslating from Italian to English exceeds the toxicity of the initial input texts. This could be due to semantic shift that might occur during translation (Beinborn and Choenni, 2020), which leads to change or even amplification of certain meanings of the text. Similarly, the multilingual model scores the toxicity of the Italian translation as more toxic compared to the original English text.

Overall, when examining the toxicity with the multilingual model, there is less fluctuation among translations and backtranslations compared to the monolingual models, with the values ranging from 0.39 to 0.49. Still, the toxicity scores of the backtranslations are lower than the toxicity scores of the original text, hinting that indeed some toxicity is lost during the translation process. This can be considered as a more fair comparison as a monolingual model is used for producing the toxicity scores. Yet, we must take into account that potential errors in machine translation could influence the toxicity score in subsequent phases of the pipeline. Hence, we also conduct a qualitative assessment (see Section 5) to address this concern.

4.3. Correlations

To further explore the relation between the quality scores and the toxicity scores, we conduct a correlation analysis, using Pearson's correlation coefficient (Freedman et al., 2007). The results are summarised in Table 2. The correlations observed are generally weak with coefficients close to 0. These values suggest that as both the quality scores (BLEU and BERTScore) increase, i.e., the translations are of higher quality, the toxicity score tends to slightly decrease, indicating lower toxicity levels. Therefore, taking into account that all the correlations are weak, it is difficult to draw definitive conclusions about a linear relationship between these variables, as the results indicate that the relationships might not be very strong or reliable. The values are close to zero, indicating that the change in BLEU or BERTScore does not strongly predict the change in the toxicity score. On the other hand, for the correlations with the backtranslation, we notice a slightly positive correlation both with BERTScore and BLEU score, although still close to zero, which is expected as the toxicity levels remained almost intact. The same holds true when we evaluate with the multilingual model, with only the Italian multilingual translation BERTScore suggesting a weak negative correlation between the MT quality and the toxicity score.

In practical terms, this analysis suggests that while there is some tendency for better translations to be associated with lower toxicity levels, other factors might also influence toxicity, including which models are used to assess both the translation quality and the toxicity scores.

5. Quality Evaluation

5.1. Threshold Filtering

To explore the possibility of filtering the sentences and keeping those that have maintained their toxicity, as well as assessing whether they are adequately translated, we define several thresholds that allow us to filter a number of sentences and manually evaluate them in terms of meaning and

Translation	BERTvsTox	BLEUvsTox
Greek_mono_t	-0.09	-0.07
Italian_mono_t	-0.07	0.03
Greek_mono_bt	0.07	0.04
Italian_mono_bt	0.03	0.00
Greek_multi_t	0.00	0.03
Italian_multi_t	-0.15	-0.02
Greek_multi_bt	0.03	0.07
Italian_multi _bt	-0.05	0.05

Table 2: Correlations between BERT Score/BLEU Score and toxicity levels from monolingual (tag *_mono*) and multilingual (tag *_multi*) models. Tag *_t* refers to translation while tag *_bt* refers to backtranslation.

toxicity, using the scores produced from the multilingual model. More specifically, the steps that we follow for the filtering and the manual evaluation are the following:

1. Calculate the absolute difference between the source sentence's toxicity score and the score of the translated sentence into Greek/Italian.
2. Calculate the absolute difference between the source sentence's toxicity score and the score of the backtranslated sentence from Greek/Italian.
3. If both calculated absolute differences are less than or equal to the specified threshold, we add the sentence to the list of maintained sentences.

Using a threshold of **0.1**, we identify 141 instances that maintained their toxicity scores. When the threshold is increased to **0.2**, we find 333 sentences that meet the criteria for maintained toxicity scores. Further increasing the threshold to **0.3** results in the identification of 611 instances that continue to meet the criteria for maintained toxicity scores, and is in fact more than half of the initial dataset. We decide to keep the 0.2 threshold, as setting the threshold too low might result in capturing instances with maintained toxicity but poor translation quality. A threshold of 0.2 allows a compromise between quality and toxicity maintenance. We then proceed to qualitatively evaluate a small sample of these instances.

5.2. Manual Evaluation

From the 333 filtered instances, 60 instances are randomly selected for further evaluation by two language experts, who are also authors of this paper. One is proficient in both English and Greek, holding a linguistics degree in both languages, and possesses Greek as their mother tongue. Likewise, the language expert who evaluated the translations from English to Italian holds a degree in Languages, with Italian as their mother tongue. In this section, we discuss different linguistic and cultural features of some of the borderline cases, as well as some of the successfully translated parallel sentences.

One of our initial observations in both languages is the overall semantic quality of the translations, as corroborated by the results of our quality control analysis in Section 5. On the contrary, grammatical and syntactical inconsistencies create more certain complexities in achieving the parallelisation of the sentences. We present and analyse some noteworthy examples in Table 3.

Example 1 is one of the cases where translation was successful conveying the intended meaning but generated grammatical issues. In particular, in the Greek translation there is an incorrect use of the person in the main verb ('kill'), which disagrees with the subject 'ugly ass monkey'. In Italian, on the other hand, the problem arises with the phrase 'looking shit', which refers to the appearance which is erroneously translated as 'cercando', meaning 'to look for'. *Example 2*, in contrast, is one of the successful cases of the translation where both the meaning and the grammatical structure of the sentence are preserved in both Greek and Italian. In Greek, the word 'beaner' which is a racist slur towards Mexican people, is translated as φασολιάριδες which is a Greek adaptation that sounds natural. In Italian, however, the word beaner remains unchanged which is acceptable, as English words sometimes are integrated intact into other languages, especially in modern usage.

Example 3 has also managed to capture the intended meaning of the original instance in both target languages, yet there is a shared error in terms of the person. This error might possibly be due to 'ill' which could be texting language for 'i'll', and which the models falsely substituted with the third person in both languages. *Example 4* is one of those examples that the words were translated accurately yet the meaning changed because of the literal aspect of the translations. More specifically, in the source text, the author of the tweet uses a series of objectifying insults to make up a misogynistic comment. These sort of insults are not common in neither Greek nor Italian, and therefore the translations do not sound natural. Finally, *Example 5* is similar to Example 1 where the translation was successful for both Greek and Italian. The specific example is a racial slur, mainly towards Asian people, therefore, in both target languages, the translation referred to the color 'yellow' (κιτρινιάρη, muso giallo) which is usually associated with East Asian people, thus capturing the toxicity of the original text.

	Source English	Translation into Greek	Translation into Italian	BERTScore Greek	BERTScore Italian
1.	kill yourself ugly ass monkey looking shit	αυτοκτονήστε άσχημη μαϊμού που δείχνει σκατά	uccidi te stesso brutto culo scimmia cercando merda	0.96	0.97
2.	all us beaners were pissed like robert when the jew ref stopped it	όλοι εμείς οι φασολιάρηδες τσατιστήκαμε σαν τον Ρόμπερτ όταν ο Εβραίος διαιτητής το σταμάτησε	tutti noi beaners eravamo incazzati come Robert quando l'arbitro ebreo lo fermò	0.96	0.96
3.	shut the fuck up fag ill beat the shit outta you	βγάλε τον σκασμό αδερφάρα να σε πλακώσει στο ξύλο	chiudi quella cazzo di bocca da frocio ti ha pestato a sangue	0.93	0.96
4.	shut your dirty dirt trap you limp knees, ashy footed, blue balls, isolated hamster wheel, no nipples bitch	κλείστε τη βρώμικη παγίδα σας, χαλαρά γόνατα, στάχτη, μπλε μπάλες, απομονωμένη ρόδα χάμστερ, χωρίς θηλές σκύλα	chiudi la tua sporca trappola per sporcizia, ginocchia zoppicanti, piedi cenere, palle blu, ruota isolata per criceti, niente capezzoli, stronza	0.94	0.94
5.	Fucking gook	Γαμημένε κιτρινιάρη	Fottuto muso giallo	1.00	0.92

Table 3: Noteworthy translation cases for the qualitative analysis.

6. Discussion

The findings of this study, as well as previous research endeavors (Lee et al., 2023b,a) show that the creation of a parallel hate speech corpus is feasible, however there must be a degree of compromise to overcome several obstacles.

Challenge 1: The quality of the translation is not certified. In Section 4.1, we observe that the performance of the machine translation system is adequate. However, although the semantic evaluation yielded high scores, there were serious grammatical and syntactical issues with the sentences, leading to lower BLEU scores. Creating a comprehensive parallel corpus demands both semantic and grammatical/structural correctness. As explained by Hershcovich et al. (2022), linguistic form and style are associated with social and cultural factors, and any linguistic variations must be correctly represented in datasets. Therefore, both choosing the right translation method and the right metrics for the evaluation are paramount to ensure quality. Given this, our method allows us to filter out higher quality sentences.

Challenge 2: Toxicity fluctuates from source to target language. Only preserving the intended meaning is not sufficient when creating a parallel hate speech dataset; the levels of the toxicity of the original text must also be maintained. This is not an easy task because, as we saw in Sections 4.2 and 4.3, the toxicity fluctuates when translating to another language while also the toxicity scores do not necessarily correlate with the quality of the translation. Therefore, an individual study should be conducted on the toxicity levels in order to analyze toxicity fluctuations to ensure that, beyond meaning and sentence structure, toxicity remains on similar levels of that of the original text.

Challenge 3: There are cultural nuances that are hard to be translated. As mentioned in Section 1, a major issue in NLP and hate speech detection revolves around the lack of culture awareness. Our qualitative analysis was an attempt to shed some light on the cultural dimensions of hate speech. We showed that there are instances that were effectively translated, yet there might be missing cultural context. In our case, Example 2, which employs the derogatory term 'beaner' in reference to Mexican people, is adequately translated, yet it resonates primarily with a specific geographical population, namely individuals from the US. People from other countries and cultures may struggle to comprehend why this specific example constitutes hate speech. This is an example of the task of *adaptation* in translation, as described in Peskov et al. (2021). The authors assert that while computational techniques for this task have advanced, there is still room for improvement. They still advocate for the automatizing of the task and they recommend using available datasets in high resource languages by adapting content instead of just translating it literally. Our approach paves the way for this endeavor.

Taking all these challenges into account, it is clear why creating a parallel hate speech corpus

is not an easy task. In our study, the amount of sentences that were filtered and could be compiled in a parallel dataset were limited but they existed, rendering the task hard but still feasible. Especially, since machine translation is becoming more and more effective, it should be "used for bridging between cultures, investigating cross-cultural communication" (Hershcovich et al., 2022).

7. Conclusion

In this paper, we touch upon some challenges about creating a parallel hate speech corpus. Specifically, we use machine translation to translate English tweets into other languages —Greek and Italian— and then we use backtranslation in conjunction with translation metrics in order to evaluate the quality of the translation. Additionally, we conduct some evaluation tests on toxicity levels of the translations. Finally, we perform a qualitative analysis on a sample of our used instances. Our results show that machine translation can achieve adequate results with regard to the intended meaning of the sentence but will still produce grammatical and syntactical errors that cannot be inserted into a parallel corpus. Only few examples maintained their meaning and toxicity of the original text and did not have grammatical or syntactical issues, which underscores the difficulty of the task.

In future research, we plan to expand the dataset to potentially produce a comprehensive parallel hate speech corpus. We plan to conduct additional experiments involving a wider array of languages, including employing English as a target language to investigate whether toxicity levels experience a similar decrease as observed in this study. Moreover, we intend to explore whether expert translators can discern cross-cultural differences and whether the task poses similar challenges to those encountered with machine translation.

Limitations

The limitations of our work primarily pertain to practical issues. The dataset used for our translation experiments is relatively limited, consisting of only 1000 instances, which, in turn, restrict the number of sentences suitable for the inclusion in a parallel hate speech corpus. Furthermore, the nature of the instances, which is Tweets, is usually characterized by specific features such as abbreviations and text-message-style language. Finally, each challenge we describe needs a more comprehensive analysis, involving a broader selection of models and metrics to determine the most suitable approach for addressing each specific challenge.

Ethical Consideration

All the data that we used are publicly available and the relevant information can be found in their corresponding references and repositories. In the examples of the qualitative analysis, we have omitted any information that could identify the author of each comment, protecting their anonymity.

Acknowledgments

K. Korre's research is carried out under the project "RACHS: Rilevazione e Analisi Computazionale dell'Hate Speech in rete", in the framework of the PON programme FSE REACT-EU, Ref. DOT1303118. A. Muti's research is carried out under project "DL4AMI–Deep Learning models for Automatic Misogyny Identification", in the framework of Progetti di formazione per la ricerca: Big Data per una regione europea più ecologica, 'digitale e resiliente—Alma Mater Studiorum–Universita di 'Bologna, Ref. 2021-15854.

8. Bibliographical References

References

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2021. Cross-lingual hate speech detection based on multilingual domain-specific word embeddings. *CoRR*, abs/2104.14728.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A factory of comparable corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13, Beijing, China. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic Drift in Multilingual Representations. *Computational Linguistics*, 46(3):571–603.

Nicola Bertoldi, Davide Caroselli, and Marcello Federico. 2018. The ModernMT project. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 365, Alicante, Spain.

Irina Bigoulaeva, Viktor Hangya, and Alexander Fraser. 2021. Cross-lingual transfer learning for hate speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 15–25, Kyiv. Association for Computational Linguistics.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2022. Addressing the challenges of cross-lingual hate speech detection. *CoRR*, abs/2201.05922.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Neha Deshpande, Nicholas Farris, and Vidhur Kumar. 2022. Highly generalizable models for multilingual hate speech detection.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Caterina Flick. 2020. The legal framework on hate speech and the internet good practices to prevent and counter the spread of illegal hate speech online. In *Language, Gender and Hate Speech A Multidisciplinary Approach*. Fondazione Università Ca' Foscari.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Lisa Hilte and et al. 2023. Who are the haters? a corpus-based demographic analysis of authors of hate speech. *Frontiers in Artificial Intelligence*, 6:986890.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023a. Crehate: Cross-cultural re-annotation of english hate speech dataset.

Nayeon Lee, Chani Jung, and Alice Oh. 2023b. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nedjma Ousidhoum, Yangqiu Song, and Dit-Yan Yeung. 2020. Comparative evaluation of label-agnostic selection bias in multilingual hate speech datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2532–2542, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Maurizio Sanguinetti, Giacomo Comandini, Elia di Nuovo, Simone Frenda, Maria Stranisci, Cristina Bosco, and Irene Russo. 2020. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task. In *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian Final Workshop*, Torino. Accademia University Press.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Motoki Sato, Hiroki Ouchi, and Yuta Tsuboi. 2018. Addressee and response selection for multilingual conversation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3631–3644, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dennis Spohr, Laura Hollink, and Philipp Cimiano. 2011. A machine learning approach to multilingual and cross-lingual ontology matching. In *The Semantic Web – ISWC 2011*, pages 665–680, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL. *arXiv e-prints*, page arXiv:2004.13850.

Teodor Tita and Arkaitz Zubiaga. 2021. Cross-lingual hate speech detection using transformer models. *CoRR*, abs/2111.00981.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32, Prague, Czech

Republic. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1435–1439.

9. Language Resource References

Thomas Davidson and Dana Warmsley and Michael Macy and Ingmar Weber. 2017. *Automated Hate Speech Detection and the Problem of Offensive Language*.

Pitenis, Zesis and Zampieri, Marcos and Ranasinghe, Tharindu. 2020. *Offensive Greek Tweet Dataset*.

Sachdeva, Pratik and Barreto, Renata and Bacon, Geoff and Sahn, Alexander and von Vacano, Claudia and Kennedy, Chris. 2022. *The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism*.

Sanguinetti, Maurizio and Comandini, Giacomo and di Nuovo, Elia and Frenda, Simone and Stranisci, Maria and Bosco, Cristina and Russo, Irene. 2020. *HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task*.