Harmful Language Datasets: An Assessment of Robustness

Katerina Korre
University of Bologna
aikaterini.korre2@unibo.it

John Pavlopoulos
Athens University of Economics and Business
annis@aueb.gr

Jeffrey Sorensen
Google Jigsaw
sorenj@google.com

Léo Laugier
EPFL
leo.laugier@epfl.ch

Ion Androutsopoulos
Athens University of Economics and Business
ion@aueb.gr

Lucas Dixon
Google Research
ldixon@google.com

Alberto Barrón-Cedeño
University of Bologna
a.barron@unibo.it

Abstract

The automated detection of harmful language has been of great importance for the online world, especially with the growing importance of social media and, consequently, polarisation. There are many open challenges to high quality detection of harmful text, from dataset creation to generalisable application, thus calling for more systematic studies. In this paper, we explore re-annotation as a means of examining the robustness of already existing labelled datasets, showing that, despite using alternative definitions, the inter-annotator agreement remains very inconsistent, highlighting the intrinsically subjective and variable nature of the task. In addition, we build automatic toxicity detectors using the existing datasets, with their original labels, and we evaluate them on our multi-definition and multi-source datasets. Surprisingly, while other studies show that hate speech detection models perform better on data that are derived from the same distribution as the training set, our analysis demonstrates this is not necessarily true.

1 Introduction

Many forms of harmful language impact social media despite efforts —legal and technological— to suppress it.[1] Social media has been under significant scrutiny with regard to the effectiveness of their anti-hate speech policies, which usually involve users manually reporting a potentially malicious post in order to trigger a human review, and platforms adjusting their community guidelines by, for example, banning hateful comments, and employing automated moderation assistants.

A robust and general solution to the problem does not yet exist, and given that there are many factors that influence the phenomenon of online hate speech, we expect this area of research to continue to pose significant challenges. One of the main reasons is that harmful language detection is an inherently subjective task. There have been many attempts to approach harmful language detection by introducing or selecting specific definitions (Fortuna et al., 2020). From blanket terms, such as abusiveness and offensiveness to sub-categories, such as misogyny and cyber-bullying, researchers have explored many variants. However, this begs the question of how to select and compare the possible definitions, especially when some categories are more efficient for cross-dataset training than others (Fortuna et al., 2021). The problem gets more intricate when multiple languages are involved, and when the translation of a term does not necessarily carry the same implications as in the source language. This can have significant implications for the development of cross-lingual systems (Bigoulaeva et al., 2021; Deshpande et al., 2022).

In this study, we attempt to shed light on the effectiveness of different definitions of harmful language both for annotation purposes and model development. We use the term "harmful language" as a wildcard term that can be potentially replaced with terms like toxic, hate speech, and offensiveness, among others. We perform a re-annotation of existing datasets with a range of definitions and replicate the experiments to assess robustness. Then, we perform a qualitative error analysis on the re-annotations, showing that even instances that contain potentially harmful terms might not be perceived as harmful by annotators, underlining the subjectivity of the task. Finally, we analyse the generalisability of the existing datasets across the different definitions by training BERT-based classifiers with the original annotations and with our re-annotations, concluding that evaluating on broader definitions can yield higher accuracy.

The rest of this article is structured as follows. Section 2 overviews existing studies on the issue of the definition of harmful language and its implications, as well as how state-of-the-art (SOTA) sys-

[1] https://edition.cnn.com/2022/06/14/asia/japan-cyberbullying-law-intl-hnk-scli/index.html

The 7th Workshop on Online Abuse and Harms (WOAH), pages 221–230

tems handle generalisability. Section 3 presents our re-annotation strategy. In Section 4, we describe our experimental setup for training and evaluating with the the original and the re-annotated datasets. Finally, after presenting our results in Section 4.3, we assess our contribution in Section 5, concluding by speculating on limitations and future work.

Disclaimer: This paper contains potentially offensive, toxic, or otherwise harmful language.

2 Related Work

Harmful language is becoming all the more frequent due to the widespread use of social media and the Internet, thus creating a vicious cycle that compromises the civility of the online community and threatens a healthy user experience (Nobata et al., 2016). The need for automatically moderating toxic language has led to the development of a considerable body of related work, proposing solutions and highlighting existing problems.

2.1 Generalisability

One of the most frequently discussed problems is the inability of toxicity detection models to generalise, namely the fact that models underperform when tested on a test set from different source than the training set (Swamy et al., 2019; Karan and Šnajder, 2018; Gröndahl et al., 2018). Yin and Zubiaga (2021) claim that, when models are applied cross-lingually, this performance drop indicates that model performance had been severely over-estimated as testing on the same dataset the training set derived from is not a realistic representation of the distribution of unseen data. Attempts to improve the performance of such models involve merging seen and unseen datasets, using transfer learning, and re-labelling (Talat et al., 2018; Karan and Šnajder, 2018). However, in the majority of cases, instances from the source dataset are needed to achieve high performance (Fortuna et al., 2021). In addition, various characteristics of datasets have been examined as variables for an effective generalisation, including the work of Swamy et al. (2019), who suggested that more balanced datasets are healthier for generalisation, and that datasets need to be as representative as possible of all facets of harmful language, in order for detection models to generalise better.

2.2 The Challenge of Definitions

Properly defining toxic content poses a great challenge, not only in computational linguistics but also in socio-linguistics and discourse studies. Discussing two important terms 'trolling' and 'flaming', KhosraviNik and Esposito (2018) very eloquently suggest that "[d]espite the widespread (and often overlapping) use of these two terms, the utmost complexity of the discursive practices and behaviours of online hostility has somehow managed to hinder the development of principled definitions and univocal terminology". Regarding hate speech, according to Davidson et al. (2017), no formal definition exists yet, while also legislation differs from place to place, rendering the creation of a universal framework very difficult. The NLP community usually deals with this problem by adapting definitions to their specific purposes. However, Fortuna et al. (2020) suggest that this can lead to the use of ambiguous or misleading terms for equivalent categories. The authors come to the conclusion that it is necessary to accurately define 'keyterms' in order to achieve better communication and collaboration in the field.

3 Methodology

Our methodology is divided in two parts. The first part investigates whether closely-related definitions have an effect on inter-annotator agreement while the second part examines the compatibility and versatility of the present datasets by using them to train models.

3.1 Annotation Experiments

In order to study the effect of the definition on inter-annotator agreement, we re-annotated toxicity datasets by using alternating definitions and by repeating the annotation in rounds for robustness.

Datasets For this study we try to use the same data used in Fortuna et al. (2020) in order to produce comparable results. However, not all of the datasets could be used, as the classes used would make it harder for the models to generalise since they were referring to specific target groups. For example, the AMI (Fersini et al., 2018) and HatEval (Basile et al., 2019) datasets referred specifically to women or immigrant minorities. Therefore, the final selection of datasets includes Davidson (2017), TRAC-1 (Kumar et al., 2018b), and Toxkaggle (Jigsaw, 2019). It must also be noted

Term	Definitions of harmful language	Citation
TOXIC	A rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.	Jigsaw (2019))
ABUSIVE	Hurtful language, including hate speech, derogatory language and also profanity	Founta et al. (2018)
OFFENSIVE	Containing "any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct."	Zampieri et al. (2019)
HATE	Expressing hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. In extreme cases this may also be language that threatens or incites violence.	Davidson et al. (2017)
HOTA	*Any* of the following: Hateful, Offensive, Toxic, Abusive language (HOTA)	Ours

Table 1: The terms and definitions of harmful language that were provided to the annotators during re-annotation.

that, for this research, the Davidson dataset is split into two subsets: DavidsonHS (for hate speech) and DavidsonOFF (for offensiveness), as the two classes correspond to two different definitions.

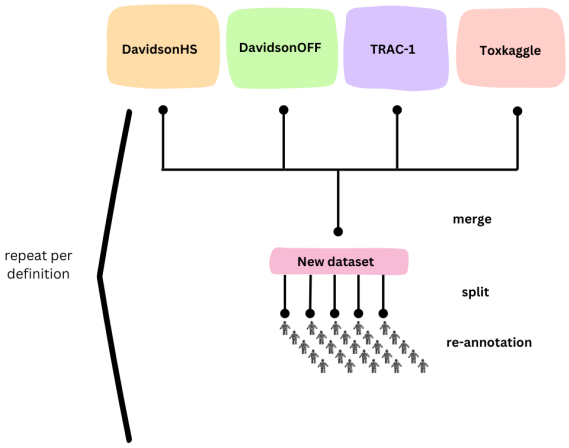

Figure 1: Annotation procedure. Instances from the 4 datasets were used to create a new dataset that would later be divided into 5 annotation batches.

Data Compilation For our annotation purposes, we create 5 different batches of data that contain instances from all aforementioned datasets. Each batch contains an equal number of different instances from each dataset, while the instances are also shuffled. To be able to map the datasets with the corresponding instances later in the analysis, a code is given for each dataset, as well as to anonymise it. The total number of instances of each of the batches was 200 (out of which we randlomly selected 80 as test questions, for quality control). In each batch we keep a balanced distribution between positive and negative instances, while we also keep the balance among the classes derived from each dataset, following the suggestions of Swamy et al. (2019) for better generalisation. Information about class distribution for each batch is presented in brackets in the column Classes in Table 2.

Annotation Procedure The annotation procedure consists of five annotation experiments, each relating to a different definition for potentially harmful content. For the annotation, we used crowdsourcing via the Appen platform.[2] The guidelines for the annotations can be found in the Appendix A. Since this project was carried out in collaboration with Jigsaw, [3] the raters were compensated according to the company's regulations, namely a compensation above minimum wage for the annotator region (USA), based on estimates of time to task completion. Jigsaw's regulations with regard to Appen annotations include reviewing feedback from raters to insure that the task is considered doable and that the raters feel they are compensated fairly. Each annotation experiment was repeated 5 times with different data each time. This variation in the data helps to ensure that the results are not specific to a particular dataset and can be generalized. Regarding the guidelines, annotators were instructed to read carefully the given definition and examples, and decide whether each text was harmful or not according to the definition provided. The same examples were provided to the annotators across all annotation experiments, and the only thing changed was the term and the definition of harmful language, presented in Table 1. Since we used crowdsourcing, each batch is not necessarily annotated by the same annotators. The quality of the annotators was ensured provided they answer correctly the aforementioned test questions. The annotation procedure is also summarised in Figure 1.

3.2 Annotation analysis

An initial exploratory analysis of the results of the annotation not only shows low inter-annotator agreement in general but also inconsistency both across datasets and across repetitions. This is evi-

[2] https://appen.com/
[3] https://jigsaw.google.com/

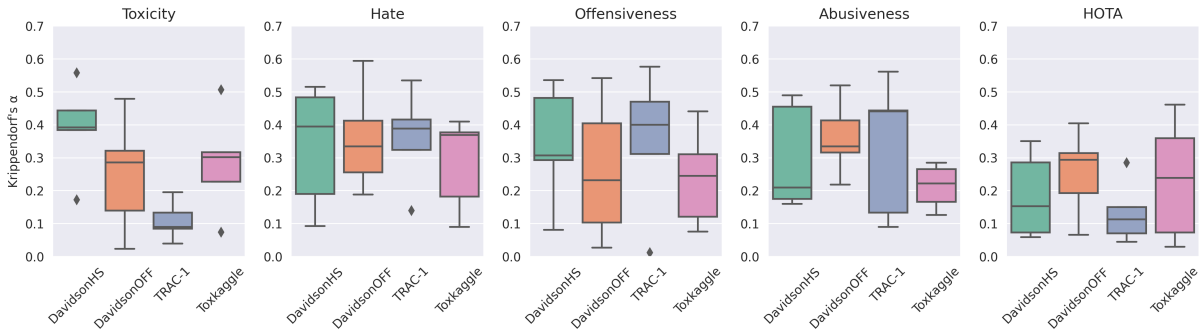

Figure 2: Boxplots showing Krippendorf's alpha inter-annotator agreement. The y axis shows the Krippendorf's alpha values while the x axis shows the different datasets. Each plot refers to a different definition.

dent in Figure 2. Among the 5 definitions, Toxicity and HOTA (see Table 1 for the acronym explanation) show more consistent annotation despite the low inter-annotator agreement, which is under 0.5. This poses the question of whether we should trust high inter-annotator agreement and potential inconsistency among repetitions or accept a lower but more robust inter-annotator agreement. Moreover, looking at the inter-annotator agreement per dataset, we see that instances of datasets that were originally annotated with a given definition present a more consistent annotation when re-annotated with another definition. For example, we would expect DavidsonHS to have a more consistent inter-annotator agreement when annotated for hate speech, but we see that it is when it is annotated for toxicity that the result is more robust. Similarly, DavidsonOFF presents slightly more consistent results when annotated for hate speech and abusiveness rather than offensiveness.

Annotation variance can be used to isolate instances with high disagreement. Table 3 presents a subset out of the 10 instances with the highest variance per definition that were sampled for the analysis. When annotated for toxicity, these posts included forms of irony. For instance, the example of the 1st row is possibly written by a woman, which might mean that the intention is not to be toxic but to cauterize misogynistic behaviours. In addition, many posts contained vocabulary that is associated with negative sentiments, such as "crazy", "cheater", and "hate". With regard to abusive language, annotators disagreed even for instances that present raw profanity ("bitch", "cocksucker"), potential racism as seen in the 2nd example of the table, and ableism as seen in the third. Similarly, when annotating for offensiveness, the raters did not necessarily annotate positively an instance that contained profanity. Also, racist instances that do not contain obscenities might have been trickier to classify. For example, the author of the 4th example resorts to ostensibly logical reasoning that might disguise the racism that pervades the sentence. Compared to the other definitions that were given during the re-annotation, the sampled re-annotations for hate speech did not show any clear pattern possibly because the definition of hate speech is more restricting referring to specific target groups. However, the same holds true for HOTA, which was the broader term during the re-annotation. The sample that we checked during this qualitative analysis included profanity, references to homosexuality or racism and misogyny, as well as instances that did not contain any harmful language. Noteworthy is also the fact that the sentence in Example 5 appeared with high variance in 3 out of 5 definitions, possibly because of the mixed language use and modified words.

4 Experimental setup

4.1 Datasets

We use the same four datasets that were used in annotation (Davidson et al., 2017; Kumar et al., 2018b) to perform toxicity/hate speech/offensiveness/aggressiveness classification. More specifically, we first extracted the 1,000 (200 per definition) instances used for the human annotation from the original datasets. Then, with the remaining instances we created 4 balanced datasets that contained an equal amount of positive and negative instances (2650 in total), with 10% of the data used for development. The evaluation of the model was carried out by calculating the accuracy with respect to the original annotation labels and the ones produced for the new annotation.

Dataset	Annotation Procedure	Classes	Source
DavidsonHS (Davidson et al., 2017)	Begining with the hatebase lexicon then CrowdFlower, users coded each tweet (minimum number of annotations per tweet is 3 , sometimes more users coded a tweet when judgments were determined to be unreliable by CF).	Hate speech (25), Not-Hate Speech (25)	Twitter
DavidsonOFF (Davidson et al., 2017)	>>	Offensiveness (25), Not-offensiveness (25)	Twitter
TRAC-1 (Zampieri et al., 2019; Kumar et al., 2018b,a)	The annotation was done using the Crowdflower platform but by what is known as 'internal' annotators in the Crowdflower lingo. The whole of annotation was done by 4 annotators – all of them were native speakers of Hindi, with a nativelike competence in English and were pursuing a doctoral degree in Linguistics.	Overtly Aggressive (OAG) (13), Covertly Aggressive (CAG) (12), Non-Aggressive (NAG) (25)	Facebook
Toxkaggle (Jigsaw, 2019)	Not provided.	Threat (3), Identity hate (3), Severe Toxic (3), Insult (3), Obscene (4), Toxic (9), NonToxic (25)	Wikipedia

Table 2: Basic description of dataset. This table was inspired by a similar table found in Fortuna et al. (2020). Davidson (2017) dataset was split into two separate datasets as Hate Speech and Offensives are too different as definitions.

4.2 Model training

We fine-tuned BERT with early stopping,[4] using patience of 3 and a max length defined per dataset, i.e., the mean length with one unit of standard deviation: 30 tokens for DavidsonHS, 37 for DavidsonOFF, 70 for Trac-1, and 100 for Toxkaggle. The code is publicly available.[5]

4.3 Results

We assess the classifiers using both the original and the re-annotated ground truth.

Using the source annotations as our evaluation ground truth, the accuracy of the classifiers is presented in Figure 3. We observe that when the model is trained on DavidsonHS datasets, it reaches an accuracy of more than 0.75 in all test sets. As expected the accuracy is higher when the model is also tested on DavidsonHS. When the model is trained on DavidsonOFF the accuracy is high only when tested again on DavidsonOFF. Training on

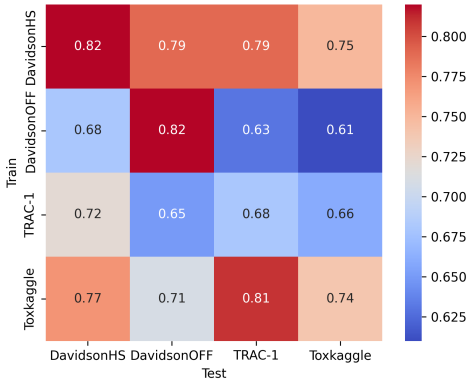

Figure 3: Heatmap showing the accuracy on the different test sets using the original ground truth (horizontally) when the model is trained on each corresponding dataset (vertically).

Toxkaggle results in more than 0.70 accuracy in all test sets, with the the highest accuracy in the TRAC-1 test set (0.81). TRAC-1, on the other hand, shows the lowest accuracy across all test sets (0.65-0.72), with the highest accuracy obtained when tesing on DavidsonHS.

[4] We used the base model and the KTRAIN library.

[5] https://github.com/katkorre/reannotation-robustness.git

	Example	Source
1.	*according to you my place is in kitchen. Excuse me while I take out the trash*	Davidson
2.	*It's because you can't bend me, you can't buy me and you can't make me into your nigger!*	Davidson
3.	*California's biggest retards*	Davidson
4.	*"Just because she has light skin doesn't mean anything... I know some Asian with really dark brown skin but that doesn't make them any less Chinese"*	Toxkaggle
5.	*and we shud also destroy taj mahal... qutub minar ....laal qila aftr that visiters seen only GBRoad vd hvng a fun vd ur neighbours*	TRAC-1

Table 3: Texts from the evaluation sets with the highest variance of inter-annotator agreement.

		Evaluation (re-annotated)			
Training	**Definition**	**DavidsonHS**	**DavidsonOFF**	**TRAC-1**	**Toxkaggle**
DavidsonHS	Toxicity	0.75 (-0.07)	0.69 (-0.10)	0.75 (-0.04)	**0.83** (+0.08)
	Hate Speech	0.64 (-0.18)	0.59 (-0.20)	0.58 (-0.21)	0.59 (-0.16)
	Offensiveness	0.64 (-0.18)	0.64 (-0.15)	0.56 (-0.23)	0.62 (-0.13)
	Abusiveness	0.63 (-0.19)	0.57 (-0.22)	0.62 (-0.17)	0.59 (-0.16)
	HOTA	**0.76** (-0.06)	**0.78** (-0.01)	**0.78** (-0.01)	0.82 (+0.07)
DavidsonOFF	Toxicity	**0.64** (+0.04)	0.72 (-0.10)	0.83 (+0.20)	**0.75** (+0.14)
	Hate Speech	0.58 (-0.10)	0.63 (-0.19)	0.59 (-0.04)	0.59 (-0.02)
	Offensiveness	0.50 (-0.18)	0.66 (-0.16)	0.56 (-0.07)	0.59 (-0.02)
	Abusiveness	0.57 (-0.11)	0.61 (-0.21)	0.62 (-0.01)	0.60 (-0.01)
	HOTA	0.62 (-0.06)	**0.76** (-0.06)	**0.84** (+0.21)	0.74 (+0.13)
TRAC-1	Toxicity	0.67 (-0.05)	0.59 (-0.06)	0.50 (-0.18)	0.53 (-0.13)
	Hate Speech	0.69 (-0.03)	**0.66** (+0.01)	0.53 (-0.15)	0.63 (-0.03)
	Offensiveness	0.69 (-0.03)	0.63 (-0.02)	**0.55** (-0.13)	**0.66** (=)
	Abusiveness	0.70 (-0.02)	0.64 (-0.01)	0.51 (-0.17)	0.65 (+0.01)
	HOTA	**0.71** (-0.01)	**0.66** (+0.01)	0.47 (-0.21)	0.57 (-0.09)
Toxkaggle	Toxicity	0.73 (-0.04)	0.67 (-0.04)	0.77 (-0.04)	**0.85** (+11)
	Hate Speech	0.68 (-0.09)	0.67 (-0.09)	0.63(-0.18)	0.61 (-0.13)
	Offensiveness	0.67(-0.10)	0.69 (-0.02)	0.63(-0.18)	0.68(-0.06)
	Abusiveness	0.71 (-0.06)	0.65 (-0.06)	0.63 (-0.18)	0.61 (-0.13)
	HOTA	**0.79** (+0.02)	**0.77** (+0.06)	**0.81** (=)	0.83 (+0.08)

Table 4: Accuracy of BERT trained per dataset (1st column), using the original annotations, and evaluated on our re-annotations per definition. In parentheses is the accuracy increase (green) or decrease (red) compared to the scores obtained on the evaluation data with the original annotations (Figure 3).

Using our re-annotations as the evaluation ground truth, is shown in Table 4. Models did not manage to generalise across datasets consistently, which is shown by the fact that accuracy decreases, in comparison to the scores obtained when the original annotations were used for testing our models. There are sparse exceptions were the accuracy increases, for example, when training on Toxkaggle and testing on re-annotations of HOTA, where results were equal (TRAC-1) or better (DavidsonHS, DavidsonOFF, Toxkaggle). In general, the highest accuracy, although still low in terms of what current language models can achieve, is achieved when testing either on the toxicity or HOTA re-annotations. Excluding Toxkaggle, however, we observe that accuracy deteriorated in our re-annotations even when evaluating on test sets derived from the same source as the training set, except for TRAC-1 that it presents a slight increase of 0.01 when testing on hate speech and HOTA.

5 Discussion

Taking into account the existing literature (Fortuna et al., 2020; Karan and Šnajder, 2018; Swamy et al., 2019; Yin and Zubiaga, 2021), this study confirms

that models face a serious difficulty generalising. Yet, our results show a promising aspect when it comes to model reproducibility for harmful language detection purposes, as well as building robust datasets though a robust annotation procedure.

5.1 Accuracy per definition

Models perform better in the two most general definitions, i.e., Toxicity and HOTA (Table 4). This can be due to pragmatic reasons, namely classifying items using broad definitions can be an easier task for both the annotators and the models. On the other hand, it might be a matter of compatibility between the training data and the testing data. For example, the classes used in the re-annotation procedure were more similar to the ones used in the two Davidson sub-sets and Toxkaggle, while they were more different compared to TRAC-1, where another definition was originally used (aggressiveness), which we did not include in our experiments.

5.2 Robustness and reproducibility

If we consider the evaluation on the original gold labels (Figure 3) as the baseline of the experiment, and compare with the re-annotations (Figure 4, we see that in many cases the performance fluctuates when the models are tested on our re-annotated data. Specifically, the performance drops when the models are tested on the re-annotations of the same source as the training set, while it can occasionally increase when tested on the re-annotations of a different source from that of the training set. This implies that the models' performance is sensitive to the specific data sources used for re-annotation. It suggests that it is possible that the models may struggle to generalise well to new data from the same source, resulting in a drop in performance and contrasting previous studies. On the other hand, there are cases that when presented with re-annotations from a different source and under certain conditions (providing a specific definition), the models might perform better, indicating a potential capability to generalise across different data sources, even when the source of the test set is different from that of the training set.

5.3 Drawing the line

Focusing on such differences among different datasets could enable researchers to outline the DOs and DON'Ts for annotations and dataset creation. Finding the correct combination between the appropriate definition to use and the correct data source can be pivotal for an efficient harmful language detection model. Moreover, we underline the need for parallel annotation (both longitudinal and by increasing the number of annotators) as "collecting the opinions of more users gives a more detailed picture of objective (or intersubjective) hatefulness" (Roß et al., 2016). According to Fortuna et al. (2020), fine-grained toxicity categories are not the optimum option, while more general categories yield better results. Considering that, for the purposes of this experiment, we tried to binarise and simplify the datasets, as much as possible, by separating the Davidson dataset and by merging the subcategories in TRAC-1 and Toxkaggle. However, this did not help the performance when it comes to TRAC-1. One possible reason behind this could be the fact that TRAC-1 contains implicit aggressiveness that is harder to detect, even when the model is trained on the respective dataset. The difficulty to detect implicit aggressiveness or other forms of harmful language is not only true for models, but also for human annotators, as we saw in Section 3.1.

6 Conclusion

In spite of recent advances, model generalisation and method robustness still has a long way to go especially regarding harmful language online. In this study, we attempt to shed some light on the issue, first, by performing a re-annotation experiment with existing datasets employing crowdsourcing annotators and, second, by using the same datasets to train a baseline model as an automatic annotator. The human annotation shows that, although in most cases the annotations were inconsistent, Toxicity and HOTA (any of the following: Hateful, Offensive, Toxic, Abusive language) appear to be the most consistent definitions, indicating that the broader the term used the more robust the annotations The experimental model, on the other hand, showed that, assessing on data from the same source as the training set, when using the original ground truth, can yield higher accuracy compared to assessing data from a different source, confirming previous studies. Yet, this cannot be used as a rule of thumb since testing on the re-annotations showed that the performance can drop when testing on the data from the same source as the training set and it can increase when testing on previously completely unseen data.

Limitations

Our study is limited in three perspectives. First, not all datasets relevant to toxicity have been studied. Also, we only experimented with BERT-based classifiers. We let the study of more datasets and algorithms for future work. Another limitation is that our annotation is only based on crowdsourcing, but the opinion of expert annotators could also be acquired. We note that such an extension would also allow a study of the effect of the quality of the two different approaches (crowdraters vs. experts) on model performance.

Ethical statement

The ethical considerations of this study mainly concern the re-annotation procedure. The original datasets were anonymised before re-annotating. After the re-annotation, and as instructed by the Appen platform, we avoided including any sensitive information of the annotators by only using their IDs for identifying any particular instance.

Acknowledgements

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.

References

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.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Neha Deshpande, Nicholas Farris, and Vidhur Kumar. 2022. Highly generalizable models for multilingual hate speech detection. *ArXiv*, abs/2201.11294.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA@CLiC-it*.

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 Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

Jigsaw. 2019. Jigsaw toxic comment classification challenge. `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`. Accessed: [8 May 2023].

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Majid KhosraviNik and Eleonora Esposito. 2018. Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz Papers in Pragmatics*, 14(1):45–68.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chikashi Nobata, Joel R. Tetreault, Achint Oommen Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.

Björn Roß, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

A Annotation instructions and interface

Figures 4 and 5 in the next page of this Appendix present the instructions (only for toxicity shown) and the interface the annotators were provided with during their re-annotation tasks.

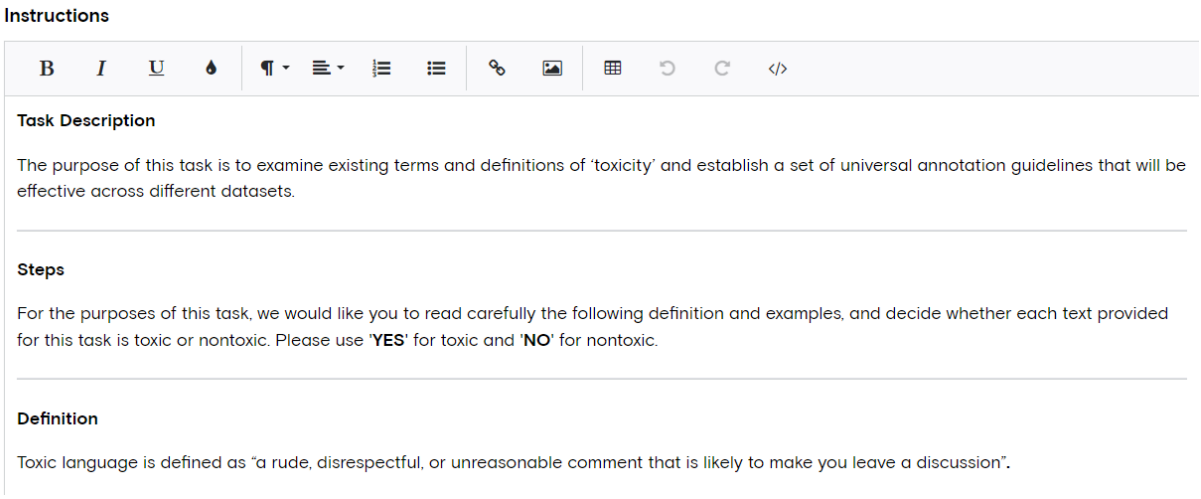

Figure 4: Instructions during re-annotation, using the term and definition of Toxicity.

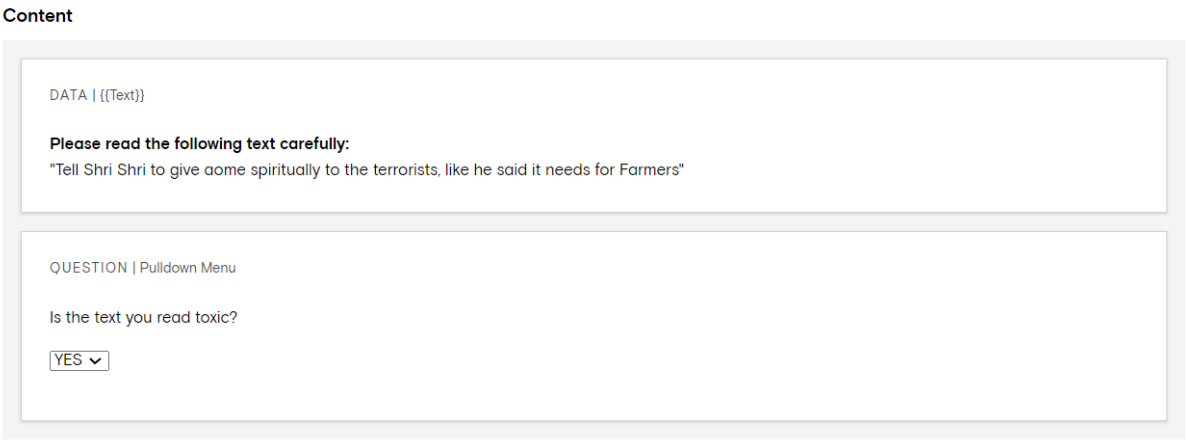

Figure 5: Interface for re-annotation.