Investigating the Effect of Machine-Translation on Automated Classification of Toxic Comments

James Roy[1], Siddhi Suresh[2], Mohamed Elsayed[3], Ronie Rocca[3],
Ziqian Dong[3], Huanying Gu[3], and N. Sertac Artan[3]
[1]University of Cambridge, Department of Engineering, Cambridge, UK
[2]Southern Connecticut State University, Department of Computer Science, New Haven, CT, USA
[3]New York Institute of Technology, College of Engineering and Computing Sciences, New York, NY, USA
E-mails: jr818@cam.ac.uk, anantakriss1@southernct.edu, {melsay04, rrocca, ziqian.dong, hgu03, nartan}@nyit.edu

***Abstract*—This paper discusses the research findings on the performance of automated toxic comment classification following machine translation. We tested Google Perspective API first on comments from non-English Wikipedia talk pages in five languages, and then on their English translation (generated with Google's Cloud Translate API). In addition to giving baselines on the current performance of Perspective in five languages, this allows for comparison on how machine-translation alters the classification. We show that the level of disagreement between pre- and post-translation classification is heavily dependent on the language used. The comments come from a Kaggle dataset and we filter them to ensure monolingual comments with simple punctuation. Results show above 84% of the French, Italian and Spanish comments received the same class pre- and post-translation, while Portuguese and Russian performed the worst of the five languages tested, with F-scores below 0.6.**

***Index Terms*—Toxic Comment Detection, Machine Translation, Perspective API**

I. Introduction

Identification of toxic comments is important to maintain safe and productive space online for discussion and debate. Tools which automate this classification such as Perspective API [1] offer a solution to this issue, but do not consider the final language the user is viewing content in. From the reader's perspective, comments which become toxic following machine translation (regardless of the author's intention) could discourage participation in productive online debate. Conversely, the author of a comment should not be held liable for the distress and repercussions caused by *toxification* of their text as a result of translation. There would be great injustice in the scenario where a user is boycotted due to this *unintentional* toxification. This paper seeks to answer the question: does machine-translation alter the toxicity of comments and cause misclassification? We investigated this issue by classifying comments with Perspective API, which is a joint effort from Jigsaw and Google to automate online toxic comment classification. Perspective is widely used, with partners including Reddit and The New York Times [2]. As well as its frequent use, Perspective API is available in 17 languages which makes it suitable for this investigation.

Social media platforms have identified translation as a key feature, given users are likely to view content written in a language they do not understand while browsing [3]. This combination of automated classification of machine translated content for detecting toxic comments introduces ambiguity over what should be classified: the original comment or its translation? By comparing Perspective API's classification of comments prior to translation and post-translation, we show that this decision is heavily language dependent. Somewhat worryingly, we also identify a significant proportion of comments which become toxic following translation. Should social media platforms only consider automated classification of comments in their original language, users who view the translation will be exposed to online toxicity.

To the best of our knowledge, this is the first study to investigate Perspective API's use on translated content, and aims to inform platforms of the dangers of unsupervised translation services. Previous studies [4] [5] have considered how machine translation impacts sentiment analysis, but none have taken advantage of Perspective API. With regard to the type of sentiment analysis, only one paper [6] was found to specifically consider toxic sentiment for translated comments. Beyond preserving the safety of social media platforms, a secondary motivation comes from assessing if a multilingual toxic classifier can be built from a monolingual classifier combined with a machine-translation service.

The rest of this paper is organized as follows. Section II reviews the related work in sentiment analysis and machine translation. The process of filtering the datasets for unusable comments prior to classification, and the methods used, are discussed in Section III. Section IV presents and analyses the results of the classification. Finally, conclusions are discussed in Section V.

II. Related Work

The focus of this paper falls under sentiment analysis following machine-translation. Balahur et al. claimed that statistical machine-translation (SMT) services (including Google

This project is funded by National Science Foundation Grant No. CNS-1852316 and New York Institute of Technology.

Translate) are advanced enough to convey the sentiment of the original text [5]. This is evaluated through the ability to train a polarity classifier on translated text. While the results obtained by the paper show it is possible for a boutique classifier to be built, it does not test current off-the-shelf classifiers. Araújo et al. and Sagnika et al. analyzed various English sentiment analysis methods and non-English methods on machine translated sentences [7] [8]. They evaluated the performance of such methods using several languages. Their results demonstrated that English sentiment analysis methods performed better than non-English sentiment analysis methods, showcasing that there is much room for improvement in the development of non-English sentiment analysis methods.

Makhnytkina et al. considered translating comments from Russian to English and then training a model to classify the English version of the comments for toxicity [6]. Their findings suggest that translation did preserve enough of the toxic sentiment, such that classification in English correlated with classification in Russian. Mohammad et al. assessed automated detection of the polarity (positive/negative/neutral) of Arabic comments and their English translation [9]. It was concluded that English classifiers (ones which first translate the Arabic comments) perform equally to translators in Arabic. Wang et al. translated a multilingual (English-Chinese) dataset to a monolingual (Chinese) dataset, while preserving semantic information by considering English and Chinese words with similar polarity [10]. A successful multilingual emotion detection model is then trained on this monolingual dataset.

Given Perspective API's popularity, previous work has investigated its limitations in other areas, as well as assessing its performance on non-translated comments [11] [12]. Specifically, Jain et al. assessed its susceptibility to adversarial attacks generated by using common techniques to perturb toxic comments [13]. Similarly, Brown et al. focused on adversarial attacks that preserve semantic meaning through acoustic and visual similarity ("hate" to "h@te" to "hayte") [14]. While this paper is not focusing on adversarial attacks, it is conceivable that machine-translation could be used to generate adversarial attacks insofar as we show benign comments can become toxic after being translated into another language.

III. Methodology

A. Datasets

In this study, we use the "test" dataset from Kaggle's Jigsaw Multilingual Toxic Comment Classification competition [15]. This dataset is taken from various non-English Wikipedia talk-pages. The Kaggle dataset used binary labels (toxic and non-toxic), which is consistent with the labeling used by Perspective API. From this multilingual dataset, five monolingual datasets were extracted using the language tags provided, namely: French, Italian, Portuguese, Russian, and Spanish.

TABLE I: Number of toxic and non-toxic comments before and after filtering

	Initial Kaggle Dataset		Dataset After Filtering	
Language	Toxic /Non-Toxic	Total	Toxic /Non-Toxic	Total
French	7,580/3,340	10920	2,303/1,380	3,683
Italian	6,857/1,637	8494	1,312/492	1,804
Porteguese	9,264/1,748	11012	2,821/801	3,622
Russian	8,312/2,636	10948	2,588/1,209	3,797
Spanish	5,080/3,358	8438	1,868/1,515	3,383

B. Translation and Toxic Comment Classification Pipeline

In this work, we aim to capture the effect of machine translation on toxicity preservation in text. Thus, we limit our investigation to monolingual comments. Additionally, comments containing features which may corrupt translations and throw errors in classification are removed. These features include: the lack of words (i.e. purely punctuation), incorrect use of punctuation, text-based emoticons, and comments with text from multiple languages. In future work we aim to extend this pipeline to incorporate features common to social media comments (such as usernames preceded with '@', hashtags, and URLs).

In this section, we outline our machine translation and toxic comment classification pipeline using Google Translate and Google Perspective API. Each dataset is checked to remove comments which contain English words (unless those words are present in both languages, such as "no" which is present in both Italian and English). This is done to avoid dropping comments if the language naturally shares words with English. The pipeline will output a filtered, translated dataset consisting of English comments.

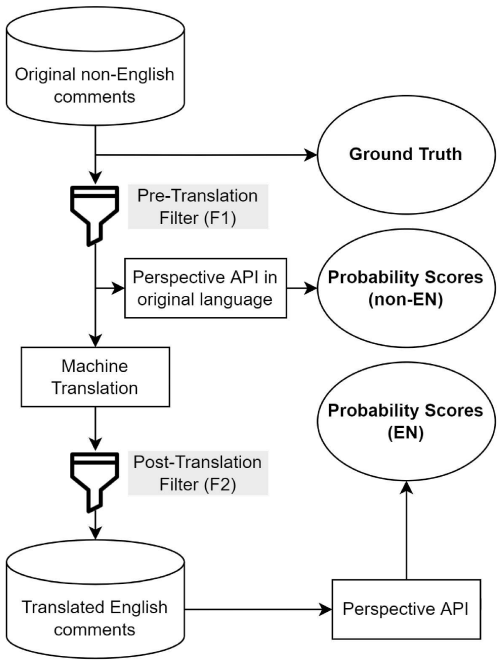

Fig. 1: Overview of the Machine Translation and Toxic Comment Classification Pipeline.

Figure 1 shows an overview of the pipeline. The pipeline results in three lists of toxicity labels, created when the dataset passes through different stages. First, the original labels from the dataset are extracted and stored as the *ground truth*. The *Pre-Translation Filter (F1)* filters non-English comments to only allow monolingual sentences to proceed through the pipeline. The filtered datasets are then classified by Perspective API in their respective languages, resulting in a list of probability scores (shown as *Probability Scores (non-EN)* in the figure). The filtered datasets are then machine-translated via Google's Cloud Translation API [16]. After translation, a second filter (*Post-Translation Filter (F2)*) ensures only English comments will be classified, in case foreign words and characters are kept after translation. Perspective API classification results after F2 are given as *Probability Scores (EN)* in the figure. The two probability scores (non-EN, and EN) are compared for final evaluation (Section IV).

1) Pre-Translation Filtering Model: This portion of the filtering model is referred to as "F1."

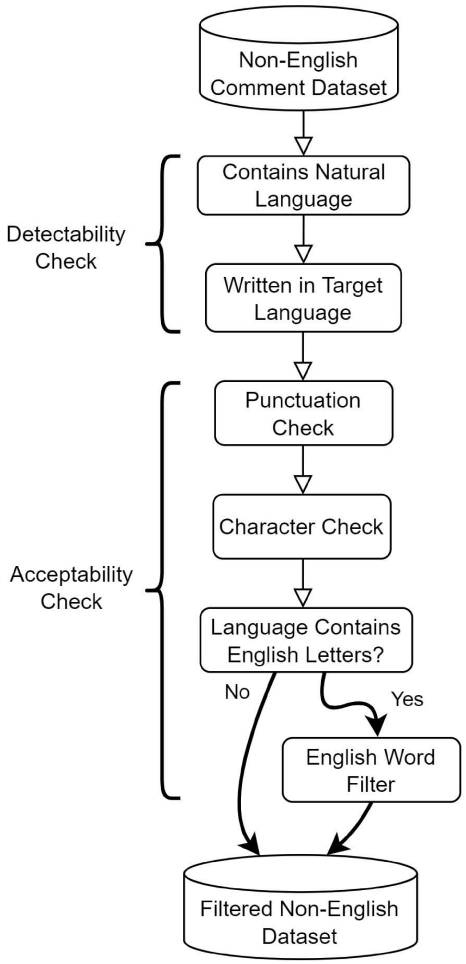

Fig. 2: Pre-Translation Filtering Model (F1).

The non-English dataset is first filtered to remove comments composed only of punctuation. Each comment is then checked to be written in the specified language of the dataset. These two steps are referred to as "Detectability" in Figure 2. A more sophisticated check is then performed to ensure correct use of punctuation, and to filter comments containing text-based emoticons (e.g. ":)"). The range of acceptable punctuation is expanded for certain languages, for example "¿" and "¡" in Spanish.

The next step in filtering depends on the writing system of the dataset's language (for example Latin script). This is due to the difference in characters in non-Latin script languages.

Comments are removed if they contain characters not present in their language's writing systems. For example, the Italian Alphabet officially does not contain the letters "J", "K", "W", or "X". In the case that the language contains letters present in the English Language, the comments are passed through an English Word Filter. For example, Russian utilizes the Cyrillic writing system, however, letters such as "A" are present in both the English and Russian alphabet. For these languages, the difference between the English word list [17] and the foreign word list are extracted and used as a list of red flag words to detect English in a comment. The list of words in French[1], Italian[2], Portuguese[3], and Spanish[4] were collected from GitHub.

The filtering model ensures that only monolingual sentences are used. Detectability quickly determines if a comment should be further analyzed in the "Acceptability Check". The language filter is an alternative to Language Detection API's. These APIs would only work effectively to determine the most likely language of an entire sentence and not to pick out specific words, or would produce misleading results since words used in multiple languages would be assigned to only one language, commonly English. Our method is also easily modifiable and may be used to include more colloquial terms used in texting and social media. English is set as a basis for comparison, however, the option of creating lists of red flags with another language set as a basis is possible.

TABLE II: Filtering Results before Translation

Kaggle Dataset	F1 Output	
Language	Acceptable	Removed
French	4,203 (38%)	6,717 (62%)
Italian	1,883 (22%)	6,611 (78%)
Portuguese	4,221 (38%)	6,791 (62%)
Russian	4,054 (37%)	6,894 (63%)
Spanish	3,774 (45%)	4,664 (55%)

2) Translation: For translation, Python's implementation of Google's Cloud Translation API (the google-cloud-translate package) was used. The filtered dataset after F1 stage as shown in Figure 2, is split into batches and translated. Translations are done from the target language to English. Google Translate was chosen because it is widely used in social media translation, including Twitter.

[1]https://github.com/words/an-array-of-french-words
[2]https://github.com/napolux/paroleitaliane
[3]https://github.com/jfoclpf/words-pt
[4]https://github.com/words/an-array-of-spanish-words

TABLE III: Original Labels ($\mathbf{y}$) vs. Non-English Perspective ($\hat{\mathbf{y}}$)

		$\hat{y}$									
		French		Italian		Russian		Portuguese		Spanish	
		Non-Toxic	Toxic	Non-Toxic	Toxic	Non-Toxic	Toxic	Non-Toxic	Toxic	Non-Toxic	Toxic
y	Non-Toxic	2236 (60.74%)	67 (1.82%)	1261 (69.90%)	51 (2.83%)	2546 (67.11%)	40 (1.05%)	1305 (36.03%)	1516 (41.86%)	1770 (52.87%)	71 (2.12%)
	Toxic	**714 (19.40%)**	664 (18.04%)	**224 (12.42%)**	268 (14.86%)	**751 (19.79%)**	457 (12.05%)	12 (0.33%)	789 (21.78%)	**657 (19.67%)**	850 (25.39%)

TABLE IV: Non-English Perspective ($\hat{\mathbf{y}}$) vs. English Perspective ($\hat{\mathbf{x}}$)

		$\hat{x}$									
		French		Italian		Russian		Portuguese		Spanish	
		Non-Toxic	Toxic	Non-Toxic	Toxic	Non-Toxic	Toxic	Non-Toxic	Toxic	Non-Toxic	Toxic
$\hat{y}$	Non-Toxic	2470 (67.12%)	**480 (13.04%)**	1281 (71.05%)	**204 (11.31%)**	2660 (70.15%)	**636 (16.77%)**	1307 (36.10%)	9 (0.25%)	2036 (52.87%)	**391 (11.68%)**
	Toxic	107 (2.91%)	623 (16.93%)	34 (1.89%)	284 (15.75%)	65 (1.71%)	431 (11.37%)	1390 (38.40%)	914 (25.25%)	95 (19.67%)	825 (25.39%)

3) Post-Translation Filter (F2): The translated dataset is then passed through the Post-Translation Filter (F2) as shown in Figure 3.

This filtering stage checks for non-English letters and characters to ensure that comments that failed to be translated are not included in the final dataset. The dataset is then checked for correct punctuation, as described in the F1 filtering stage to ensure that the translator does not add punctuation and other symbols.

Tables II and V show a comprehensive view of how many comments were removed and accepted after each phase of filtering, while Table I shows an overview of the class distribution of the filtered and translated datasets. As expected, most comments were removed by the pre-translation filter as opposed to the post-translation filter. Most comments were removed due to having unacceptable characters/symbols that did not fit our punctuation criteria. After filtering and translating, 21-40% of the original comments remained.

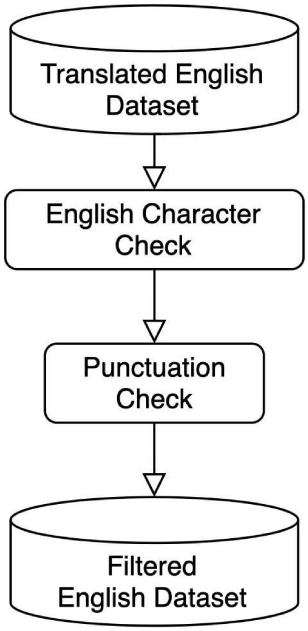

Fig. 3: Post-Translation Filtering Model (F2)

C. Classification

Perspective API was used to classify labeled non-English datasets both before and after machine translation. This allowed for testing of how well the translation API preserved toxicity of comments, as well as comparing the non-English Perspective API with the English Perspective API. Perspective API was setup utilizing its Python implementation (from the google-api-python-client package), and the requested attribute set to "TOXICITY."

When classifying the datasets outlined in Section III-A with Perspective API, 0.3% of comments from the Kaggle dataset resulted in errors. Table VI gives a breakdown of Perspective API errors by language. For example, in the French dataset, which includes the translated comments, only 3 comments were flagged as an error(0.08%). The Spanish dataset had the highest number of errors (1.12%). The most common of these errors was "Languages not supported," which is typically caused by unrecognizable characters or when comments in languages that use non-Latin characters have been translated to contain Latin characters [1]. For instance, a comment from the Russian dataset when translated is converted to a hybrid language known as Runglish. This results in English Perspective API to throw an error.

TABLE V: Filtering Results after Translation

Kaggle Dataset	F2 Output	
Language	Acceptable	Removed
French	3,683 (87%)	520 (13%)
Italian	1,804 (96%)	79 (4%)
Porteguese	3,622 (86%)	599 (14%)
Russian	3,797 (94%)	257 (6%)
Spanish	3,383 (90%)	391 (10%)

TABLE VI: Total number of classification errors after translation.

Language	Number of Comments
French	3 (0.08%)
Spanish	38 (1.12%)
Italian	1 (0.06%)
Portuguese	2 (0.06%)
Russian	6 (0.16%)

IV. Results and Discussion

The confusion matrices presented in Table III compare the number of comments classified as toxic and non-toxic by non-English Perspective API to the ground truth (original labels). The comments with classification errors (Table VI) are excluded from this comparison.

The confusion matrices presented in Table IV compare the number of comments classified as toxic and non-toxic by English and non-English Perspective API. Similarly, the comments with classification errors (Table VI) are excluded from this comparison.

Standard classification task metrics are reported in Tables VII and VIII.

TABLE VII: Classification metrics for non-English Perspective API on the original dataset

	Accuracy	Recall	Precision	F-score
French	0.788	0.482	0.908	0.630
Italian	0.848	0.545	0.840	0.661
Russian	0.792	0.378	0.920	0.536
Portuguese	0.578	**0.985**	**0.342**	0.508
Spanish	0.783	0.564	0.923	0.700

TABLE VIII: Classification metrics for the assessment of English Perspective against the classes predicted by non-English Perspective

	Accuracy	Recall	Precision	F-score
French	0.840	0.853	0.565	0.680
Italian	0.868	0.893	0.582	0.705
Russian	0.815	0.869	0.404	0.552
Portuguese	0.614	**0.397**	**0.990**	0.566
Spanish	0.855	0.897	0.678	0.772

A. Discussion of Non-English Perspective

Table VII provides a baseline for the performance of Perspective API in the original language of the dataset (i.e. with no translation). This represents the intended use-case of Perspective API. The results show better performance in French, Italian and Spanish than in Portuguese and Russian. The binary accuracy was close to 0.8 for all languages except Portuguese (although the F-scores reveal Russian to be worse than the other three). Portuguese was classified particularly poorly, with low binary accuracy (0.578) and low F-score (0.508). It should be noted when comparing results that the Italian dataset was approximately half the size of the other four (as shown in Table V). The confusion matrices associated with this comparison, Table III, show there is also a trend amongst all languages, except from Portuguese, for more originally toxic comments to be misclassified than originally non-toxic (also indicated by the trend in high precision but low recall scores). These examples are displayed in bold in Table III. This suggests that Perspective API has a lower threshold to label comments as toxic than a human. This may be a conscious feature (i.e. to have more confidence that potentially toxic comments are not shown to users), or it may be a result of the corpora that Perspective was trained on (online forums such as Wikipedia and The New York Times comment section [18]) differing in their definition of 'toxicity' to these datasets[5]. This is in contrast to the Portuguese dataset, in which a large proportion of human-labelled non-toxic comments were misclassified as toxic.

[5]Perspective API uses the definition: "A comment that is rude, disrespectful, unreasonable, or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective"

B. Discussion of English Perspective

The metrics for assessing the level of agreement between the predicted English Perspective class ($\hat{\mathbf{x}}$) and the predicted non-English Perspective class ($\hat{\boldsymbol{y}}$) are shown in Table VIII. In order to use standard classification metrics, a ground-truth label had to be set which was chosen as the predicted non-English Perspective class ($\hat{\boldsymbol{y}}$). This comparison represents the choice between automatically classifying comments in their original language, or English. There is a greater level of agreement between both Perspective models, than there is between the ground truth and non-English Perspective. This suggests Google's Cloud Translate API is capable of maintaining toxic sentiment. Table IV, however, demonstrates that there are a significant number of comments (displayed in bold) which effectively *became* toxic following machine-translation. Using Russian as an example, 636 of the 3296 comments deemed non-toxic by Perspective API in Russian were classed as toxic by Perspective API in English. This level of disagreement calls for automated classification post-translation to be considered alongside classification of the source text. For Portuguese, the opposite is true, with 38.4% of the comments *becoming* non-Toxic post-translation (according to the two Perspective models). The high precision but low recall score for Portuguese (shown in bold in Table VIII) is an artifact of this type of misclassification.

C. Future Work

With regards to expanding the datasets used in future analysis of Perspective API, our filtering strategy can be expanded to be less discriminatory in the comments it rejects. This will result in a larger proportion of comments including common social media traits, such as usernames (preceded by '@'), URLs, and hashtags all of which are not currently considered to be valid inputs. Additionally, while we considered an off-the-shelf classifier, [7] [10] [19] all investigate training their own toxic/sentiment classifiers which could be done with the datasets we've collected and built. This would allow for assessment of how the inclusion of machine-translated text in the training set of toxic-comment classifiers affects accuracy.

V. Conclusion

In this paper we have demonstrated that machine-translation has a meaningful impact on the automated toxic classification of online comments using Perspective API and Google's Cloud Translate API. We have shown the level of agreement between classification in the original (non-English) language, and classification in English is heavily dependent on the specific language, with the best performance on comments in French, Italian and Spanish. If a user selects to translate a comment to English, we show there could be benefit in platforms classifying the translated comments again in English (particularly for Portuguese and Russian).

REFERENCES

[1] Google. (2021) How it works: Using machine learning to reduce toxicity online. [Online]. Available: https://perspectiveapi.com/how-it-works/

[2] "Case studies." [Online]. Available: https://perspectiveapi.com/case-studies/

[3] "About automatic language translation for ads." [Online]. Available: https://www.facebook.com/business/help/352340477433999 1?id=649869995454285

[4] G. Shalunts, G. Backfried, and N. Commeignes, "The impact of machine translation on sentiment analysis," *Data Analytics*, vol. 63, pp. 51–56, 2016.

[5] A. Balahur and M. Turchi, "Multilingual sentiment analysis using machine translation?" in *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. Jeju, Korea: Association for Computational Linguistics, Jul. 2012, pp. 52–60. [Online]. Available: https://aclanthology.org/W12-3709

[6] O. Makhnytkina, A. Matveev, D. Bogoradnikova, I. Lizunova, A. Maltseva, and N. Shilkina, "Detection of toxic language in short text messages," in *Speech and Computer*, A. Karpov and R. Potapova, Eds. Cham: Springer International Publishing, 2020, pp. 315–325.

[7] M. Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis," *Information Sciences*, vol. 512, pp. 1078–1102, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025519309879

[8] S. Sagnika, A. Pattanaik, B. S. P. Mishra, and S. K. Meher, "A review on multi-lingual sentiment analysis by machine learning methods," *ISI Digital Commons*, vol. 506, pp. 154–166, 2020.

[9] S. M. Mohammad, M. Salameh, and S. Kiritchenko, "How translation alters sentiment," *Journal of Artificial Intelligence Research*, vol. 55, pp. 95–130, 2016.

[10] Z. Wang, S. Lee, S. Li, and G. Zhou, "Emotion detection in code-switching texts via bilingual and sentimental information," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 763–768.

[11] S. G. Roy, U. Narayan, T. Raha, Z. Abid, and V. Varma, "Leveraging multilingual transformers for hate speech detection," *CoRR*, vol. abs/2101.03207, 2021. [Online]. Available: https://arxiv.org/abs/2101.03207

[12] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, Aug. 2021, pp. 299–318. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/kumar

[13] E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan, "Adversarial text generation for google's perspective api," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018, pp. 1136–1141.

[14] S. Brown, P. Milkov, S. Patel, Y. Z. Looi, Z. Dong, H. Gu, N. S. Artan, and E. Jain, "Acoustic and visual approaches to adversarial text generation for google perspective," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 355–360.

[15] "Jigsaw. 2020. jigsaw multilingual toxic comment classification." [Online]. Available: https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification,July.

[16] "Translation api." [Online]. Available: https://cloud.google.com/translate

[17] "The corncob list of more than 58 000 english words." [Online]. Available: http://www.mieliestronk.com/wordlist.html

[18] "Training data." [Online]. Available: https://developers.perspectiveapi.com/s/about-the-api-training-data

[19] G. Xie, "An ensemble multilingual model for toxic comment classification," in *International Conference on Algorithms, Microchips and Network Applications*, N. Sun and F. Cen, Eds., vol. 12176, International Society for Optics and Photonics. SPIE, 2022, pp. 429 – 433. [Online]. Available: https://doi.org/10.1117/12.2636419