META: Text Detoxification by leveraging METAmorphic Relations and Deep Learning Methods

Alika Choo*, Arghya Pal†, Sailaja Rajanala‡ and Arkendu Sen§
Monash University Malaysia, Malaysia
E-mail: {alika.choo*, arghya.pal†, sailaja.rajanala‡, arkendu.sen§}@monash.edu

***Abstract*—In the world of online interactions, social communities face a significant challenge: the spread of offensive content and hate speech through toxic languages. Such issues led to growing research on text detoxification systems that can automatically rewrite toxic content. A systematic evaluation is required to ensure these systems produce high-quality detoxified text that modifies the original text to be non-toxic while preserving its content. However, this often relies on large amounts of labelled data and human judgement, which may not always be feasible. This limitation is typically known as the *oracle problem*. Metamorphic testing (MT) has conventionally been used to solve the oracle problem by deriving metamorphic relations (MRs) to test a program's functionality. A new MT approach focused on data validation showed that MRs incorporated with tools can be used to identify defects in machine translation services. This paper draws inspiration from this new MT perspective by presenting four metamorphic relations incorporated with tools to evaluate style transfer accuracy, content preservation, fluency, and a joint of these three. Our proposed approach effectively identifies defective behaviour in state-of-the-art text detoxification systems.**

I. Introduction

Warning: *This paper contains offensive language that commonly appears in hate speeches. Hate speech text present in this paper does not represent the views of the authors!*

♠ **Divisive hate speech and ideologies on a global scale** Social media has been serving as an outlet for people to display hate speech and harmful and offensive behaviour online, facilitated by user anonymity and lack of regulations on social media platforms. Such toxic content often contributes to cyberbullying – using technology to send hostile messages intended to inflict harm to a target individual or group – which has been postulated to be the cause of recent rising suicide rates in adolescents. Furthermore, the situation is so severe that the United Nations resolution on 18 June proclaimed the date to be remembered as the International Day for Countering Hate Speech. This paper, along with other sister works [2], [3], recognises the need to develop an AI-driven text detoxification method.

♠ **AI-driven Deep Learning based Hatetext Detoxification** This prevalence of online toxicity and the potential harm that it can bring to communities has prompted significant research in automatic detection systems for toxic speech and text detoxification systems [2][3]. Text detoxification is the autonomous rewriting of toxic content. It is understood as a style transfer task where the original text is rewritten to change its style (i.e. non-toxic) while preserving its content. Text detoxification systems are an encouraging approach for reducing harmful online behaviour and neutralising emotional comments [4].

♠ **Challenges to Find Correct Pipeline for AI-driven Detoxification** The evaluation of the degree of detoxification, i.e. how many words or sentences have been detoxified, mostly depends on metrics like the Human Turing Test [4] or word embedding-based methods [4], thus making it time-consuming, costly, and subjective [5]. To automate this testing process for text detoxification systems, parallel datasets such as ParaDetox [6] and APPDIA [7] have recently served as benchmark detoxification datasets. Nevertheless, referring to a single ground truth may not be ideal given the ever-growing complexity of online comments [8]. Thus, relating those studies to famous ***oracle problem*** [9]; An oracle (i.e. the evaluation) is a procedure determining whether a program (i.e. the AI-driven Deep Learning model) has produced the correct output (i.e. the detoxified text) [9]. The context of toxic language detoxification arises from the subjective nature of what constitutes an appropriate "detoxification" for a given text.

♠ **Metamorphic Testing: An Introduction** Metamorphic testing (MT) has conventionally addressed the oracle problem for software through metamorphic relations (MRs). MRs are relations we expect to hold over multiple inputs and their expected outputs [9] [10]. These relationships are more general compared to having single-ground truths, yet they set an expectation for what a valid output should be. A violation of an MR highlights an erroneous behaviour in the program. Conventional MT helps software test coverage. It relies on source test cases (a selected set of program inputs) and transformations [9] to build metamorphic test cases that share similar specifications with the source test cases. Finally, MT verifies the source and metamorphic test cases with their outputs against an appropriate MR [10]. However, conventional MT is not without its practical challenges. It often requires expert knowledge and the time-consuming generation and validation of metamorphic test cases [11]. Furthermore, with advancements in large language models (LLMs) like BERT,

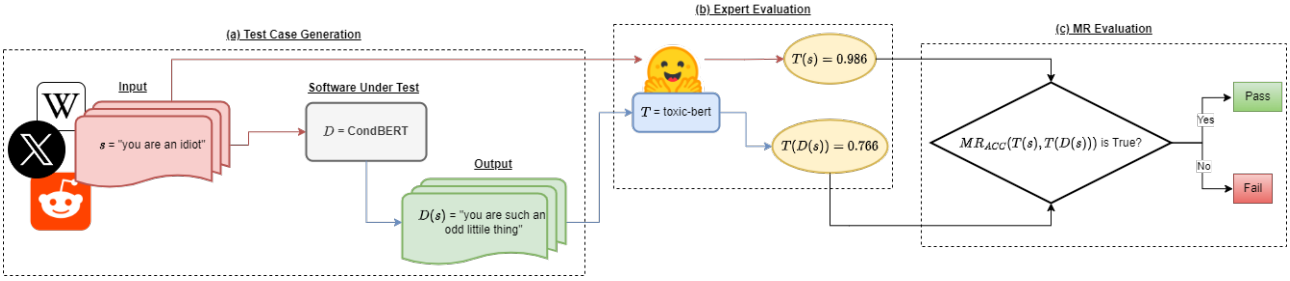

Fig. 1. MR for Data Validation approach [1] adapted for the text detoxification domain to evaluate style transfer accuracy. Our methodology, as discussed in Section IV has three distinct parts: (a) source inputs are passed to software under test (SUT) to generate detoxified outputs (b) experts compute a score for the source input and outputs individually (c) MR uses expert scores to determine whether the detoxified output is valid detoxification of the source input

the oracle problem is becoming less pronounced as these tools approach human-level judgment [3][4][6]. Hence, this leads us to the research question we aim to tackle for this paper *How can we incorporate such tools to act as "experts" for our testing processes to identify defects?*

♠ **Rethinking Metamorphic Testing for AI-driven Text Detoxification** Yan et al. [1] proposes a promising perspective focused on data validation, suggesting that MR violations might stem from input data quality rather than program faults. They demonstrated that this approach was helpful in automatically identifying poorly translated text messages using a sentiment analysis tool.

♠ **Our Contributions** In our paper, we emphasise less on data validation and focus back on identifying defects in AI-driven models under test (SUT), a concept not involved in the data validation MT approach [1]. We adapt their approach for the text detoxification domain and show the effectiveness of four metamorphic relations while incorporating state-of-the-art tools. The objectives of our study are to:

- Propose four metamorphic relations based on the Joint metric score proposed by Krishna et al. [12]:
  1) ACC - which classifies the level of non-toxicity
  2) SIM - which calculates the similarity between the original and translated text
  3) FL - which assesses fluency
  4) JOINT - combines ACC, SIM, and FL
- Suggest tools for evaluating detoxification models against toxicity removal, content preservation, and fluency

By adopting this approach, we can systematically test text detoxification systems for defects and quickly determine the reason for the defect (e.g., poor toxicity removal). Given that the process is autonomous, we limit the need for human intervention.

II. Text Detoxification Models

Text detoxification models can be categorised as supervised or unsupervised. To show that our methodology applies to different text detoxification systems, we exemplify our approach on the following three state-of-the-art text detoxification models that exhibit unique characteristics:

- CondBERT [4]: An unsupervised method where toxic words in a sentence are masked, and the candidates are reranked based on their non-toxicity scores.
- ParaGedi [4]: An unsupervised method where the detoxification task is viewed as a paraphrasing task, but toxicity scores of the next token prediction are also taken into account during the generation step.
- ParaDetox-BART [6] : A BART (base) model trained on the parallel detoxification dataset ParaDetox.

These models have benchmarked past research works [3][13]. If we show that our metamorphic relations are effective on these state-of-the-art models, it may apply to similar architectures.

III. Dataset

We selected various datasets to test our three models and validate the effectiveness of our proposed metamorphic relations. We chose these datasets according to the criteria:

- Language - Limited to English datasets
- Release Date - Not older than ten years
- Distribution - Most comments are toxic
- Test Dataset Size - Allows sampling of 10,000 sentences
- Toxicity Rating - Scores can be deduced to 0-1 toxicity rating

From these criteria, we selected the following datasets:

- Jigsaw 2018[1] - An annotated dataset of comments from Wikipedia's talk page edits that includes scores measuring toxicity and several subtypes of toxicity.
- WikiDetox 2017[2] - A human-annotated dataset sourced from English Wikipedia of over 100,000 discussion comments.

To prepare the test dataset from Jigsaw, we took the same test dataset used by Dale et al. [4]. For WikiDetox, we preprocessed the dataset (removing irrelevant tokens, removing duplicate comments) then created a test dataset by first standardising the sentences to 200 characters or fewer (including whitespaces), then taking 10,000 of the most toxic sentences based on our chosen pre-trained classifier that achieved an

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[2] https://figshare.com/articles/dataset/Wikipedia_Talk_Labels_Toxicity/4563973/2

AUC score of 98.64[3]. Finally, we will have two test datasets that we can use to evaluate each model.

Most of each test dataset consisted of content with considerable levels of toxicity, validated by the toxicity distributions we analysed for Jigsaw and WikiDetox test datasets, which were right-skewed. Most of its sentences were highly toxic (> 0.8).

IV. Methodology

This paper adapts the MT approach presented by Yan et al. [1] for testing text detoxification models. Our approach focuses less on validating given datasets and more on identifying defects in text detoxification models. Figure 1 illustrates our approach in evaluating a detoxified output for style transfer accuracy. The process looks similar when evaluating content preservation and fluency. In the following sections, we discuss the distinct phases of our MT methodology.

A. Test Case Generation

This step generates outputs from a given text detoxification model D that, when paired together with a source input, represents a *test case* as shown in Figure 1(a). D can be referred to as SUT in this context, although Yan et al. mention SUT is not present in their data validation methodology [1]. Nonetheless, since D shares the closest meaning to SUT in this context, we labelled D under SUT to align with conventional MT terminology. Each test case is then passed onto the next stage of evaluation.

B. Expert Evaluation

This phase differentiates this MT methodology from conventional MT, as depicted in Figure 1(b). The Expert Evaluation step simulates the manual process an expert undertakes to validate input quality, except in this approach, this step is autonomous using an appropriate tool. In this paper, we introduce the concept of 'Experts', whose role is to evaluate the quality of its input against a particular metric. An Expert in this context refers to a toxicity classifier, a similarity scorer, or a fluency scorer. Each of these Experts returns scores that are passed to the next step for evaluation against an appropriate MR.

C. MR Evaluation

Finally, our test cases are evaluated against one MR at a time, as shown in Figure 1(c). Test cases are considered to have passed if an MR holds and fails if it does not. For example, in Figure 1, given that MR_{ACC} requires $T(D(s))$ to be lower than $T(s)$, the test case passed. In later sections, we describe other MRs to evaluate content preservation, fluency, and joint scores.

[3]https://www.https://huggingface.co/unitary/toxic-bert

V. Metamorphic Relations For Data Validation

In this section, we describe our proposed MRs and mention the tools we have chosen to use for our experiments. The main goal is to select a tool that can perform closest to human judgement concerning one of the three style transfer metrics. Hence, the violations highlighted by tools must precisely contain erroneous outputs. For the following MRs, let L_1 and L_2 be languages prone to toxic content and devoid of toxic content, respectively. Let s_1 and s_2 be texts written in L_1 and L_2, respectively.

A. Style Transfer Accuracy

The goal of MR_{ACC} is to evaluate the disparity in toxicity ratings between two comments:

$$MR_{ACC} : T(s_2) - T(s_1) < 0 \tag{1}$$

Let T be an Expert in toxic comment classification that returns a score between 0-1 for both languages. Generally, for s_2 to be a valid detoxification of s_1, the toxicity rating of s_2, $T(s_2)$, should be less than $T(s_1)$. The detoxified text should not be more toxic than its original text. **Choosing Classifier T**: We use the BERT classifier from Section III and a roBERTa classifier, trained on Jigsaw's Wikipedia comments (2018) and Civil comments (2019) datasets. The roBERTa classifier achieves an AUC of 93.74.

Our analysis showed no significant differences between the classifiers: both highlighted over 70% of the same CondBert violations and 100% of the ParaDetox-BART violations. However, roBERTa missed more ParaGedi violations. Given the negligible differences and BERT's slightly better performance, we choose the BERT model for our experiments.

B. Content Preservation

The goal of MR_{SIM} is to evaluate the similarity between two comments:

$$MR_{SIM} : S(s_1, s_2) \geq \alpha \tag{2}$$

Let S be an Expert in similarity ratings that returns a score between 0-1 for both languages. For s_2 to a valid detoxification of s_1, the similarity score given for s_1 and s_2 given by S should be reasonable (i.e. $\alpha > 0.5$). Although preserving too much content risks the persistence of toxic content, simply removing toxic content inappropriately hinders content preservation. **Choosing Similarity Metric:** We compared three state-of-the-art metrics for MR_{SIM} at different α values in 0.1 increments: BLEU (bilingual evaluation understudy), METEOR (metric for evaluation of translation with explicit ordering), and LaBSE (language agnostic BERT sentence encoder). Our analysis shows that BLEU consistently underestimates the similarity of detoxified texts to the original. METEOR has a similar issue but to a lesser degree. These issues align with common pitfalls of n-gram-based metrics, which penalise translations for surface differences [14]. LaBSE performed closest to human judgment in assessing semantic similarity across all models.

C. Fluency

The goal of MR_{FL} is to evaluate the fluency difference between two comments:

$$MR_{FL} : F(s_2) - F(s_1) \geq -\beta \tag{3}$$

Let F be an Expert in fluency ratings that returns a score between 0-1 for both languages. For s_2 to a valid detoxification of s_1, the fluency of s_2 given by F should not be worse than that of s_1 given by F. Substituting away toxic content at the expense of worse grammar may be suitable at times. However, output text with abysmal grammar may seem fake and unpleasant to read.

Choosing Fluency Expert: We compared BERT and roBERTa classifiers, both fine-tuned on the CoLA corpus, for judging grammar fluency. Analysis across different β thresholds revealed that roBERTa generally identified more violations and detected more extreme grammar issues in the detoxified text at $\beta = -0.8$. Therefore, we selected the roBERTa classifier for our experiments.

D. Joint

This MR combines the conditions from MR_{ACC}, MR_{SIM}, and MR_{FL}. Each comment is evaluated against

$$MR_{JOINT} : A \wedge S \wedge F \tag{4}$$

where the statements A, S, and F are statements where the detoxified output complies with MR_{ACC}, MR_{SIM}, and MR_{FL} respectively.

VI. Experiment Design

This section aims to evaluate each MR from Section V in identifying defects in various models. Moreover, we consider the following metamorphic relations to be effective *if the violations that are highlighted precisely contain an erroneous detoxified output*.

A. Ablation Study

We determined reasonable thresholds for MR_{SIM} and MR_{FL} by experimenting with a range of α and β values to balance the detection of violations. We considered that if α is set too low, non-erroneous detoxifications might be flagged as violations, while a too-high α may risk missing some erroneous detoxifications. The same is true for β. We experimented with a scoring system where we selected the α/β values for a given model and dataset that produced a number of violations closest to the average number of violations across all α/β values tested. This approach helped us choose an α/β value roughly representing the average violations among different values.

B. Metamorphic Relation Design

The sections below define the conditions and Experts used for each metamorphic relation from Section V. To verify the effectiveness of each MR in highlighting defects, we generated sets of MR violations across different models and datasets, then analysed them in Section VII.

1) MR_{ACC}: Using the BERT-based classifier from Section V, each comment is evaluated against the condition defined in Section V-A. Moreover, if $T(s_2) - T(s_1)$ is zero or greater, s_2 is more toxic than s_1, and a violation has occurred.

2) MR_{SIM}: From Section VI-A, we found that the following condition highlighted a reasonable number of violations for poor content preservation across all test datasets and models. We use the LaBSE metric from Section V with an α value of 0.6. Thus, if $S(s_1, s_2)$ is less than 0.6, s_2 has not sufficiently preserved its original content from s_1, and a violation has occurred.

3) MR_{FL}: From Section VI-A, the following condition highlighted a reasonable number of violations for noticeable degradation of grammar across our test datasets and models. We use the roBERTa classifier from Section V and a β of 0.4. Thus, if $F(s_2) - F(s_1)$ is less than -0.4, the grammar of s_2 has degraded substantially after the detoxification of s_1, and a violation has occurred.

4) MR_{JOINT}: MR_{JOINT} follows the condition defined in (4). Thus, any ACC, SIM, or FL violation would also be a violation in MR_{JOINT}.

VII. Experimental Analysis

TABLE I
Metamorphic Relation Violations

Model	MR	Jigsaw	WikiDetox	All
CondBert	MR_{ACC}	106	2142	2248
	MR_{SIM}	571	368	939
	MR_{FL}	869	581	1450
	MR_{JOINT}	1467	2990	4457
ParaDetox-BART	MR_{ACC}	1479	1367	2846
	MR_{SIM}	969	2404	3373
	MR_{FL}	710	735	1445
	MR_{JOINT}	3046	4178	7224
ParaGedi	MR_{ACC}	82	280	362
	MR_{SIM}	1504	1839	3343
	MR_{FL}	1304	862	2166
	MR_{JOINT}	2690	2518	5208

This section demonstrates how the MRs from Section VI-B can be used on several models to highlight possible defects. Table I shows the frequency of MR violations across the three models on the two test datasets. We also categorised the errors found across all three violations, checked which errors each model made, and provided examples from our two test datasets[4] into Table II. For brevity, we labelled the models that are affected beside the error's description where CB, PD, and PG refer to CondBert, ParaDetox-BART, and ParaGedi.

A. MR_{ACC} Analysis

MR_{ACC} was used to evaluate the change in toxicity level in the detoxified comment. In general, most MR_{ACC} violations stemmed from the lack of replacement for toxic words (error 6) as exhibited a minimal difference in toxicity between the original and the output text.

[4]For more examples, visit https://github.com/alikachoo/mt_research_2024

TABLE II
MODEL ERRORS AND TEST CASE EXAMPLES

ID	Error	CB	PD	PG	Example
1	replacement of toxic words for words with similar or greater level of toxicity	✓	✓	✓	“you are an embarrassment to the rest of us canadians” → “you are an enemy even to the rest of us canadians”
2	replacement of toxic words with different meaning	✓		✓	“dirty mind” → “well-read mind”, “an idiot” → “an excellent writer”, “F*cking nerd” → “Fick ’Nerd”
3	replacement of sentence meaning	✓		✓	“:F*ck off you idiot” → “if you’re like me, do it.”
4	avoiding the toxic or difficult part by generalising		✓	✓	“You’re an idiot. Nuff said.” → “that’s what I meant.”, “and apparently, youre a pile of horsesh*te” → “You are not good”
5	replacement of toxic words degrading overall grammar	✓	✓	✓	“you must really suck at life” → “you must really bad at life” “your country is sh*t poor and ugly” → “Your country’s poor and unhappily”
6	lack of replacements	✓	✓	✓	“GOOD RIDDANCE. GET LOST.” → “good rid . get lost .”

CondBert: MR violations were primarily due to the failure to replace toxic words (error 6). In the Jigsaw and WikiDetox datasets, 71 out of 106 and 1718 out of 2142 violations had zero toxicity difference, respectively. The model primarily made stylistic changes, such as modifying capitalisations and adding/removing spacing and punctuation, which did not sufficiently detoxify the content. The remaining errors were simply due to replacements of words (error 1) that were just as toxic or more toxic.

ParaDetox-BART: Error 6 here was also caused by the ParaDetox-BART only making mostly stylistic changes, which made minimal difference to the toxicity in the detoxified text. In other cases, ParaDetox-BART may also remove fragments of toxic content without replacing its entirety (e.g., "you die you die and you go to hell" → "you die you go to hell").

ParaGedi: Similar to the previous two models, ParaGedi failed to remove abusive language and hateful speech (racism, sexism, misogyny) in the output. Most text outputs showed the same types of stylistic changes as mentioned earlier and minor paraphrasing changes such as use of contractions (e.g., "are not" → "aren’t").

Upon inspection of the violations that MR_{ACC} highlights, the detoxified outputs were erroneous. Additionally, we found that MR_{ACC} helps set expectations for each model’s attempt to remove toxicity content.

B. MR_{SIM} Analysis

MR_{SIM} aims to quantify content preservation from the original text to the detoxified output. Most violations arose from replacement errors where the modified text diverged significantly from the original meaning.

CondBert: With regards to error 2, CondBert erroneously replaced toxic content for antonyms (e.g., "idiot" → "an even better friend"). As for error 3, CondBert only rewrote fragments of a sentence, changing the overall sentence meaning. Another case relates to how CondBert replaces toxic text with different punctuations when encountering sentences that are largely infested with abusive language. However, these behaviours may sometimes be appropriate, given that they often remove offensive words from the original text.

ParaDetox-BART: Most replacement errors were caused by avoiding toxic parts by generalisation (error 4). This behaviour often resulted in multiple distinct toxic texts being transformed into the same detoxified output, such as "go to hell with that idiot" and "so f*ck the f*ck off" both being reduced to "Go away.". While such generalisations can sometimes suffice in removing toxic content, they tend to strip away the specific target and content, undermining the comprehension of the original text’s meaning.

ParaGedi: MR_{SIM} violations highlighted errors that overlap with a quantitative analysis Dale et al. [4] conducted (errors 2, 3, and 4). This can be attributed to ParaGedi’s tendency to paraphrase the whole sentence, add punctuation or spacing, modify capitalisations, and replace fragments of toxic content.

We inspected the violations that MR_{SIM} highlighted and found that most precisely failed to preserve content sufficiently in the output. Nonetheless, these models prioritise removing toxicity at the expense of worse content preservation.

C. MR_{FL} Analysis

MR_{FL} assesses whether a degradation of fluency has occurred during the detoxification process. MR_{FL} violations were predominantly due to errors 5.

CondBert: MR_{FL} violations in this case were primarily due to replacement errors (error 5) which commonly involved inappropriately replacing proper nouns with abstract nouns (e.g., lesbians → lesbian, unfairly → unfair). Similar to some CondBert MR_{SIM} violations, this model may replace toxic text with punctuation and degrade fluency in the output.

ParaDetox-BART: Most MR_{FL} violations here stemmed from error 5, which typically removed toxic content by degrading grammar. This commonly involved ParaDetox-BART neglecting to put articles (e.g.,‘a’, ‘an’, ‘the’) in the output text to make it more fluent, even though the source input had high fluency (e.g., “This guy is jew” and “You are robot”).

ParaGedi: ParaGedi exhibited the highest number of MR_{FL} violations across the two test sets due to poor paraphrasing that degraded grammatical fluency (error 5). This model often misused articles (e.g., “successful in the life”) and inappropriately substituted noun types (e.g. "loser" → "lost"). Another common erroneous behaviour that degrades sentence fluency is the inappropriate use of apostrophe s’s (i.e. ’s) and other contractions.

Upon analysing the violations that MR_{FL} highlights, we found that most outputs from the test cases precisely showed noticeable fluency degradation. Furthermore, these models

may prioritise removing toxic content despite worsening fluency.

D. MR_{JOINT} Analysis

MR_{JOINT} violations consist of the violations from the other three MRs with no duplications. Although MR_{JOINT} does not provide new findings, it is still helpful in getting a consensus of how well a given text detoxification performed in transferring style on a test set.

VIII. Discussion

Our findings suggest utility in evaluating source inputs that led to our MR violations (erroneous outputs), as was the focus in the study by Yan et al. [1]. An extensive qualitative analysis comparing how tools match against human evaluators would have been more compelling. Finally, given the subjectivity of human judgement, MR violations in the previous section may not precisely contain erroneous outputs to specific individuals. Thus, expanding analyses across a greater diversity of communities would be a priority in future works.

IX. Conclusion

The proposed metamorphic relations effectively identified erroneous output and defective behaviour across three models on different datasets. Moreover, we found that the MRs helped us gain a concrete understanding of each model's behaviour and limitations toward text detoxification without detailed knowledge of its underlying architecture. This capability can enable developers to build and improve on the robustness of text detoxification systems, contributing to a safer world for everyone.

Acknowledgment

This work was supported by JC School of Medicine & Health Sciences, Monash University Malaysia Seed Grant (T & L) (Grant No: I-M010-STG-000192 and I-M010-SED-000229).

References

[1] Yan et al., "Metamorphic relations for data validation: A case study of translated text messages," 2019. DOI: 10.1109/met.2019.00018.

[2] P. Fortuna, J. Soler, and L. Wanner, "Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets," English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Calzolari et al., Ed., Marseille, France: European Language Resources Association, May 2020, pp. 6786–6794, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.838.

[3] Floto et al., "Diffudetox: A mixed diffusion model for text detoxification," in *Findings of the Association for Computational Linguistics: ACL-23*, Association for Computational Linguistics, 2023, pp. 7566–7574.

[4] Dale et al., "Text detoxification using large pre-trained neural models," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021. DOI: 10.18653/v1/2021.emnlp-main.629.

[5] L. Sun and Z. Q. Zhou, "Metamorphic testing for machine translations: Mt4mt," in *2018 25th Australasian Software Engineering Conference (ASWEC)*, 2018, pp. 96–100. DOI: 10.1109/ASWEC.2018.00021.

[6] Logacheva et al., "ParaDetox: Detoxification with parallel data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6804–6818. [Online]. Available: https://aclanthology.org/2022.acl-long.469.

[7] K. Atwell, S. Hassan, and M. Alikhani, "Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations," *ArXiv*, vol. abs/2209.08207, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252368208.

[8] I. Smirnov, "The digital flynn effect: Complexity of posts on social media increases over time," in *Social Informatics*. Springer International Publishing, 2017, pp. 24–30, ISBN: 9783319672564. DOI: 10.1007/978-3-319-67256-4_3.

[9] Barr et al., "The oracle problem in software testing: A survey," *IEEE Transactions on Software Engineering*, vol. 41, no. 5, pp. 507–525, 2015. DOI: 10.1109/TSE.2014.2372785.

[10] Chen et al., "Metamorphic testing: A review of challenges and opportunities," *ACM Comput. Surv.*, vol. 51, no. 1, Jan. 2018, ISSN: 0360-0300. DOI: 10.1145/3143561.

[11] Altamimi et al., "Metamorphic relation automation: Rationale, challenges, and solution directions," *Journal of Software: Evolution and Process*, vol. 35, no. 1, e2509, 2023. DOI: https://doi.org/10.1002/smr.2509.

[12] Krishna et al., "Reformulating unsupervised style transfer as paraphrase generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 737–762. DOI: 10.18653/v1/2020.emnlp-main.55.

[13] Pesaranghader et al., "Gpt-detox: An in-context learning-based paraphraser for text detoxification," in *IEE 2023 International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2023. DOI: 10.1109/icmla58977.2023.00230.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *Bertscore: Evaluating text generation with bert*, 2020. arXiv: 1904.09675 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1904.09675.