Validating Multimedia Content Moderation Software via Semantic Fusion

Wenxuan Wang
The Chinese University of Hong Kong
Hong Kong, China
wwxwang@cse.cuhk.edu.hk

Jingyuan Huang
The Chinese University of Hong Kong
Hong Kong, China
1155173905@link.cuhk.edu.hk

Chang Chen
The Chinese University of Hong Kong
Hong Kong, China
ccchen@link.cuhk.edu.hk

Jiazhen Gu
The Chinese University of Hong Kong
Hong Kong, China
jiazhengu@cuhk.edu.hk

Jianping Zhang
The Chinese University of Hong Kong
Hong Kong, China
jpzhang@cse.cuhk.edu.hk

Weibin Wu
Sun Yat-sen University
Zhuhai, China
wuwb36@mail.sysu.edu.cn

Pinjia He∗
The Chinese University of Hong Kong
Shenzhen, China
hepinjia@cuhk.edu.cn

Michael Lyu
The Chinese University of Hong Kong
Hong Kong, China
lyu@cse.cuhk.edu.hk

ABSTRACT
The exponential growth of social media platforms, such as Facebook, Instagram, Youtube, and TikTok, has revolutionized communication and content publication in human society. Users on these platforms can publish multimedia content that delivers information via the combination of text, audio, images, and video. Meanwhile, the multimedia content release facility has been increasingly exploited to propagate toxic content, such as hate speech, malicious advertisement, and pornography. To this end, content moderation software has been widely deployed on these platforms to detect and blocks toxic content. However, due to the complexity of content moderation models and the difficulty of understanding information across multiple modalities, existing content moderation software can fail to detect toxic content, which often leads to extremely negative impacts (e.g., harmful effects on teen mental health).

We introduce Semantic Fusion, a general, effective methodology for validating multimedia content moderation software. Our key idea is to fuse two or more existing single-modal inputs (e.g., a textual sentence and an image) into a new input that combines the semantics of its ancestors in a novel manner and has toxic nature by construction. This fused input is then used for validating multimedia content moderation software. We realized Semantic Fusion as DUO, a practical content moderation software testing tool. In our evaluation, we employ DUO to test five commercial content moderation software and two state-of-the-art models against three kinds of toxic contents. The results show that DUO achieves up to 100% error finding rate (EFR) when testing moderation software and it obtains up to 94.1% EFR when testing the state-of-the-art models. In addition, we leverage the test cases generated by DUO to retrain the two models we explored, which largely improves model robustness (2.5%–5.7% EFR) while maintaining the accuracy on the original test set.

CCS CONCEPTS
• Software and its engineering → Software testing and debugging.

KEYWORDS
Software testing, semantic fusion, multimedia content moderation, metamorphic testing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION
Multimedia contents, such as Internet memes and videos, play an important role in online communication and content publication on social media platforms. For example, in 2020, there were more than one million posts mentioning "meme" being shared on Instagram, one of the most popular social media platforms [3], every day. Moreover, Cisco reports that 82% of global Internet traffic will come from either video streaming or video downloads in 2022 [55]. Fig. 1 presents two examples of such multimedia content from the web.12. Although the exponential growth of multimedia content has
greatly facilitate user communication and content distribution in the world, it has also exacerbated the propagation of toxic content. Toxic contents generally refer to harmful contents that can cause negative affect on reader’s attitudes, behavior or health. In particular, toxic contents are mainly including but not limited to into the following categories: (1) abusive language and hate speech, which are abusive contents targeting specific individuals, such as politicians, celebrities, religions, nations, and the LGBTIQ+; (2) malicious advertisements, which are online advertisements with illegal purposes, including phishing and scam links, malware download, and illegal information dissemination; and (3) pornography, which is often sexually explicit, associative, and arousing.

Such toxic contents have significant negative impacts on users. For instance, Munro [43] concluded that online hate speech may develop depression, anxiety, and other mental health problems in children. Malicious advertisements for illegal purposes also remain a global burden, accounting for up to 85% of daily message traffic [8]. Pornography may cause significant undesirable effects on both the physical and psychological health of children [76]. Statistics showed that adult content sites accounted for 0.67% of all website categories accessed by Latin American children from May 2019 to May 2020 [9]. Moreover, such widely disseminated toxic contents greatly affect social harmony and increase the number of criminal cases to a certain extent [6].

Due to the harmfulness of toxic content, content moderation software for detecting and blocking toxic content has attracted massive interest from both academia and industry. Existing methods typically formulate toxic content detection as a classification task and resort to Artificial Intelligence (AI) techniques, such as convolutional neural networks (CNN), long-short-term-memory models (LSTM), and Transformer-based models [65], and achieving considerable performance on corresponding datasets [41, 53, 73]. Because of the importance of content moderation, large-scale online service providers, such as Google [17], Meta (Facebook) [66], Twitter [14], and Baidu [26], have extensively deployed commercial-level content moderation software on their products. In particular, Meta reports that they remove millions of violating content on Facebook and Instagram every day, among which more than 90% are detected by AI-based moderation software.

Although tremendous efforts have been spent on developing toxic content moderation models, existing content moderation software sometimes fails to detect inputs from malicious users, exposing toxic content to other users. For example, a New Zealand terrorist live-streamed a massacre on Facebook [57], YouTube Kids, an app for children, was reported to contain a significant amount of inappropriate content, which was made readily available for unsuspecting kids [33]. Due to the huge number of Internet users, even a 0.01% failure rate may cause serious consequences.

Despite its apparent importance, validating the robustness of multimedia content moderation software is very difficult and has, therefore, been much under-explored. First, existing high-quality multimedia data have already been utilized in the development of the moderation software and models, while the construction of a new test case typically incurs extensive manual labeling effort. Second, previous studies mainly generate test cases from only one specific modality, such as visual modality [77], audio modality [22], and textual modality [47, 48]. While these approaches can generate interesting test cases, they fail to stress-test the core ability of multimedia content moderation software: understanding multimodal inputs.

Inspired by an SMT solver validation method [71] that fuses two existing formulas into a new formula, this paper introduces Semantic Fusion, a general, effective methodology for validating multimedia content moderation software. Our key insight is to fuse two or more single-modal inputs into a new multi-modal input that combines the semantics of its ancestors and is toxic by construction. Fig. 2 presents the high-level idea of Semantic Fusion via two sketches of the test cases. To realize this concept, we implement DUO, a tool that can generate test cases that cover all three typical categories of toxic contents (i.e., hate speech, malicious advertisement, and pornography) in two widely-used languages (i.e., English and Chinese). Specifically, DUO first adopts a template-based approach to construct seed toxic sentences. Then DUO generates multi-modal toxic contents as test cases by fusing single-modal inputs accordingly. These toxic test cases will be fed into the multimedia content moderation software as test cases. If a test case evades the detection of the software under test, an error will be reported.

To evaluate the performance of DUO, we apply DUO to test five widely-deployed commercial content moderation software from famous software providers, including Google Cloud, Amazon Web Service, Baidu Cloud, Tencent Cloud, and Alibaba Cloud, and two state-of-the-art models (i.e., Vision-Transformer-based [12] and ResNet-based models [18]). The results show that the software under test fails to detect most of the test cases generated by DUO. Notably, up to 100% and 94.1% of generated toxic image and video test cases can bypass the content moderation software and research models, respectively. In addition, we leverage the test cases generated by DUO to retrain the model we explored, which largely improves

---

These are monkeys

Figure 2: The sketches of two multimedia contents generated by Semantic Fusion. Both contents are toxic (hate speech) by construction if: (1) “xxx” was replaced by the name of a specific group (e.g., racial, gender, religion, or nation) for the left content; and (2) the people were replaced by a photo of a specific group for the right content.

model robustness (2.5%~5.7% EFR) while maintaining the accuracy on the original test set.

The contributions of this paper are summarized as follows:

- We introduce Semantic Fusion, a general, effective methodology for testing multimedia content moderation software.
- Based on the Semantic Fusion methodology, we design and develop the first tool, DUO, for multimedia content moderation software validation.
- DUO effectively reported errors in five widely-deployed commercial software products and two state-of-the-art research models with consistently high error finding rates.
- We successfully improved the robustness of the two moderation models by retraining them with the failed test cases.

Content Warning: We apologize that this article presents examples of aggressive, abusive, and pornographic expressions to demonstrate the results of our method.

2 BACKGROUND

2.1 Content Moderation Software

Big companies, for example, Google, Meta, Twitter, Amazon, Baidu, Tencent and Alibaba, have developed and deployed commercial-level content moderation software on their products. According to their official technical documents, the backbone of their software is usually a complicated engineering system containing neural network-based models and rule-based methods. For example, Baidu Commercial Content Moderation Software is powered by a deep neural network and a huge pre-defined banned word list. This kind of hybrid approach can leverage the best of different methods. Neural network-based methods can effectively mine contexts and semantic information, while rule-based methods easily implement user-defined functionality.

2.2 Multi-modal Multimedia Content

Multimedia is a form of communication that uses a combination of different content forms such as text, audio, images, animations, or video into a single presentation1. A model that can deal with multimedia data is called multi-modal AI and each channel is called a modality. There are two main issues in multi-modal AI, processing single modality and understanding information across different modalities. Our work mainly involves three modalities: visual modality, audio modality, and textual modality. For example, a meme contains visual (the image) and textual (the top text and bottom text) information, and a video usually involves visual modality (the video screen), audio modality (the soundtrack), and textual modality (the subtitles).

2.3 Multi-modal Fusion

Multimedia content (e.g., a meme or video) has different modalities to convey information. Therefore, to understand the whole picture of multimedia content and determine its toxicity, one needs not only to process the information in every single modality but also to fuse the information from different modalities. The fusion of different modalities is generally performed at two levels: feature level and decision level. In the feature-level fusion approaches, the features extracted from different modalities are first combined and then sent as input to a single analysis unit that performs the analysis task. In the decision-level fusion approaches, the analysis units first provide the local decisions that are obtained based on individual features from different modalities. The local decisions are then combined using a decision fusion unit to make a fused decision. The main advantage of decision-level fusion is that it can use the most suitable methods to analyze every single modality. However, it fails to utilize the feature-level correlation among modalities.

3 APPROACH AND IMPLEMENTATION

In this section, we introduce the design and implementation of DUO, a novel tool to validate content moderation software. Figure 3 overviews the workflow of DUO, which consists of two main modules, test case generation and error detection. In particular, the test case generation module adopts a semantic fusion-based method to generate toxic multimedia contents as test cases. Then, the generated cases are fed into the error detection module, which performs metamorphic testing to reveal errors in content moderation software.

More specifically, to generate toxic multimedia content, DUO first adopts a template-based method to extract keyword pairs from existing toxic datasets. Each keyword pair can constitute toxic sentences, which are further used as the seeds to generate toxic multimedia contents (Section 3.1). For each extracted keyword pair, DUO distributes the keyword pairs into different modalities (Section 3.2), and then fuses the multi-modal information to generate multimedia contents (Section 3.3). Finally, DUO feeds the generated cases into the content moderation software under test, and detects errors based on metamorphic relations, i.e., the content toxicity is invariant under modulation transformation (Section 3.4).

3.1 Keyword Pair Extraction

To generate test cases, DUO first extracts keyword pairs, which can constitute toxic contents, from existing datasets. Specifically, DUO adopts a template-based approach: It first extracts keyword pairs that can generate toxic sentences for different toxic categories based on templates. Then, an NLP-based filtering method is adopted

---

1https://en.wikipedia.org/wiki/Multimedia
1. Keyword Pair Extraction

Template: A are B

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>Women</td>
</tr>
<tr>
<td>White People</td>
<td>Black People</td>
</tr>
<tr>
<td>Shit</td>
<td>Bugs</td>
</tr>
<tr>
<td>Monkey</td>
<td>Terrorist</td>
</tr>
</tbody>
</table>

2. Modality Transformation

Google Images & Detection Filter

3. Multi-modal Fusion

Text Resizing & Location Determination & Font Color Selection

4. Toxicity Collection

Content Moderation Software Under Test

5. Error Detection

Our MRs: Error Detected

Table 1: Statistics of toxic datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Sent</th>
<th>Lang</th>
<th>Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Bias</td>
<td>150k</td>
<td>English</td>
<td>Abuse</td>
<td>Twitter</td>
</tr>
<tr>
<td>Dirty</td>
<td>2.5K</td>
<td>Chinese</td>
<td>Abuse</td>
<td>Weibo</td>
</tr>
<tr>
<td>SMSSpam</td>
<td>5.5K</td>
<td>English</td>
<td>Ad.</td>
<td>Grumbletext</td>
</tr>
<tr>
<td>SpamMessage</td>
<td>60K</td>
<td>Chinese</td>
<td>Ad.</td>
<td>Taobao</td>
</tr>
<tr>
<td>Sexting</td>
<td>0.5K</td>
<td>English</td>
<td>Porno</td>
<td>Github</td>
</tr>
<tr>
<td>Midu</td>
<td>7.3K</td>
<td>Chinese</td>
<td>Porno</td>
<td>Midu</td>
</tr>
</tbody>
</table>

Figure 3: The overview of DUO.

to drop invalid keyword pairs, so as to avoid the generation of low-quality cases.

3.1.1 Template Designing. DUO utilizes the following templates to extract keyword pairs for each kind of toxicity, i.e., hate, advertisement and pornography.

For abuse and hate speech, DUO utilizes the template "A is/are B", where A is a group, e.g., a specific race or gender, and B is a negative adjective, such as “stupid” or “lazy”, or noun, such as criminal or pig.

For malicious advertisement, we design a template of "A: B", where A is a product, such as “Tobacco” or “Alcohol” and B is the contact information, such as telephone number, email address or WhatsApp number.

For pornography, the template "A your/my B" is adopted, where A is a verb and B is a sex-related organ.

Even though DUO only adopts one simple template for each type of toxicity, it can detect a large number of errors in practical content moderation software (Section 4). Since DUO can achieve satisfying performance even with simple templates, we do not investigate the effectiveness of more complicated templates, which can be studied in future work.

3.1.2 Keyword Selection. Based on the above templates for different kinds of toxicity, DUO extracts keyword pairs that can constitute corresponding toxic texts from existing toxic datasets. In order to find keywords similar to real-world cases, DUO utilizes a total of 6 manually labeled datasets collected from practical Internet platforms, 2 for each toxicity type. The statistics of the datasets are shown in Table 1.

Specifically, for abuse and hate speech, DUO extracts keyword pairs from Social Bias Corpus [52] and Dirty [5]. Social Bias Corpus contains 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups. Dirty is an open GitHub repository containing 2.5k Chinese toxic sentences with abusive and sexual words.

For malicious advertisement, DUO extracts keyword pairs from SMS Spam Collection [6] and SpamMessage [7]. SMS Spam Collection is a set of tagged SMS messages, containing 5,374 SMS messages in English, tagged as being ham (legitimate) or spam. The data was manually extracted from the Grumbletext website, a UK forum in which cell phone users make public claims about SMS spam messages, while SpamMessage is an open GitHub repository containing 60k malicious advertisement messages.

For pornography, DUO utilizes Sexting [8] and Midu [58], where Sexting is an English pornographic text dataset containing 537 sexual texting messages, while Midu is a Chinese novel paragraph dataset collected by ourselves from an online literature reading platform called MiDu App [9]. It is a corpus with 62,876 paragraphs including 7,360 pornographic paragraphs and 55,516 normal paragraphs.

It is worth noting that not all words in the above datasets are potential keywords. An ideal keyword should be frequently used in toxic content while less frequently in a general domain corpus so that it is more likely to contain toxicity. Therefore, we use TF-IDF, a numerical statistic that reflects how important a word is to a document in a collection or corpus, to select potential keywords from the above datasets. The TF–IDF value increases proportionally to the number of times a word appears in the document and is offset from the number of documents in the corpus that contain the word. Particularly, DUO utilizes sklearn [10] (English) and Jieba library [11] (Chinese) to filter out the stop words, followed by calculating the TF-IDF score, and select the top 20 words with the highest TF-IDF score as the candidate keywords for each dataset.

3.1.3 Keyword Pair Extraction and Filtering. After obtaining the candidate keywords, it is non-trivial to extract keyword pairs that

\[ \text{https://github.com/pokemonchw/Dirty} \]
\[ \text{https://www.kaggle.com/uciml/sms-spam-collection-dataset} \]
\[ \text{https://github.com/hrwhisper/SpamMessage} \]
can constitute toxic texts according to the templates, since not all keyword pairs are suitable. For example, for the template of pornographic (i.e., “A you/my B”), A should be a verb and B should be a sexual-related organ. It is inappropriate to fill a verb keyword into slot B, leading to meaningless sentences.

In order to find proper keyword pairs suitable for different templates, DUO has to first obtain the property of every keyword. Considering that a keyword may have multiple properties, for each keyword, DUO first retrieves 5 sentences containing the keyword from the corresponding dataset. Then, it utilizes the language analysis method to perform Part-of-Speech tagging (PoS tagging), which identifies the word property (e.g., noun, verb, adjective or adverb), and Named Entity Recognition (NER), which determines whether the keyword is a belongs to a pre-defined category such as group names or location. The results of each keyword are voted on by the results of the five sentences. In the implementation, DUO adopts Flair toolkit12 for English and Baidu NLP API13 for Chinese.

After obtaining the keyword property, DUO extracts keyword pairs for three types of toxic contents as follows:

For abuse and hate speech, the template is “A is/are B”, where A is a group and B is a negative noun or adjective. Hence, for A DUO selects the keywords being identified as nouns by PoS tagging and group names by NER. And for B, DUO selects the keyword which is either a noun or adjective and also has a negative sentiment, which is obtained from the APIs provided by Google14 (for English) and Baidu15 (for Chinese). In addition, we require A and B to be different words, aiming to prevent the “monkey is monkey” situation that generates non-toxic seed sentences.

For malicious advertisement, the template is “A your/my B”, where A is a product and B is the contact information. Hence, DUO selects the keyword identified as the noun for A, according to the PoS tagging toolkit. For B, DUO extracts the keywords with a prefix of a kind of contact way. We collect a candidate list, including “Tel”, “Email”, “WhatsApp” and “Ins”. In this way, DUO can extract keyword pairs that can constitute the prefix contact way followed by the specific contact information, such as “Tel: 12345678”.

For pornography, the template is “A your/my B”, where A is a verb and B is a sexual-related organ. Hence, DUO extracts the keyword identified as verb for A and noun for B, according to the PoS tagging toolkit.

In this way, DUO can extract keyword pairs as seeds for the further test case (i.e., multimedia content) generation. Particularly, DUO generates 100 keyword pairs for each type of toxicity and each language, ending up with 100*3*2=600 keyword pairs.

3.2 Modality Transformation

Since the extracted keyword pairs only contain single modality information (i.e., textual), in order to generate multi-modal contents, DUO has to perform modality transformation. Particularly, this work focuses on three kinds of modalities, i.e., visual, textual, and audio. For each extracted keyword pair (A, B), DUO transforms the information of A and B to different modalities as follows:

Visual modality. For the keyword pair (A, B), DUO transforms the textual information to visual modality via calling the Google figure API16 to search the top 5 images with A or B as the query. To ensure that the returned figures correctly contain the information of the queried keyword, DUO further utilizes Baidu Image Recognition API16 to check the recognized salient object in each figure. If the object is not equal to the keyword, DUO discards the figure.

Textual modality. Since the keywords are texts per se, DUO directly uses the templates for keyword extraction (Section 3.1) to obtain the textual information.

Audio modality. In order to transform the textual information to audio modality data (i.e., speech), DUO calls the Baidu text-to-speech synthesis API17.

3.3 Multi-modal Semantic Fusion

After obtaining the information of keyword pairs from different modalities, DUO performs semantic fusion, which fuses the multi-modal information together and generates test cases (i.e., multimedia contents). In particular, DUO generates two kinds of multimedia data, i.e., images and videos. The images are generated by fusing visual and textual information, while the videos can be constituted by any combination of the three modalities, i.e., visual & textual, visual & audio, audio & text, and visual & textual & audio. Remembering that, besides the (A, B) keyword pair, there is a middle word in our template that represents the logical connection between A and B. For example, in abuse and hate speech, there is an “are” between A and B. We add this middle word to text or audio modality.

3.3.1 Image Generation. Image generation is performed via fusing visual and textual information (i.e., image and text), more specifically, inserting the text into the image. In this process, we need to address three main issues: (1) how to decide the size of the text to avoid being too big or too small; (2) how to decide the location for the insertion, which should not affect the image content; (3) how to decide the color of the text so that it can be recognized in the image. In the following, we introduce three algorithms adopted by DUO to address these issues, respectively.

Text resizing. The goal of text resizing is to resize the text object to make the inserted text a comparable size with the salient object in the image. To achieve this target, DUO first utilizes Baidu Image Recognition API to detect the salient object in the image and obtains the coordinates of the four vertices of its bounding box (x1, y1), (x2, y2), (x2, y1), (x1, y2). Then, the height h and the width w of the salient object can be approximated as w = x2 - x1 and h = y2 - y1, respectively. Finally, DUO sets the area of text within the range of [0.8 * w * h, 1.2 * w * h], where 0.8 and 1.2 are hyper-parameters that we manually set based on empirical experiences.

Location determination. After resizing the text, DUO should select a suitable location in the image to insert the text. An ideal location should (1) have few overlaps with the salient object in the image, which is beneficial for humans to recognize both the text and the image; (2) the relative positions of the image and text follow the reading habits which is easier for human to understand the logic relation between the image information and the text information.

12https://github.com/flairNLP/flair
13https://ai.baidu.com/tech/nlp_api
14https://ai.baidu.com/tech/nlp_basic
15https://cloud.google.com/natural-language
16https://ai.baidu.com/ai-doc/IMAGERECOGNITION/Xk3bxdum
17https://ai.baidu.com/ai-doc/SPEECH/jk3bygno
To find the location with few overlaps with the salient object, we first define 9 candidate insert positions: top left, top middle, top right, middle left, middle middle, middle right, bottom left, bottom middle and bottom right. Since DUO has obtained the coordinates of the four vertices of the salient object in text resizing, it can directly calculate the overlapping area between the salient object and each candidate position. If the overlapping area is larger than 30%, DUO discards this candidate position.

To find the position in line with human reading habits, DUO utilizes a rule-based method based on the consideration that humans typically read from top to bottom and from left to right. It is worth noting that the keyword pair (A, B) is extracted according to the templates (i.e., “A is/are B”, “A: B”, and “A your/my B”). In all the used templates, a human reads A first and reads B later. Hence, object A should be above or to the left of object B, such constraints help DUO filter out some candidate positions. For example, suppose we need to insert text B into image A, the candidate positions of B following human reading habits are middle right, bottom middle and bottom right. For each (image, text) pair, if there is no ideal location candidate left based on the criteria above, DUO discards this pair. If more than one candidates are suitable, DUO randomly selects one as the location to insert the text.

**Font color selection.** After determining a location to insert the text, another issue DUO needs to decide is the font color. If the font color is too similar to the background image color, the text will be hard for humans to recognize, leading to an invalid image. To mitigate this issue, DUO adopts a special webkit property called Text Stroke, which adds an exterior border around each character of the text. Text stroke can change the outline of the text, such as setting a color different from the original font color, so that the text can be recognized easily whatever the background color.

3.3.2 **Video Generation.** Unlike generating image data, video data can be generated by fusing any two or more information in different modalities. DUO conducts semantic fusion on different modality data as follows:

**Fusing vision and text.** Generating video test cases that fuse vision ad text is similar to image generation, except that we should take the order of showing the image and the text into consideration. Similar to the position selection when generating images, for a keyword pair (A, B), A should come first and B should come later. Hence DUO generates the video that presents A first and then presents B. For example, if A is n image and B is the text. The generated video shows A first for a while and then shows the text of B. On the other hand, if A is the text and B is an image, the video would show the text of A and then show the image B. In the implementation, DUO adopts ffmpeg\(^{18}\), a complete, cross-platform API to record, convert and stream audio and video.

**Fusing vision and audio.** DUO generates the video based on vision and audio by showing the image and playing the audio (a synthesized speech generated from the text). To make the video easier for humans to understand, again, we consider the order of showing the image and playing the audio. Specifically, for a keyword pair (A, B), the generated video shows A first and then shows B. For example, if A is an image and B is the audio. We show A first for a while and then play the audio of B. On the other hand, if A is the audio and B is an image, the video first plays the audio of A and meanwhile shows a blank video screen. The image B will not be displayed until audio A has finished playing.

**Fusing text and audio.** Generating the video test cases with audio and text information is similar to generating the video that fuses image and audio. The main difference is that here DUO shows the text, rather than the image, and plays audio. Again, if A is audio and B is text, the generated video plays the audio first and then shows the text, and vice versa. Since this kind of fusion contains both text and audio modality, we randomly add the middle word to either of the modality.

Finally, it is also feasible to generate videos by fusing all three modalities. To achieve this, DUO adopts the procedure of fusing image and audio. The only difference is that here DUO shows the middle words in text format between showing A and showing B, rather than showing accompanied with the keyword in audio modality.

As such, DUO can generate multi-modal images and videos, which are used as test cases to detect errors in content moderation software.

3.4 **Toxicity Collection and Error Detection**

After modality transformation and semantic fusion, DUO constructs test cases (i.e., multi-modal images and videos). Each test case has a corresponding keyword pair (A, B), which is extracted based on a specific template. By filling the keyword pair into the template, we can get a seed sentence, which is further used for solving the test oracle problem. During testing, DUO first feeds the generated test cases, as well as the corresponding seed sentences, to the content moderation software under test. Since all seed sentences are supposed to be toxic, any test case whose corresponding seed sentence is classified as non-toxic by the textual content moderation software will be discarded.

After toxicity collection, DUO inspects the predicted toxicity of each test case. Any case that violates the metamorphic relation (MR) will be reported as a suspicious error. The MR is designed based on the following simple consideration: the toxicity should remain the same when some of the information is transformed into another modality. Since DUO assumes seed sentences are all toxic ones, all the remaining cases should be classified as toxic, i.e., an error is found if the test case is categorized as non-toxic.

In the implementation, we test 5 commercial software products provided by large Internet companies, i.e., Google Cloud\(^ {20}\), Amazon Web Service\(^ {21}\), Baidu Cloud\(^ {21}\), Tencent Cloud\(^ {22}\) and Alibaba Cloud\(^ {25}\), all of which are the official content moderation software from big technology companies with more than 100 millions of users. In particular, all the software products are the latest version by Nov. 1st, 2022, when the experiments were conducted. The version information of the software under test is listed in Table 2. Besides commercial software products, we also test popular research models. The core of multimedia content moderation software is

---

18https://ffmpeg.org/
20https://cloud.google.com/video-intelligence/docs/analyze-safesearch
21https://docs.aws.amazon.com/rekognition/latest/dg/moderation.html
22https://cloud.baidu.com/doc/ANTIPORN/s/6ki012lqu
23https://cloud.tencent.com/document/product/1235
4 EVALUATION

To validate the effectiveness of DUO and get more insights on enhancing content moderation software, we use our method to test five commercial software products and two state-of-the-art research models for multimedia content moderation. In this section, we detail the evaluation process and empirically explore the following four research questions (RQs).

- RQ1: Are the test cases generated by DUO toxic and realistic?
- RQ2: Can DUO find errors in content moderation software?
- RQ3: What factors affect the performance of DUO?
- RQ4: Can we use DUO to improve the performance of multimedia content moderation?

4.1 RQ1: Are the test cases generated by DUO toxic and realistic?

DUO aims to generate test cases that are toxic and are as realistic as the ones real-world users produce to evade moderation. Thus, in this section, we evaluate whether the generated test cases are still toxic (i.e., semantic-preserving) and whether they are realistic.

We conduct human annotation via crowdsourcing. First, we generate 10 images and 30 videos (10 vision + text, 10 vision + audio, 10 audio + text) for each task and each language, ending up with 240 test cases for annotation. For each test case, we ask three questions: 1) From "I strongly disagree" to "5 strongly agree", to what extent do you agree that the image/video is semantically equivalent to the sentence? 2) From "I strongly disagree" to "5 strongly agree", to what extent do you agree that the image/video is toxic (hate speech, malicious advertisement, or pornography, according to the dataset)? 3) From "I strongly disagree" to "5 strongly agree", to what extent do you agree that this kind of image/video is realistic that Internet users would use? For English/Chinese, we distribute the questionnaire and recruit 20/20 crowd workers on Proflc/Tencent Wenjuan, who have English/Chinese as their first language. Before annotation, we provide instructions about the type of questions and asked them to make subjective judgments in the annotation. We do not provide additional training to avoid potential bias from us. Then the annotators are asked to annotate the test cases. Annotation results show that: 1) the generated test cases are semantically equivalent to the seed sentence, with an average score of 4.46/4.59; 2) the generated test cases are toxic, with an average score of 4.19/4.29; 3) the generated test cases are realistic, with the average score of 3.96/4.15. We followed [32] to measure the inter-annotator agreement using Randolph’s Kappa, obtaining a value of 0.84/0.81 for the test cases, which indicates “almost perfect agreement”. There are a few cases that the generated images are annotated as non-toxic subjectively. For example, sometimes Google Image API returns a lovely cartoon image which could make the annotator feel less offensive.

Answer to RQ1: Test cases generated by DUO are toxic and commonly seen in real-world scenarios.

4.2 RQ2: Can DUO find errors in content moderation software?

DUO aims to automatically generate test cases to find potential bugs in current content moderation software. Hence, in this section, we evaluate the number of errors that DUO can find in the outputs of commercial content moderation software and academic models.

4.2.1 Software Products and Models under Test. We use DUO to test five commercial content moderation software products and two state-of-the-art research models.

Image Content Moderation. For commercial software products, we choose the products that have been deployed in the cloud of large Internet companies, including Google, Amazon, Baidu, Tencent, and Alibaba. All of them can be accessed by registered users via an API. In particular, since Amazon Web Service does not provide advertisement detection services, Google Cloud does not provide hateful image detection and advertisement detection services, and Tencent Cloud does not provide hateful and abusive image detection services, we do not include them in our experiments. For research models, we choose the Vision-Transformer Model and ResNet-18 Model, which are the most downloaded image classification models in Hugging Face, and model zoo. For hate speech detection, we use the Hateful Memes Dataset, a hate and abuse detection dataset containing 8,500 multimodal memes with labeled toxicity. For malicious advertisement detection, we use the Advertisement Understanding Dataset, an advertisement dataset containing 64,832 Ad and 13,597 non-Ad images. Since there is no publicly available pornographic image detection dataset, we do not include this in our experiments.

We follow the official hyper-parameters setting to train the research model. More specifically, we fine-tuned the model with 15 epochs, a batch size of 8, and a weight decay of 0.01. We use the model with the highest accuracy on the validation set, and test its performance on the test set.

Video Content Moderation. All the five commercial software products mentioned above provide video content moderation services. Similar to the image content moderation experimental settings, we conduct experiments on the toxic content category if the APIs provide the corresponding detection services. Thus, there are no results
on hate detection in Google Cloud and advertisement detection in Amazon Web Service, Google Cloud, and Tencent Cloud. Since the video content moderation service in Amazon Web Service and Google Cloud do not support audio modality analysis, we consider the modality fusion between vision and textual for these two APIs. Since there is no publicly available multi-modal video classification dataset for hate, pornographic, and advertisement detection, we do not conduct testing on the research model for video content moderation.

For each tested software product or model, we first input all the seed data and filter out the sentence that cannot be identified as toxic. In other words, we only used those that have already been recognized as toxic contents as seed data to generate test cases. The statistics of seed data are shown in Table 3, which shows that most of the generated seed sentences are toxic. Then, we generate image and video test cases accordingly. Finally, we use the generated test cases to test the software products and research models.

To evaluate how well DUO does on generating test cases that trigger errors, we calculate Error Finding Rate (EFR), which is defined as follows:

\[
EFR = \frac{\text{the number of misclassified test cases}}{\text{the number of generated test cases}}
\]

4.2.2 Analysis. We list the results in Table 4 and 5.

**Overall Analysis.** In general, DUO is able to find errors in software products and research models with a relatively high EFR. Since most of these test cases are annotated as toxic according to our annotators, we believe such high EFR implies the effectiveness of DUO and reveals the unexpected vulnerability of widely-deployed software products and research models.

One common concern about AI software testing is whether the software performs well on existing test cases, which are toxic single-modal inputs in content moderation. To address this concern, we conduct a lightweight experiment to evaluate the effectiveness of the software under test in detecting toxic contents from the Internet. Since there is no publicly-available toxic (hateful, porn, and malicious ad) image benchmark dataset (probably due to the toxic nature), we manually collect a dataset with 50 hateful images, 50 porno images, and 50 ad images from the Internet. The average detection rate of five content moderation software is 97.8%, indicating the effectiveness of the software. Thus, we think the high EFR achieved by DUO is exciting.

The comparison between image and video. The robustness of image content moderation and that of video content moderation do not strictly hold a positive correlation. For example, Baidu’s image moderation is much worse than its own video moderation, even though the testing logic should be similar. We think the algorithms behind them are different.

The comparison between different software. Since the seed sentence and test cases are different, it is difficult to compare the APIs for English content, i.e., Amazon Web Service and Google Cloud, with the APIs for Chinese content, i.e., Baidu Cloud, Tencent Cloud, and Alibaba Cloud. Even if the APIs for the same language are using the same test cases, it is also difficult to obtain an interesting conclusion because the functionalities of different software could vary a lot.

The comparison within a product. The performance of moderating different kinds of toxic data for one software product is different. For example, the robustness of pornography detection is much better than that of advertisement detection for the image content moderation APIs provided by Baidu, Tencent, and Alibaba. We think it is because the bulk of Internet Companies’ revenue comes from advertising so it is common to share the link, image, or video of advertisement between the users on these platforms. In addition, in Chinese culture, advertisement is not as toxic as pornography. Therefore, these companies seem to pay less attention to advertisement detection than pornography detection when developing their content moderation software.

**Answer to RQ2:** DUO can find substantial errors in both commercial software and research models.

### 4.3 RQ3: What factors affect the performance of DUO?

This section explores the impact of two factors on the DUO’s performance. First, we studied the impact of the location of the image and text. In theory, a video should be identified as toxic content if it contains a toxic image or text, no matter where the image or text is located. However, we found that changing the location of the image and text can affect the prediction of content moderation software. More specifically, we randomly select 20 seed sentences for each kind of toxicity and generate two kinds of vision + text video test cases with different locations: 1) the image and the text are showing on the left of the screen; 2) the image and the text are showing on the right of the screen. The test cases are fed into the video content moderation software products. We found that on average 3.2% of video test cases can bypass the moderation after we move the image and the text from the left to the right of the screen. This result implies that we can find the best location for toxic images and text if we want to find more cases that can bypass moderation.

<table>
<thead>
<tr>
<th>Software</th>
<th>Tasks</th>
<th>Seed Num</th>
<th>Toxic Num</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Abuse&amp;Hate</td>
<td>100</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>pornography</td>
<td>100</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Advertisement</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>Baidu</td>
<td>Abuse&amp;Hate</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>pornography</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Advertisement</td>
<td>100</td>
<td>91</td>
</tr>
<tr>
<td>Tencent</td>
<td>Abuse&amp;Hate</td>
<td>100</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>pornography</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>Advertisement</td>
<td>100</td>
<td>91</td>
</tr>
<tr>
<td>Alibaba</td>
<td>Abuse&amp;Hate</td>
<td>100</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>pornography</td>
<td>100</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Advertisement</td>
<td>100</td>
<td>84</td>
</tr>
</tbody>
</table>
Table 4: Error finding rates of image content moderation software and academic models (AM).

<table>
<thead>
<tr>
<th></th>
<th>Amazon</th>
<th>Google</th>
<th>AM_ResNet</th>
<th>AM_ViT</th>
<th>Baidu</th>
<th>Tencent</th>
<th>Alibaba</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abuse Detection</td>
<td>91.7%</td>
<td>-</td>
<td>94.1%</td>
<td>90.6%</td>
<td>99.55%</td>
<td>-</td>
<td>76.50%</td>
</tr>
<tr>
<td>Porno Detection</td>
<td>78.82%</td>
<td>77.18%</td>
<td>-</td>
<td>-</td>
<td>96.49%</td>
<td>76.77%</td>
<td>33.14%</td>
</tr>
<tr>
<td>Ad Detection</td>
<td>-</td>
<td>-</td>
<td>78.8%</td>
<td>75.3%</td>
<td>99.78%</td>
<td>99.29%</td>
<td>65.46%</td>
</tr>
</tbody>
</table>

Table 5: Error finding rates of video content moderation software.

<table>
<thead>
<tr>
<th>Perturbation</th>
<th>Abuse Detection</th>
<th>Porn Detection</th>
<th>Ad Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modalities</td>
<td>Amazon Baidu Tencent Alibaba</td>
<td>Amazon Google Baidu Tencent Alibaba</td>
<td>Baidu Alibaba</td>
</tr>
<tr>
<td>V + T</td>
<td>89% 71.91% 100% 60.81%</td>
<td>91.38% 100% 50.53% 100% 60%</td>
<td>0% 100%</td>
</tr>
<tr>
<td>V + A</td>
<td>- 100% 100% 54.05%</td>
<td>- - 91.58% 100% 55.29%</td>
<td>5.49% 100%</td>
</tr>
<tr>
<td>A + T</td>
<td>32.58% 100% 81.08%</td>
<td>- - 10.53% 100% 58.82%</td>
<td>0% 100%</td>
</tr>
<tr>
<td>V + A + T</td>
<td>76.67% 100% 97.29%</td>
<td>- - 55.79% 100% 97.64%</td>
<td>7.61% 100%</td>
</tr>
</tbody>
</table>

Second, we studied the impact of the selection of modality. In theory, a toxic content should be identified as toxic no matter which modality it is. For example, an abusive sentence should be categorized as toxic no matter if it is in text format or being converted to audio format. However, we found that for multimedia content moderation software, the sensitivity of different modalities is not the same. To this end, we randomly select 20 toxic seed sentences for each kind of toxicity and generate two kinds of video test cases: 1) pure text video that shows the sentence from the beginning to the end; 2) pure audio video that uses text-to-speech APIs to convert the seed sentence to audio and playing the audio. The test cases are fed into the video content moderation software products. We found that on average 94.7% of the pure text video can be detected as toxic while only 84.2% of the pure audio video can be detected. This result implies that we can distribute more toxic information to the less sensitive modality if we want to find more cases that can bypass moderation.

Answer to RQ3: The location of the image and text and the modality can affect the performance of DUO.

4.4 RQ4: Can we use DUO to improve the performance of multimedia content moderation?

In this section, we discuss how to use DUO to improve the robustness of multimedia content moderation. A natural way would be to retrain the models using the test cases generated by DUO and check whether the retrained models become more robust. This method is known as the robust retraining method. For each research model, we randomly select 100 test cases that successfully bypass moderation and label them as toxic. After that, we add these 100 labeled images to the corresponding training set and re-train the ResNet-18 model and the Vision-Transformer model. We adopt the default fine-tuning settings from the Hugging Face website. In other words, the setting of robust training is identical to the setting of normal training. The only difference is the data.

Table 6: Improvement of models with DUO

<table>
<thead>
<tr>
<th>Model</th>
<th>Task</th>
<th>Ori</th>
<th>Rob</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>Abuse</td>
<td>94.1%</td>
<td>5.7%</td>
</tr>
<tr>
<td></td>
<td>Ad</td>
<td>78.8%</td>
<td>3.6%</td>
</tr>
<tr>
<td>ViT</td>
<td>Abuse</td>
<td>90.6%</td>
<td>4.8%</td>
</tr>
<tr>
<td></td>
<td>Ad</td>
<td>75.3%</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

To validate the effectiveness of robust retraining with DUO, we use DUO to test the model after robust retraining, denoted as "Rob", and compared its EFRs with the original model, denoted as "Ori". The results are presented in Table 6. We can observe that robust training with DUO’s test cases largely reduces the EFRs on all the settings, which shows that DUO can effectively improve the robustness of multimedia content moderation models. In addition, we also evaluate their model accuracy on the original test set and the results show that in all settings, model accuracy remains the same after robust training.

Answer to RQ4: Test cases generated by DUO can effectively improve the robustness of multimedia content moderation models without affecting their accuracy.

5 DISCUSSION

5.1 Summary of Findings

5.1.1 Different Modalities Matter. In the evaluation, we can observe that different content moderation software has different specialties. For example, the content moderation software of Baidu performs quite well (i.e., low EFR) on videos fused by audio and text, while cannot achieve satisfying performance on those fused by vision and audio. In addition, the software from Alibaba can achieve 65.46% EFR on image-based Ad detection, while it is infeasible to detect video-based advertisements (i.e., 100% EFR for all kinds of fused videos). Such characteristics make it non-trivial to
comprehensively test content moderation software, e.g., test cases for software may be inappropriate for another. In contrast, DUO considers three different modalities, and performs semantic fusion to generate both image and video test cases in English and Chinese. Such design makes DUO produce comprehensive test cases for different content moderation tasks. In addition, DUO is designed for black-box testing, which does not require white-box access to the software under test. Such design makes it convenient for adapting DUO to other AI-based software testing tasks, such as image caption systems.

5.1.2 Future Direction. Although DUO is able to find numerous errors in multimedia content moderation software and research models, there are several limitations that can lead to future works.

First, we can improve the diversity of our generated test cases. In this work, we only design one simple template and extract limited keywords from a toxic dataset for each kind of toxicity to generate test cases. In future work, we can follow the framework of this paper but use more kinds of templates and keywords.

Second, we can improve the authenticity of our generated test cases. In this work, we conduct multi-modal fusion by inserting text into an image or showing an image while playing a synthesized speech. In future work, we can utilize more advanced conditional image or video generation models to generate more vivid test cases.

5.2 Simplicity of the Methodology

While our Semantic Fusion methodology is simple, it is novel, general, and effective. The idea that fuses the semantics of single-modal inputs into a new input that combines the semantics of its ancestors has not been explored by any existing studies. Meanwhile, although the templates we use are simple, the single-model inputs can be automatically extracted from media sources or datasets to generate diverse test cases. It can be easily generalized to more complicated templates and seeds. In addition, the test cases generated from our templates can already reveal many errors in real-world software, which was reported to be highly effective by the companies. We believe that our simple methodology could benefit the community and we hope our work will be a starting point that brings more attention to multimedia content moderation in the field.

5.3 Threats to Validity

Internal Validity. The major threat to internal validity is that the image and video test cases generated by DUO may no longer preserve the toxic nature of the seed sentence, which may cause false positives during testing. Besides, DUO only uses simple templates to extract keyword pairs for test case generation. The template design can also affect the quality of the generated test cases, which may not be realistic enough that seems to be applied by users and appear in the real world. To relieve these issues, we validated the generated test cases by conducting human annotations, which shows that our generated text cases are toxic and realistic. During the manual inspection of the evaluation results, we also observe few false positive. In addition, DUO adopts existing tools for data transformation, e.g., Flair for PoS tagging and ffmpeg for video processing, whose performance may affect the quality of generated cases. To relieve this issue, we only chose tools that are widely used by millions of users. Moreover, these tools are also replaceable, which means users of DUO can customize these tools for specific testing tasks.

External Validity. The major threat to external validity is whether the content moderation software products we test are good and representative of what the industry uses. To relieve this issue, we evaluated the effectiveness of these products. The average detection rate of five content moderation software products is 97.8%, indicating their effectiveness. As for representativeness, all five commercial software products are paid cloud services provided by the cloud platform from the big companies. Moderation services and other cloud services have become an important source of income for them. Meanwhile, a huge amount of downstream companies and users are using paid services provided by these companies. The customer list can be seen from Google, Amazon and Baidu.

Thus, we believe the moderation software studied in our paper has a significant impact on real users and is representative of industry practice. The other possible threat to external validity lies in the source datasets used for keyword extraction. Low-quality keywords can directly affect the testing effectiveness of generated test cases. To relieve this threat, we extracted keywords from six practical datasets, which consist of real-world toxic contents. Manual inspection and the evaluation results also prove that DUO can produce effective multimedia contents. Another threat may lie in the comprehensiveness and representation of our metamorphic relations, which might hurt the generalizability of our results and findings. To mitigate this threat, we extensively implemented splitting and fusion for all the possible permutations of different modalities. Hence, the final metamorphic relations are empirically comprehensive. To further validate the generalizability of DUO, we also conducted extensive experiments. Specifically, we tested DUO on five commercial textual software products and two research methods for content moderation. We also tested DUO with different kinds of toxic data: abuse hate speech, porn content and malicious advertisement, in both English and Chinese. The evaluation results confirm that our findings can generalize to different methods and tasks.

6 RELATED WORK

6.1 Content Moderation

There are generally two categories of research models for content moderation, i.e., feature engineering-based models and neural networks-based models. Feature engineering-based models work by manually designing features and using the features to train classification models to identify toxic contents. While a neural networks-based method is typically a neural network model trained in an end-to-end manner on a huge amount of data, which does not require much human effort on feature engineering.

6.1.1 Single-modal Content Moderation. In the literature, there are mainly three categories of data modalities, i.e., visual, audio, and textual.

Visual Content Moderation. Early works on toxic image detection mainly extract some pre-designed features and then train

https://cloud.google.com/customers/#/products=Data_Analytics
https://aws.amazon.com/machine-learning/customers/
https://cloud.baidu.com/partner/plan.html#search
a machine learning classifier. For example, Shen et al. [54] computed color histograms to detect pornography images, depending on the activity suspicion that pixels in pornographic images are mostly skin. Zhang et al. [78] extract the motion features and use a Hidden Markov Model (HMM) for video anomaly detection. With the development of deep learning techniques and the building of big datasets, the progress of image representation learning has motivated researchers to explore neural network-based models for visual content moderation. For example, Moustafa [42] adapted a CNN architecture for image classification to the pornographic video classification task. Each frame is input for being classified as porn or non-porn and then integrating the final result for a video via a majority voting process.

Audio Content Moderation. Zhang [78] extracted Mel-Frequency Cepstral Coefficients (MFCCs) feature and use an HMM to detect audio events. [55] extract various audio features, including MFCC, mean of short-time zero-crossing rates(mSTZCR), mean of the spectral centroid(SC) and high short-time zero-crossing rates ratio(HZCRR), and use in-class clustering for porno-sounds detection. Gupta [16] utilized a bidirectional Gated Recurrent Units(GRU) and Long Short-Term Memory (LSTM) model to detect abuse from multilingual audio. Recently researchers proposed a deep learning method that can detect COVID-19 from breath and cough audio [7].

Textual Content Moderation. Early works are rule-based methods that are based on pre-defined rules or dictionaries[45, 59]. However, rule-based can hardly deal with implicit abuse and sarcasm. Besides, they are vulnerable to errors in spelling, punctuation, and grammar [70]. Computation-based methods leverage some statistics of the textual data, such as TF-IDF[51, 74]. Computation-based methods require less human effort and are more robust to spelling, punctuation, and grammatical variations. Nevertheless, this kind of approach can only capture surface-level patterns, not deeper semantic properties [70]. The advancements in text representation learning have motivated researchers to explore neural network-based models, such as Feed-forward neural networks[11], LSTM[1] or the pre-training models [10, 37], on textual content moderation task and have achieved remarkable performance.

### 6.1.2 Testing Multi-modal Content Moderation

The testing of multi-modal content moderation is relatively unexplored compared with that of single-modal content moderation. Kiela et al. [28] manually created and labeled a multi-modal hateful memes dataset as a benchmark to evaluate the multi-modal understanding ability of cross-vision and text modality. Hussain et al. [21] collected and labeled a TV advertisement classification benchmark that predicts whether a frame of video is an advertisement. DUO is different from the existing papers: 1) existing work needs huge human effort to create and label the multi-modal test cases while DUO can automatically generate and label the test cases; 2) existing work only provides image test cases on English hateful and advertisement image detection, while DUO is able to comprehensively generate image and video test cases in two languages for three kinds of toxicity detection. In addition, our experiments show that DUO can find numerous errors from models trained on the datasets provided by existing work. We believe our work complements this line of existing papers.

### 6.2 Robustness of AI Software

Artificial Intelligence (AI) software has been widely adopted by many domains, such as autonomous driving and face recognition. However, AI software is not robust enough and can generate erroneous outputs that lead to fatal accidents [34, 83]. To this end, researchers have proposed a variety of methods to generate adversarial examples or test cases that can fool AI software [4, 5, 13, 23, 27, 29, 30, 38–40, 44, 50, 60–62, 64, 68, 69, 79–81].

AI software has also become adept at solving many NLP tasks, such as sentiment analysis [82], reading comprehension [75], Grammatical Error Correction [72] and machine translation [2, 24, 25]. In recent years, inspired by the work on adversarial examples in the AI field, NLP researchers have started exploring attack and defense techniques for various NLP software [15, 19, 63]. For example, Ribeiro et al. [46] designed a behavioral testing method to test NLP software for sentiment analysis, duplicate question answering, and machine comprehension. Li et al. [35] used deep learning models to generate test cases for deep learning-based NLP software. And Wang et al. [67] proposed a metamorphic testing framework for textual content moderation software. Unlike these studies, this work focuses on the robustness of multimedia content moderation software, which has not been explored by existing work.

### 7 CONCLUSION

In this paper, we propose a general, effective methodology, **Semantic Fusion**, for validating multimedia content moderation software and further realize it as a practical tool DUO. DUO can generate image and video test cases covering three kinds of toxic contents, including abusive language and hate speech, malicious advertisements, and pornography, in both two language settings (English and Chinese). We used DUO to test five commercial content moderation software products and two research models. Results show that DUO can find numerous errors from all the software and models. In addition, we showed that we can improve the models’ robustness by utilizing the test cases produced by DUO to perform robust training. We hope our framework can help identify the defects of current content moderation software and facilitate their development, contributing to a cleaner and better Internet environment.

### ACKNOWLEDGEMENT

The work described in this paper was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14206921 of the General Research Fund) and the National Natural Science Foundation of China (Grant Nos. 62102340 and 62206318).

### REFERENCES


