Robustness of Fusion-based Multimodal Classifiers to Cross-Modal Content Dilutions

Gaurav Verma Georgia Institute of Technology gverma@gatech.edu
Vishwa Vinay and **Ryan A. Rossi** Adobe Research {vinay, ryrossi}@adobe.com
Srijan Kumar Georgia Institute of Technology srijan@gatech.edu

Abstract

As multimodal learning finds applications in a wide variety of high-stakes societal tasks, investigating their robustness becomes important. Existing work has focused on understanding the robustness of vision-and-language models to *imperceptible* variations on benchmark tasks. In this work, we investigate the robustness of multimodal classifiers to *cross-modal dilutions* – a *plausible* variation. We develop a model that, given a multimodal (image + text) input, generates additional dilution text that *(a)* maintains relevance and topical coherence with the image and existing text, and *(b)* when added to the original text, leads to misclassification of the multimodal input. Via experiments on Crisis Humanitarianism and Sentiment Detection tasks, we find that the performance of task-specific fusion-based multimodal classifiers drops by 23.3% and 22.5%, respectively, in the presence of dilutions generated by our model. Metric-based comparisons with several baselines and human evaluations indicate that our dilutions show higher relevance and topical coherence, while simultaneously being more effective at demonstrating the brittleness of the multimodal classifiers. Our work aims to highlight and encourage further research on the robustness of deep multimodal models to realistic variations, especially in human-facing societal applications.

1 Introduction

Rich multimodal content understanding is crucial for several AI for Social Good applications like humanitarian information detection during crises, hate speech analyses, and fake news mitigation (Ofli et al., 2020; Kiela et al., 2020; Facebook, 2020; Khattar et al., 2019; Verma et al., 2022). In many such scenarios, the information in individual modalities, either image or text, is designed to be complementary to information in the other modality. As such, joint modeling of both modalities is of fundamental importance, and consequently, technologies that enable multimodal understanding are advancing rapidly and are being deployed at scale (Nayak, 2021; Grauman et al., 2021).

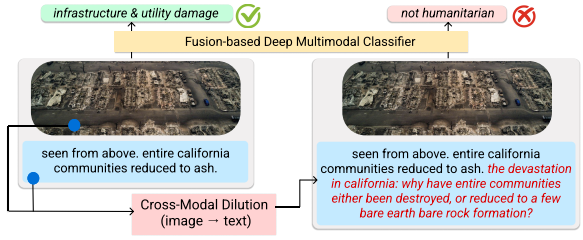

Figure 1: **Overview of our study.** We investigate the robustness of fusion-based deep multimodal classifiers to cross-modal dilutions. We generate dilutions that maintain semantic relevance with the original text and image while causing incorrect classifications. We also demonstrate the realistic nature of cross-modal dilutions using human evaluation. The figure shows an actual example from our experiments.

It is desirable that deep learning models are robust to dilution-based variations in input. *Dilution* is defined as the addition of related content that dilutes the effect of the original information. Naik et al. (2018) and Ribeiro et al. (2020) argue that natural language processing (NLP) models should not alter their predictions after adding dilutions — for instance, appending statements like "and true is true" (multiple times) for the Natural Language Inference task and adding randomly created URLs for the Sentiment Analysis task.

We study the robustness of multimodal classifiers to dilutions. Compared to the simple dilutions created for NLP tasks, we aim to explore realistic dilutions for multimodal data. Since what entails plausible dilutions for multimodal data has not been established, we propose a new category of dilutions specific for multimodal content, named *cross-modal* dilutions. Cross-modal dilution involves adding relevant information from the image modality to the text modality for a multimodal input; see Figure 1. Our notion of dilution, unlike the examples above, is contextual – that is, the change

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 360 - 374
December 7-11, 2022

introduced varies for different information items. Additionally, evaluating robustness to dilutions in a multimodal setting is non-trivial because the possible additions are constrained by the semantics of both the image and the original text.

Previous research on the robustness of deep multimodal learning focuses on perturbations for Visual Question Answering (VQA) (Srivastava et al., 2020; Zhang et al., 2019; Gupta, 2017; Wu et al., 2017) and involves making minor alterations to the textual questions (Mudrakarta et al., 2018), or asking more challenging questions than what were present in the training dataset (Sheng et al., 2021; Li et al., 2021b). In contrast, we focus on multimodal classification and study dilution-based variations. To this end, we propose a method that leverages a large language model to generate *additional* text that is *(i)* related to the information in the image, *(ii)* semantically aligned to the existing user-provided textual description, and *(iii)* is adversarial in nature (i.e., when added to the existing description, leads to incorrect predictions by multimodal models). The first two constraints ensure that the additional text is realistic, while the third constraint enables us to assess the robustness of multimodal classifiers under these settings.

Our contributions are summarized as follows:

• We propose and investigate the robustness of multimodal classifiers to cross-modal dilutions. We develop an approach that leverages keywords from image and text to perform controlled generation of semantically relevant text that can be appended to the original text to cause misclassification.

• Via extensive evaluation covering aspects like adversarial effectiveness, content relevance, diversity, and coherence, we establish that the dilutions generated by our proposed model are better than several rule-based and model-based baselines. We release our code to aid future research.[1]

• We conduct human evaluations to *(a)* assess the quality of generated dilutions over the most competitive baseline and *(b)* establish the realistic nature of diluted multimodal examples. We find that our cross-modal dilutions are perceived by humans as better than the baseline dilutions and more realistic.

2 Related Work

Robustness of Multimodal Models: Existing research studies the robustness of multimodal models by making imperceptible adversarial changes to the individual input modalities using unimodal perturbations (Li et al., 2020a; Chen et al., 2020). However, while adversarial perturbations to images are often deemed as imperceptible to humans, the adversarial perturbations in text often compromise the semantic meaning and its category to notable extents (Wang et al., 2021). In the context of multimodal learning, the problem of introducing textual perturbations that lead to semantically poor changes has been tackled by developing careful automated approaches – for instance, by synthesizing counterfactual samples using language models (Chen et al., 2020), or by conducting human-in-the-loop curation of adversarial examples (Sheng et al., 2021; Li et al., 2021b). However, these studies only focus on VQA (Antol et al., 2015). Additionally, as Gilmer et al. (2018) argue, the imperceptibility criterion does not constrain the plausible action space in human-facing applications. For instance, it has been shown that the human-provided description of an image can vary notably with the personality, age, and location of the writer in terms of its length, emotion, and vocabulary; all the while preserving the cross-modal semantic interaction (Shuster et al., 2019; Chunseong Park et al., 2017; Denton et al., 2015). Consequently, in this work, we focus on the robustness of multimodal classifiers to *plausible* variations, specifically cross-modal dilutions.

Adversarial Perturbations: Our investigation concerns adding related text in a multimodal example to the existing textual information. Several methods have been proposed to introduce *imperceptible* and *adversarial* perturbations in text (Li et al., 2021a, 2020b; Garg and Ramakrishnan, 2020), focusing on word-level or phrase-level automated insertions, replacements, and merging. Moving beyond the imperceptibility constraint, to estimate robustness to perceptible but plausible changes in text, recent research has investigated the robustness of NLP models to *rule-based* distractions that are added to the original text (Naik et al., 2018; Ribeiro et al., 2020). As the constraints that govern textual dilutions in a multimodal setting are different, we propose a model to generate cross-modal (image → text) dilutions that maintain semantic and topical coherence with the existing image and text, while also demonstrating adversarial properties with respect to the multimodal classifiers. This provides us with a realistic estimate of the robustness of multimodal classifiers.

[1]Project webpage with code: https://claws-lab.github.io/multimodal-robustness/

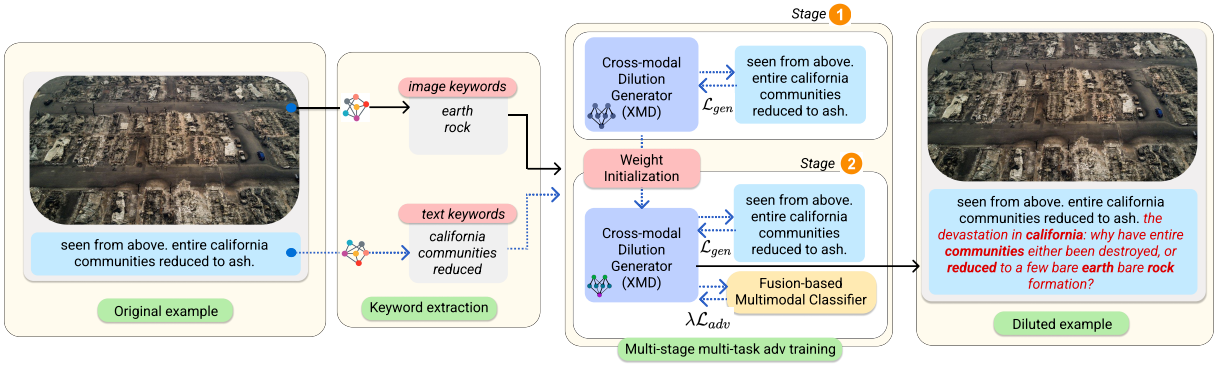

Figure 2: **Overview of our proposed method.** We propose XMD — Cross-Modal Dilution Generator. Our approach extracts keywords from the image and text of a multimodal example and generates dilution text that causes incorrect classifications by the multimodal classifier when appended to the original text. The generation model is trained in a multi-stage multi-task setup, where the adversarial loss component (stage 2) encourages the generation of dilution words that cause incorrect categorization. The blue dashed lines depict the training pathway.

3 Cross-Modal Dilutions

Related work on language-only models (Naik et al., 2018; Ribeiro et al., 2020) inspires us to study the robustness of deep multimodal classifiers to dilution. In the context of multimodal learning, dilutions can be introduced by adding information from the associated image to the original text. Since multimodal fusion models are expected to consider the information in images and text jointly, they should, in principle, be robust to the expression of additional information regarding the image in the form of text. This is, however, challenging to study because a plausible dilution should have semantic similarity with both the image and the original text. While a rule-based dilution like “and true is true" (investigated by Naik et al. (2018)) are plausible for specific language-only tasks like Natural Language Inference (Bowman et al., 2015), they do not cover the action space of plausible cross-modal dilutions for multimodal content. Therefore, we develop an approach to generate dilutions that are semantically aligned with original text and image.

Our proposed approach follows the following framework to generate dilutions; see Figure 2.
(i) Extract keywords from image and text based on their prominence in their respective modalities.
(ii) Train a language model to fill words around the extracted keywords from the original text to generate dilutions (Zhang et al., 2020). The generation model is trained using a multi-stage multi-task approach. The first stage fine-tunes the model to generate in-domain text using textual keywords in a self-supervised manner. The second stage involves training the model on an objective that combines generation loss with adversarial loss.
(iii) The trained model is then used to generate text based on the keywords combined from both text and image modalities. The generated text is then appended to the original text as dilution.

3.1 Method for Generating Dilutions

Multimodal classifier: We design a fusion-based multimodal classification model ($\mathcal{M}_{mm}$) following widely adopted architectures in both academic research and industrial applications (Agarwal et al., 2020; Dataminr, 2020). $\mathcal{M}_{mm}$ takes the concatenation of modality-specific representations as input and makes a joint classification. To model individual modalities, we first train an image-only classifier $\mathcal{M}_{image}$ and a text-only classifier $\mathcal{M}_{text}$ for the same classification task. We then concatenate the output of the penultimate layers of the modality-specific models to feed them into a fully-connected network that is trained to fuse the modality-specific representations to perform joint classification based on the multimodal input.
Keyword extraction: Our dilution generation approach is centered around keywords in the original image and text as that will ensure semantic relatedness of the dilution text with both the associated modalities. We use Yet Another Keyword Extractor (YAKE) (Campos et al., 2018) to extract the most important keywords from the textual description for each example. For extracting keywords from the image, we consider the top 150 objects in the Visual Genome dataset (Krishna et al., 2017) and identify these objects in our dataset using a pre-trained image to Scene Graph generator (Tang et al., 2020). We further filtered the list of all identified objects by only considering objects with a

bounding box that occupies at least 10% of the total image area to ensure prominence in the image. These objects are considered the keywords of the image. We denote the keywords from text and image as K_{text} and K_{image}, respectively.

Constrained text generation: Once we have the keywords from text and image for each of the examples, the goal is to generate dilution around these keywords. For this, we extend the constrained text generation approach proposed by Zhang et al. (2020). We fine-tune a BERT language model to progressively predict [MASK] tokens around the initial set of keywords until only a special token (i.e., no-insertion token [NOI]) is predicted at all places to indicate no further insertions. We consider the original descriptions of the training examples in our target dataset and fine-tune the pre-trained model to reconstruct the original examples using keywords in the text, i.e., K_{text}. We adopt the same generation objective as Zhang et al. (2020) and denote it as $\mathcal{L}_{gen}$. The fine-tuned model can generate domain-specific text using the supplied keywords during inference.

Adversarial training: While the above fine-tuning enables constrained generation of target-domain text based on the supplied keywords, we need to ensure that the generated dilutions also cause incorrect classifications by the trained multimodal classifier $\mathcal{M}_{mm}$. Explicitly designing the generation process to exhibit adversarial nature provides an estimate of the possible drop in performance in the presence of cross-modal dilutions. To this end, we consider the POINTER model after domain-specific fine-tuning and fine-tune it further using a combined loss function. The combined loss function takes into account not only the original generation loss but also a weighted component of the adversarial loss $\mathcal{L}_{adv}$. More formally,

$$\mathcal{L}_{combined} = \mathcal{L}_{gen} + \lambda \mathcal{L}_{adv} \quad (1)$$

where λ controls the contribution of the adversarial loss towards the generation process. The incorporation of $\mathcal{L}_{adv}$ encourages the generation model to fill the [MASK] tokens with words that would cause incorrect classifications by the multimodal classifier $\mathcal{M}_{mm}$. More formally, $\mathcal{L}_{adv}$ is computed for each training example as:

$$\mathcal{L}_{adv} = -(y \log(\widehat{y}) + (1-y) \log(1-\widehat{y}))$$

where $y = 1$ when the predicted class by $\mathcal{M}_{mm}$ is different from the ground-truth class and $y = 0$ when the predicted and ground-truth labels are the same. The probability of incorrect classification, i.e., $\hat{y}$, is obtained by adding the class probabilities of incorrect classes (Le et al., 2020; He et al., 2021). The training of the generation model is done in a multi-stage manner — in the first stage, the model is fine-tuned to generate related in-domain text from keywords using $\mathcal{L}_{gen}$ and then in the second stage, it is trained in a multi-task fashion using a weighted combination of $\mathcal{L}_{gen}$ and $\mathcal{L}_{adv}$. This ensures that the model maintains the quality and coherence in the generated text while learning adversarial behavior.

Inference-time dilution generation: We use the constrained text generation model described above to generate text based on combined keywords from both text and image (i.e., $K_{text} \oplus K_{image}$ — where $\oplus$ denotes the concatenation of keywords). These generated textual dilutions are added to the original text to obtain examples with cross-modal dilutions. Our evaluation aims to assess the impact of these cross-modal dilutions on the performance of the trained multimodal classifier M_{mm} along with various attributes of the generated text.

4 Multimodal Datasets

We conduct experiments on two user-generated datasets that have real-world societal applications.

Crisis Humanitarianism Dataset: During crises, affected parties often use social media to communicate with humanitarian organizations that process the available information to provide timely and effective interventions. To aid development of related computational methods, Alam et al. (2018) curated the CrisisMMD dataset. This multimodal dataset comprises 7,216 Twitter posts in English (images + text) that are categorized into 5 humanitarian categories.[2] We formulate the task of humanitarian information detection as a multi-class classification problem, and use the standard training ($n = 5263$), evaluation ($n = 998$), and test ($n = 955$) sets in our experiments.

Sentiment Detection Dataset: User-generated content has been frequently used to infer sentiments of individuals for various applications, including detection of mental health indicators (De Choudhury et al., 2013). We collect the dataset introduced by Duong et al. (2017) for the task of sentiment

[2] infrastructure and utility damage: 10%, rescue volunteering or donation effort: 14%, affected individuals: 1%, other relevant information: 22%, & not humanitarian: 53%

detection. The dataset comprises multimodal posts (in English) from Reddit that are categorized into 4 classes.[3] We crawled the images from Reddit URLs provided by the authors and split the dataset in a 80:10:10 ratio to obtain the train ($n = 2568$), validation ($n = 321$), and test ($n = 318$) sets.

5 Experiments

We first discuss the training of our proposed cross-modal dilutions (XMD) generator model. Then, we discuss multiple baselines that dilute the original text using various rule- and model-based approaches. Finally, we evaluate XMD and compare its performance with the baselines.

5.1 Training Details

Multimodal Classifier ($\mathcal{M}_{mm}$): $\mathcal{M}_{mm}$ is a fusion of text-only and image-only classifiers. For text-only classifier $\mathcal{M}_{text}$, we fine-tune and evaluate a BERT (Devlin et al., 2018) model on the target dataset. Similarly, we fine-tune a VGG-16 model (Simonyan and Zisserman, 2015) pre-trained on ImageNet (Deng et al., 2009) to train an image-only classifier $\mathcal{M}_{image}$. We refer the reader to Appendix A.1 and A.2 for details and evaluation of the modality-specific classifiers. We feed the concatenation of fine-tuned text and image representations to the multimodal classifier, which is essentially a series of fully-connected layers with ReLU activation (Agarap, 2018). The architecture of the multimodal classifier comprises an input layer (1024 neurons), 3 hidden layers (512, 128, 32 neurons), and an output layer (neurons = number of classes in the dataset). We use Adam optimizer (Kingma and Ba, 2014) with a learning rate initialized at 10^{-4} and adopt early stopping based on the validation set loss to avoid overfitting.

Cross-modal dilutions generator (XMD): For keyword extraction from YAKE, we set the maximum n-gram size to 1, the de-duplication threshold to 0.9 with 'seqm' function, and the window size to 1. The rest of the hyper-parameters were set to their default values used in previous studies (Zhang et al., 2020; Tang et al., 2020). We fine-tune the POINTER model pre-trained on Wikipedia text (Zhang et al., 2020) using default hyper-parameters for 5 epochs. During this stage of the training, the objective is $\mathcal{L}_{gen}$ and the model learns to generate text from keywords that aligns with the target domain. Following this, we further train the generation model for another 1 epoch using the combined objective in Equation 1, while setting $\lambda = 0.01$ (based on results on the validation set). This adversarial adaptation of the model encourages generations that could cause misclassifications by the trained $\mathcal{M}_{mm}$. Finally, the keywords from images and text (i.e., $K_{text} \oplus K_{image}$) are passed as input to XMD to generate dilutions for the examples in the test set.

[3] creepy: 22%, rage: 19%, gore: 25%, & happy: 34%

5.2 Baselines

Rule-based dilutions:
(i) Random URL: As proposed by Ribeiro et al. (2020), we append a randomly generated twitter URL (e.g., https:t.co/gXvDrs) to the original text.
(ii) Relevant keywords: We experiment with adding extracted keywords from the image, text, and both together to the original text.
(iii) Most similar image's description: We add the textual description of the most similar image (computed using cosine similarity between fine-tuned VGG-16 embeddings of images in the test set) to the original text. This mimics scenarios where the user dilutes the original text by adding the description of a highly similar image; see Appendix A.4.

Model-based dilutions:
(i) GPT: We use the original text as the prompt for a GPT-2 (Radford et al., 2019; Wolf et al., 2020) model and add the generated text to it for dilution.
(ii) GPT Fine-tuned: We first fine-tune a GPT-2 model using the text in the training set of the dataset (using default hyper-parameters) for domain adaptation, and then use the original text as the prompt to obtain the dilution text.
(iii) Image Captioning: We use two trained image captioning models (SCST (Rennie et al., 2017) & XLAN (Pan et al., 2020)) to generate the captions for the images in the test set. We append the generated captions to the original text for dilution.

5.3 Evaluation metrics

Our evaluation is focused on assessing two aspects of the dilutions: *(a)* are the dilutions effective in deteriorating the classification performance of the multimodal classifier?, and *(b)* are the added dilutions relevant to the original text + image, and maintain topical coherence with the existing text? To this end, we compute standard classification metrics for the former evaluation and compute embedding-based similarity measures for the latter. **Sim**$_{text}$ denotes the similarity between the original

text and the generated dilution and is computed using the cosine similarity between the embeddings from the fine-tuned BERT classifier. Similarly, **Sim**$_{img}$ denotes the similarity between the generated dilution and the image and is computed using cosine similarity between CLIP embeddings (Radford et al., 2021). For topical coherence, we compute the KL Divergence (**KL Div**) between the topic distributions of the original text and the generated text. For details regarding the training of the topic model, please see Appendix A.6. Additionally, we quantify the correspondence similarity between image and final text (i.e., original + dilution) using a learned metric **Sim**$_{corr}$ that quantifies the correspondence between *diluted descriptions and original images* based on the correspondence between *original text and images*; see Appendix A.5 for further details. Furthermore, we compute Self-BLEU (Zhu et al., 2018) scores for the sentences in the generated dilution to quantify diversity, wherever applicable. For all model-based baselines and XMD, we report the average values over 5 runs with different random seeds.

6 Main Results

Our results (Tables 1 & 2) show the following:
• Rule-based dilutions do not demonstrate adversarial effectiveness with the exception of using most similar image's description as dilution, which however, shows poor relevance and topical coherence.
• Model-based baselines show adversarial effectiveness but lack in relevance and coherence.
• XMD demonstrates the best adversarial effectiveness while generating more relevant and topically coherent dilutions with respect to all the baselines.
• Our results generalize over both the datasets under considerations — see Table 1 for Crisis Humanitarianism and Table 2 for Sentiment Detection. We elaborate on these results next.

Effect of rule-based baseline dilutions: We start by noting that the insertion of random URL, keywords from image, text, and both together, are ineffective in decreasing the classification performance of multimodal classifiers considerably. However, inserting the most similar image's description to the original text substantially lowers the classification performance, from F_1 score of 0.734 to 0.642 (12.3% drop) for Crisis Humanitarianism dataset and from 0.793 to 0.665 (16.4% drop) for the Sentiment Detection dataset. This indicates that adding text corresponding to a similar image in the dataset is a reasonably effective dilution strategy. However, since the most similar image in the dataset could correspond to a different class, using its description as dilution frequently leads to less relevance and low topical coherence, as indicated by low values of **Sim**$_{text}$, **Sim**$_{img}$, and **KL Div**.

Effect of model-based baseline dilutions: Model-based baseline dilution strategies are generally more effective than rule-based dilution strategies in lowering the classification performance of the multimodal models. The drop in F_1 scores ranges from 9.6% (0.734 $\rightarrow$ 0.684) using GPT to 15.1% (0.734 $\rightarrow$ 0.628) using GPT-FT for the Crisis Humanitarianism dataset. Similar trends are observed on the Sentiment Detection dataset. Since GPT-FT is fine-tuned on in-domain text, the inserted text demonstrates a higher relevance with the original text when compared to GPT alone. Similarly, consistently across the two datasets, the correspondence similarity and the topical similarity scores for GPT-FT based dilutions are better than those of GPT. While the caption generation-based dilution strategies are also effective, they show lower relevance with existing text and a higher topical difference due to domain mismatch. The generated captions are generic and do not cater to the domains of crises and sentiment. Given the performance of all the model-based baselines across all the metrics, we consider GPT-FT to be the most competitive baseline. Overall, these results demonstrate that model-based baseline dilutions, whether text-only (GPT and GPT-FT) or cross-modal (using SCST and XLAN caption generation models), severely affect the performance of multimodal classifiers but lack in terms of relevance and coherence.

Effect of proposed cross-modal dilutions: The cross-modal dilutions added using our approach lead to a drop in F_1 scores from 0.734 to 0.564 (23.3%) and from 0.793 to 0.614 (22.5%) for the Crisis Humanitarianism and Sentiment datasets, respectively. This is by far the most effective dilution strategy that also demonstrates high relevance with the original text and image, high correspondence similarity, and low topical difference. The observed trends are consistent across both datasets. The superior performance of XMD across all metrics can be attributed to several design choices. First, XMD is designed to exploit the model vulnerabilities by encouraging misclassification via an adversarial loss component in the training objective. Second, while dilution using GPT-FT only considers the

	Classification Performance ↓				Relevance ↑			Diversity ↓	Topical Diff. ↓
	$\mathbf{F_1}$	Prec.	Recall	Acc.	$\mathbf{Sim}_{text}$	$\mathbf{Sim}_{img}$	$\mathbf{Sim}_{corr}$	Self-BLEU	KL Div.
Original	0.734	0.742	0.725	0.828	–	0.292	0.999	0.048	–
Rule-based									
Random URL	0.705	0.747	0.672	0.817	0.467	–	0.967	–	–
Image KW	0.733	0.735	0.757	0.822	0.498	0.194	**0.989**	–	4.383
Text KW	0.736	0.744	0.736	0.823	0.703	**0.233**	0.991	–	**2.022**
Text + Image KW	0.706	0.716	0.702	0.824	0.656	0.232	0.988	–	4.618
Similar image's desc	0.642	0.624	0.677	0.751	0.597	0.204	0.962	0.049	10.104
Model-based									
GPT	0.684	0.693	0.677	0.783	0.562	0.221	0.971	0.081	9.251
GPT-FT	0.628	0.616	0.629	0.754	0.614	0.216	0.981	0.063	8.182
SCST Captions	0.666	0.696	0.663	0.774	0.502	0.200	0.979	–	11.443
XLAN Captions	0.673	0.703	0.677	0.782	0.534	0.218	0.980	–	10.733
XMD (Ours)	**0.564**	**0.571**	**0.552**	**0.718**	**0.715**	0.232	0.985	**0.035**	6.113

Table 1: **Results on the multimodal Crisis Humanitarianism dataset.** We evaluate dilution methods based on classification performance (lower values denote greater adversarial effectiveness of dilutions), relevance (higher similarity scores denote more relevance), diversity (lower Self-BLEU score denote more diverse sentences in generation), and topical differences (lower KL Divergence denotes better topical coherence). Our proposed method is compared against rule-based and model-based baselines.

	$F_1 \downarrow$	Relevance ↑			Topical Diff. ↓
		$\mathbf{Sim}_{text}$	$\mathbf{Sim}_{img}$	$\mathbf{Sim}_{corr}$	KL Div.
Original	0.793	–	0.314	0.999	–
Rule-based					
Random URL	0.773	0.538	–	0.960	–
Image KW	0.783	0.559	0.204	0.983	5.163
Text KW	0.792	0.834	0.231	**0.992**	**3.114**
Text + Image KW	0.774	0.689	0.231	0.988	5.877
Similar image's desc	0.665	0.611	0.268	0.963	11.980
Model-based					
GPT	0.691	0.620	0.274	0.978	11.638
GPT-FT	0.652	0.642	0.261	0.981	10.091
SCST Captions	0.671	0.553	0.251	0.971	13.427
XLAN Captions	0.651	0.588	0.261	0.979	14.612
XMD (Ours)	**0.614**	**0.795**	**0.298**	0.984	9.137

Table 2: **Results on the multimodal Sentiment Detection.** Similar trends on a different dataset reinforce the adversarial effectiveness of XMD while generating relevant and coherent dilutions. Complete results are presented in Appendix A.7.

original text as context, XMD generates text based on the keywords from both the original text and the image. This results in relatively higher relevance to the original text and the image. Finally, even though both GPT-FT and XMD are trained to generate in-domain text via task-specific fine-tuning, XMD exceeds in terms of topical similarity between inserted and original text (KL Div: 6.113 versus 8.182 for Crisis Humanitarianism dataset).

It is worth mentioning that our proposed method (XMD) also generates text with the highest diversity across generated sentences compared to all the baselines. This is demonstrated by lowest Self-BLEU scores in Tables 1 and 2. However, since the values for all the methods are consistently small, all the dilutions can be considered sufficiently diverse.

To summarize, we observe that deep multimodal classifiers are not overly sensitive to minor content dilutions like the insertion of random URLs or keywords from the original content. However, adding dilutions based on text-alone (GPT, GPT-FT) or cross-modal (Captions, XMD) causes a notable drop in the classification performance of multimodal models. To this end, our proposed XMD generates the most effective dilutions in terms of the observed drop in classification performance while maintaining relevance with the original image and text and topical coherence.

7 Analysis of Cross-Modal Dilutions

Next, we further analyze the dilutions generated by our proposed method (XMD). We focus on the Crisis Humanitarianism dataset for our analyses. In addition to the analyses presented here, we investigate the effect of the length of dilutions (i.e., number of inserted words) on classification performance and observe no notable difference in observed trends with similar dilution lengths; see Appendix A.10. In Appendix A.11, we analyze the sensitivity of quantified metrics with respect to variations in λ. Finally, we conduct a human evaluation to assess how realistic the *diluted* multimodal examples are when compared against *real* multimodal examples.

Subjective Assessment of Dilutions: Figure 3 shows examples of the dilutions generated by XMD from the Crisis Humanitarianism dataset along with dilutions obtained from the baselines.

To further assess the quality of generated dilutions, we conducted a survey on Amazon Mechan-

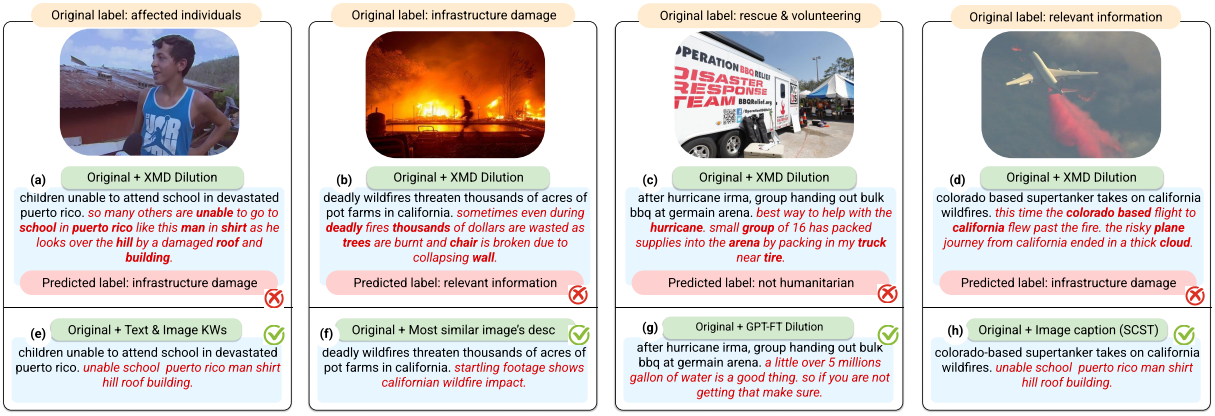

Figure 3: **Qualitative examples of XMD Dilutions and baselines.** *(a-d)*: Examples that are misclassified by the multimodal classifier *after* adding dilutions generated by XMD; the original examples were classified correctly. *(e-h)*: For each example, we also show what a baseline method would have added as a dilution which did not lead to incorrect classification. The original text is shown in black and the inserted dilutions are shown in red; extracted keywords are in bold.

ical Turk (AMT). We instructed the annotators to compare two multimodal posts — one containing dilutions from XMD and the other containing dilutions from GPT-FT for the same multimodal example. The posts were randomly ordered to mitigate position bias. Annotators were asked to respond on a 5-point Likert scale (1: strongly disagree, 5: strongly agree) to the following question: *Based on the quality of the text and its relevance with the image, is the post on the right more likely to be an actual social media post than the post on the left?* We obtained 5 annotations for each of the 200 examples that were randomly sampled from the test set of the Crisis Humanitarianism dataset. Overall, the results showed that annotators consider the dilutions generated by XMD to be more realistic than GPT-FT. The percentage of examples for which the majority of annotators preferred XMD dilutions over GPT-FT dilutions were: 32.1% (strongly) and 46.4% (moderately). For 12.2% examples, the majority of annotators were neutral and preferred GPT-FT dilutions over XMD dilutions for 9.3% examples. See Appendix A.8 for more details about human evaluation, recruitment, and compensation.

Ablations: We aim to understand the role of two key components in XMD – the incorporation of the adversarial loss component in the training objective and the inclusion of textual keywords. For the Crisis Humanitarianism dataset, Table 3 shows that *(a)* XMD without the adversarial loss component and without infusing keywords from the original text (i.e., XMD (Plain)) lacks in generating relevant and topically coherent dilutions. On adding

	Classification Performance ↓				Relevance ↑		Top. Diff. ↓
	F_1	Prec.	Recall	Acc.	Sim_{text}	Sim_{img}	KL Div.
XMD (Plain)	0.643	0.656	0.651	0.770	0.493	0.195	9.246
XMD (Adv)	0.624	0.632	0.621	0.739	0.483	0.193	9.275
XMD (Full)	0.564	0.571	0.552	0.718	0.715	0.232	6.113

Table 3: **Ablation Results for Crisis Humanitarianism Dataset. Plain** is trained with $\mathcal{L}_{gen}$ alone and only uses image keywords. **Adv** is trained using $\mathcal{L}_{gen} + \lambda\mathcal{L}_{adv}$. **Full** includes adversarial training + text & image keywords.

adversarial loss component to the objective (i.e., XMD (Adv)), the classification performance decreases further with little effect on relevance and coherence. Keeping the adversarial loss while infusing keywords from the original text (i.e., XMD (Full)) leads to the largest drop in classification performance while improving relevance (with both text and image) as well as topical coherence. Ablations on the Sentiment Detection task show the same trends; see Appendix A.9.

Are cross-modal content dilutions realistic? We now focus on assessing how *realistic* the diluted examples are when compared to the real social media examples. To this end, we conduct an AMT survey that requires users to compare multimodal examples with inserted dilutions against different but original multimodal examples. To prime the annotators, we first show them 5 unmodified multimodal examples and subsequently ask them to analyze a list of randomly-ordered multimodal examples, half diluted and the other half original unmodified examples. We use the dilutions generated by XMD. For each example in the list, the annotators are asked to respond to the following question on a

5-point Likert scale: *Do you think this post (text and image) could be a real post from social media website?* We select a subset of 100 examples from the test set of the Crisis Humanitarianism dataset and obtain 3 annotations for each example. The average Likert score for original examples is 3.61 ($\pm$ 0.53), whereas that for diluted examples is 3.38 ($\pm$ 0.39). The inter-rater agreement indicated strong reliability of annotations (Krippendorf's $\alpha = 0.83$). An independent two-sided t-test (assuming unequal variances) resulted in a p-value of 0.24, indicating no evidence that the average Likert scores of the original and diluted examples are from different distributions. These results show that the annotators asses the diluted and original examples to be similar, reinforcing the realistic nature of dilutions.

8 Conclusion and Future Work

In sum, our work is the first investigation of the robustness of multimodal classifiers to cross-modal dilutions. We establish the plausibility of such dilutions via human evaluations and develop a model to emulate adversarial scenarios reliably. We find that multimodal classifiers that fuse the state-of-the-art modality-specific representations are not robust to cross-modal dilutions generated by XMD.

Deep classifiers are increasingly being used for crucial applications that involve the joint understanding of user-generated multimodal data. Our broader goal in this work is to analyze and advocate for the robustness of multimodal models with societal applications, while focusing on the most representative fusion-based multimodal classification technique. In the future, we intend to leverage the knowledge of vulnerabilities identified in the current work to develop more robust multimodal models. We encourage interested researchers to investigate other cross-modal variations pertinent to multimodal data and assess the robustness of multimodal learning approaches to these variations.

9 Limitations and Broader Perspective

It is important to be clear about the limitations of this work. Our approach hinges on extracting informative keywords from both the image and the text to ensure the relevancy of the generated dilutions. In scenarios where the extracted keywords from images are generic (like celebrity faces for multimodal fake news detection) or the contextual relationship between image and text modalities is not straightforward (like multimodal hate speech), the proposed method does not generate semantically meaningful dilutions. We discuss the limitations in greater detail in Appendix A.12.

This work emphasizes the possibility that the lack of robustness of multimodal classification models can cause societal harm, such as delaying humanitarian interventions during crisis events. As such, the trained adversarial dilution generation models could be put to malicious use. We strongly condemn the misuse of this research. We release the code to aid reproducibility and promote future research on this topic. We believe that this research will encourage the community to investigate the robustness of multimodal classifiers and minimize real-world harm, leading to long-term benefits.

Bias of pre-trained models: It is known that pre-trained models used in our study demonstrate many biases (Bender, 2019; Hendricks et al., 2018; Garimella et al., 2021). This is often reflected in the kind of keywords that are identified in images and the resulting generated text (e.g., stereotypical gender associations). We acknowledge that the current state of deep learning research is limiting, and the consequential shortcomings are reflected in our work to some extent.

Annotations, IRB approval, and datasets: The annotators for this study were recruited via AMT. We specifically recruited 'Master' annotators located in the United States; and paid them at an hourly rate of 10 USD for their annotations. The human evaluation experiments were approved by the Institutional Review Board at Georgia Tech. The datasets used in this study are publicly available and were curated by previous research.

10 Acknowledgements

This research/material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112290102 (subcontract No. PO70745), NSF ITE-2137724, NSF ITE-2230692, Microsoft AI for Health, IDEaS at Georgia Institute of Technology, and Adobe Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the position or policy of DARPA, DoD, NSF, and SRI International and no official endorsement should be inferred. We thank Shivaen Ramshetty for sharing insights from related experiments during the rebuttal phase, the CLAWS research group members for their inputs, and the anonymous reviewers for their constructive feedback.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 346–353.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAAI conference on web and social media*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*, pages 806–810. Springer.

Rich Caruana, Steve Lawrence, and C Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *NeurIPS*.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903.

Dataminr. 2020. Multi-Modal Fusion AI for Real-time Event Detection.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *ACM WebSci*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*.

Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1731–1740.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Chi Thang Duong, Remi Lebret, and Karl Aberer. 2017. Multimodal classification for analysing social media.

Facebook. 2020. Hateful Memes Challenge and dataset for research on harmful multimodal content.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. 2018. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*.

Akshay Kumar Gupta. 2017. Survey of visual question answering: Datasets and techniques. *arXiv:1705.03865*.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2021. Petgen: Personalized text generation attack on deep sequence embedding-based classification models. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 575–584.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web conference*, pages 2915–2921.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Thai Le, Suhang Wang, and Dongwon Lee. 2020. Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 282–291. IEEE.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B Dolan. 2021a. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069.

Linjie Li, Zhe Gan, and Jingjing Liu. 2020a. A closer look at the robustness of vision-and-language pretrained models. *arXiv preprint arXiv:2012.08673*.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021b. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2042–2051.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. Bert-attack: Adversarial attack against bert using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? *arXiv:1805.05492*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Pandi Nayak. 2021. MUM: A new AI milestone for understanding information.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*.

Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. 2020. Visual question answering using deep learning: A survey and performance analysis. In *International Conference on Computer Vision and Image Processing*, pages 75–86. Springer.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Gaurav Verma, Eeshan Gunesh Dhekane, and Tanaya Guha. 2019. Learning affective correspondence between music and image. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3975–3979. IEEE.

Gaurav Verma, Rohit Mujumdar, Zijie J Wang, Munmun De Choudhury, and Srijan Kumar. 2022. Overcoming language disparity in online content classification with multimodal learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1040–1051.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Dongxiang Zhang, Rui Cao, and Sai Wu. 2019. Information fusion in visual question answering: A survey. *Information Fusion*, 52:268–280.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and William B Dolan. 2020. Pointer: Constrained progressive text generation via insertion-based generative pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8649–8670.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.

A Appendix

A.1 Text-only Classifier Training

Before training, we pre-process the text in multimodal examples to remove URLs, emoticons, platform-specific tokens (like 'RT' for indicating retweets on Twitter), and symbols like @ and #. We also expanded negatives like *can't* and *won't* to '*can not*' and '*will not*'. To train the text classifier ($\mathcal{M}_{text}$), we fine-tune a pre-trained language model, DistilBERT (Sanh et al., 2019; Devlin et al., 2018), on the two datasets discussed in Section 4 by using the respective training sets. To train the text classification models for each dataset, we use Adam optimizer (Kingma and Ba, 2014) with a learning rate initialized at 10^{-4}; hyper-parameters are set by observing the classification performance achieved on the respective validation set. We use early stopping (Caruana et al., 2000) to stop training when the loss value on the validation set stops to improve for 5 consecutive epochs. The performance of the trained classifier on the test sets of Crisis Humanitarianism and Sentiment Detection datasets are presented in Table 4.

A.2 Image-only classifier

We apply a standard image pre-processing pipeline so that images with different dimensions can fit the pre-trained VGG-16 model's input requirement. First, we resize the image so that its shorter dimension is 224. We then crop the square region in the center and normalize the square image with the mean and standard deviation of the ImageNet images (Deng et al., 2009).

⋆ **Crisis Humanitarianism**	CLASSIFICATION PERFORMANCE			
	F_1	**Prec.**	**Recall**	**Acc.**
Text-only Classifier	0.713	0.725	0.703	0.801
Image-only Classifier	0.429	0.456	0.426	0.528
Multimodal Classifier	0.734	0.742	0.725	0.828
⋆ **Sentiment Detection**	CLASSIFICATION PERFORMANCE			
	F_1	**Prec.**	**Recall**	**Acc.**
Text-only Classifier	0.732	0.739	0.733	0.742
Image-only Classifier	0.941	0.948	0.946	0.953
Multimodal Classifier	0.793	0.797	0.798	0.802

Table 4: Performance of text-only and image-only classifiers on the Crisis Humanitarianism and Sentiment Detection tasks.

To train the image-only classifier ($\mathcal{M}_{image}$), we apply a fine-tuning approach to train the task-specific image classifiers. We first freeze the weights of VGG-16 (Simonyan and Zisserman, 2015), pre-trained on ImageNet (Deng et al., 2009), and then swap the last layer from the original model to three fully connected hidden layers with dimensions `4096`, `256`, and `num-of-classes`. Finally, we retrain these three layers to adapt the image distribution in each dataset. We use Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} for each dataset. To avoid overfitting, we use early stopping to stop training when the loss value on the validation set stops to improve for 10 consecutive epochs. Table 4 shows the performance of image-only classifier.

A.3 Keyword Extraction from YAKE

For extracting keywords from the original text, we use YAKE (Campos et al., 2018). We set the following hyper-parameters: maximum N-gram size = 1; de-duplication threshold = 0.9; de-duplication algorithm: 'seqm'; window size = 1, maximum number of keywords extracted from text = 5.

A.4 Baseline: Most similar image's desc.

We create this baseline to emulate the scenario where the user could have posted the multimodal example after diluting the original text by adding a highly similar image's description. We find the most similar image in the test set to an image of a given multimodal example and append its caption to the text in the given multimodal example. As mentioned in the main text, we use the cosine similarity between the VGG-16 embeddings obtained after task-specific for computing the similarity. Overall, most similar images were found to be highly similar, with an average highest similarity score of 0.767 with a standard deviation of 0.067. Nonetheless, as discussed in Section 6, this naïve dilution strategy frequently leads to irrelevant and topically incoherent.

A.5 Evaluation: Correspondence similarity

We explain the rationale behind adopting the correspondence similarity score (i.e., **Sim**$_{corr}$) as one of our evaluation metrics. For context, the cross-modal correspondence prediction task is a binary classification task that aims to classify two input modalities as corresponding or not. For instance, if an image and text that are parts of the same multimodal example are provided as input, the correct prediction is Label 1, indicating true correspondence. Conversely, if the input text and image are from different multimodal examples, the correct prediction is Label 0, indicating false correspondence. The correspondence prediction task has been widely adopted as a pre-training step for multimodal deep learning models (Arandjelovic and Zisserman, 2017; Verma et al., 2019; Feng et al., 2014). In this work, we train correspondence prediction models using the fine-tuned image and text representations of the dataset-specific *undiluted* training set, and then report **Sim**$_{corr}$ — the average probability score for Label 1 (i.e., true correspondence) on the *diluted* dataset-specific test set examples. Effectively, the score indicates that given a model trained to predict correspondence between image and text from original unmodified training examples; the model is successful in establishing a correspondence between diluted text and images in the test set examples.

To train the cross-modal correspondence prediction model, we create negative examples by randomly sampling 3 mismatched descriptions from the training set for each image with the correct description. We then take the fine-tuned representation of the input image and text and pass them through a series of fully-connected layers of sizes (1024 (input), 512, 256, 128, 64, 32, and 2 (output)). As shown in Tables 1 and 2, the correspondence prediction model provides a nearly-perfect **Sim** score (i.e., 0.999) on undiluted test sets. However, the scores for baselines and the proposed model differ based on the dilution strategy adopted.

A.6 Topical Coherence

To measure the topical coherence between generated dilution and the original text, we compute the KL Divergence between the topic distributions of the two text segments — i.e., $D_{KL}(P_{dilution}||Q_{original})$. We train an LDA topic model (Blei et al., 2003) using the text in a task-specific training set. The presented KL divergence scores are averaged over all the examples in the test set. We set the number of topics to be 20 (based on topic coherence score) for the results presented in this paper. Additionally, we do not witness a change in the observed trends with variations in the chosen number of topics ($n \in \{5, 10, 15, 20\}$) for LDA topic modeling.

For implementing the Self-BLEU metric for quantifying diversity, we use NLTK's BLEU score

	Classification Performance ↓				Relevance ↑			Diversity ↓	Topical Diff. ↓
	F_1	Prec.	Recall	Acc.	Sim_{text}	Sim_{img}	Sim_{corr}	Self-BLEU	KL Div.
Original	0.793	0.797	0.798	0.802	–	0.314	0.999	0.053	–
Rule-based									
Random URL	0.773	0.777	0.772	0.784	0.538	–	0.960	–	–
Image KW	0.783	0.782	0.794	0.796	0.559	0.204	0.983	–	5.163
Text KW	0.792	0.791	0.798	0.801	0.834	0.231	0.992	–	3.114
Text + Image KW	0.774	0.771	0.768	0.785	0.689	0.231	0.988	–	5.877
Similar image's desc	0.665	0.662	0.676	0.680	0.611	0.268	0.963	0.052	11.980
Model-based									
GPT	0.691	0.695	0.683	0.697	0.620	0.274	0.978	0.086	11.638
GPT-FT	0.652	0.664	0.661	0.668	0.642	0.261	0.981	0.074	10.091
SCST Captions	0.671	0.675	0.681	0.680	0.553	0.251	0.971	–	13.427
XLAN Captions	0.651	0.663	0.656	0.665	0.588	0.261	0.979	–	14.612
XMD (Ours)	0.614	0.617	0.626	0.633	0.795	0.298	0.984	0.047	9.137

Table 5: Complete results for the multimodal Sentiment Detection dataset. We observe the same trends as we do with the Crisis Humanitarianism dataset, demonstrating the generalizability of our approach.

	Classification Performance ↓				Relevance ↑			Topical Diff. ↓
	F_1	Prec.	Recall	Acc.	Sim_{text}	Sim_{img}	Sim_{corr}	KL Div.
XMD (Plain)	0.663	0.654	0.669	0.671	0.586	0.237	0.979	11.012
XMD (Adv)	0.652	0.644	0.651	0.655	0.571	0.232	0.964	11.157
XMD (Full)	0.614	0.617	0.626	0.633	0.795	0.298	0.984	9.137

Table 6: Ablation results for the multimodal Sentiment Detection dataset.

	# Words (std. dev.)	Control tech.	Updated F_1
Original	12.12 (3.94)	–	0.734
Rule-based			
Random URL	–	repeat 5 times	0.693
Image KW	+ 3.18 (1.97)	repeat 5 times	0.711
Text KW	+ 2.76 (0.51)	repeat 8 times	0.726
Text + Image KW	+ 5.85 (2.13)	repeat 4 times	0.681
Similar image's desc	+ 11.72 (3.59)	repeat twice	
Model-based			
GPT	+ 20.85 (8.64)	no change	0.684
GPT-FT	+ 22.87 (6.52)	no change	0.628
MM Captions	+ 9.11 (1.12)	repeat twice	0.657
XLAN Captions	+ 8.63 1.96	repeat twice	0.662
Proposed			
XMD (Plain)	+ 36.43 (10.16)	truncate text	0.649
XMD (Adv)	+ 39.46 (11.34)	truncate text	0.632
XMD (Full)	+ 37.97 (13.62)	truncate text	0.571

Table 7: Numbers of words inserted by the dilution methods and classification performance after controlling for the number of inserted words (all methods have ~20 words after modifications).

function (Loper and Bird, 2002) and adopt the approach proposed in Zhu et al. (2018).

A.7 Results on Sentiment Detection

The main text presents an abridged version of the results on the Sentiment Detection dataset. The complete results are presented in Table 5.

A.8 Human evaluation details

For both our annotation tasks, we recruited annotators using Amazon Mechanical Turk. We set the criteria to 'Master' annotators who had at least 90% approval rate and were located in the United States. The rewards were set by assuming an hourly rate of 10 USD for all the annotators. In addition, the annotators were informed that the aggregate statistics of their annotations would be used and shared as part of academic research.

The annotators were primed to identify real social media posts by showing them 5 original multimodal examples. Previous research has demonstrated the role of providing examples in obtaining high-quality annotations (Khashabi et al., 2021). For both our human evaluations, we also inserted some "attention-check" examples during the annotation tasks to ensure the annotators read the text carefully before responding. This was done by explicitly asking the annotators to mark a randomly-chosen score on the Likert scale regardless of the actual content. We discard the annotations from annotators who did not correctly respond to all the attention-check examples.

A.9 Ablations for Sentiment Detection

The ablation results on the Sentiment Detection dataset are presented in Table 6. The results follow the same trends as discussed in Section 8 for the Crisis Humanitarianism dataset.

A.10 Length of Dilutions

To examine whether the drop in performance is contingent on the number of words inserted for dilution, we first report the number of words inserted using each of these methods (see Table 7). Then, we control for the number of words inserted by employing either repetition or truncation so that each method inserts a comparable number of ~20 words for dilution. As shown in Table 7, even with comparable number of inserted words, the trends observed in Section 6 persist. This reinforces that

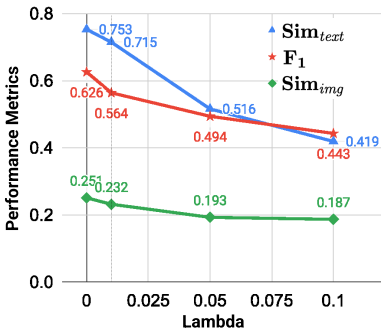

Figure 4: **Effect of varying λ.** As λ is increased, the adversarial effectiveness of the generated dilutions increases (lower F_1) but at the expense of relevance with original text & image (lower **Sim**$_{text}$ and **Sim**$_{img}$).

it is not merely the dilutions' length that precipitates the drop in classification performance but the sensitivity to the inserted content.

A.11 Effect of variations in lambda

Our main results and subsequent analyses are based on $\lambda = 0.01$, which controls the contribution of adversarial loss in the overall objective (see Equation 1). Figure 4 shows the variation in classification performance on the crisis humanitarianism dataset with respect to the variations in λ. We find that as λ increases, the classification performance deteriorates further. However, increasing λ hurts the relevance of the generated dilution with the original text and image, as well as the topical coherence – the relevance and coherence scores drop quickly as the relative contribution of $\mathcal{L}_{gen}$ is reduced.

A.12 Limitations

As indicated in Section 8, in some scenarios, the extracted keywords from the images could be generic and do not extract meaningful keywords towards the specific task at hand. For instance, for multimodal fake news detection, the extracted keywords from pictures of celebrity faces are typically: *man*, *woman*, *eye*, *smile*, *dress* etc. However, these keywords are unrelated to the larger (true/false) discourses centered around the celebrity. Similarly, for multimodal hate speech detection, the extracted keywords are often literal (such as *hat*, *clown*, *monkey*) while the original text aims to establish provocative parallels like calling a person clown or associating certain groups with animals. Our current work is best applied to settings where the contextual relationship between the visual and textual modalities is straightforward, and extracted keywords provide a good representation of the cumulative expression. As part of our future work, we intend to develop cross-modal dilution strategies that can work with a wider variety of user-generated multimodal data.