Vision and Language for Digital Forensics

Grace Luo

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-109
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-109.html

May 13, 2022

Vision and Language for Digital Forensics

Grace Luo

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Trevor Darrell
Research Advisor

5/4/2022

(Date)

* * * * * * *

Professor Hany Farid
Second Reader

May 6, 2022

(Date)

Vision and Language for Digital Forensics

by

Grace Luo

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Trevor Darrell, Chair
Professor Hany Farid

Spring 2022

Vision and Language for Digital Forensics

Abstract

Vision and Language for Digital Forensics

by

Grace Luo

Master of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Trevor Darrell, Chair

People frequently use the internet to transmit information about *events*. Whether we see a photo of a protest or a video of an airstrike, we often accept that the event occurred because of the visual evidence. However, what if this visual evidence were miscaptioned or taken out-of-context? Digital forensics, or the examination of the provenance and validity of online media, is a critical practice in fields such as human rights law, investigative journalism, and social media fact checking for this reason. Organizations manually verify textual claims about visual media via reverse image search and geolocation, which is an incredibly time consuming process. This report presents automated methods for verifying image-caption consistency, combining state-of-the-art vision-and-language neural models with real-world data relevant to digital forensics.

Chapter 1 discusses NewsCLIPpings, an approach for producing challenging instances of out-of-context images. Because such media is often unlabeled (and if detected, taken down by platform content moderators), our method can be used to benchmark and augment training data for automated verification methods.

Moving from news to social media, Chapter 2 produces out-of-context images in specific topical domains such as climate change and explores further techniques for automated verification, including methods for multimodal fusion and remedies for the domain shift between machine-made training data and human-made evaluation data. These chapters also give a glimpse into the outstanding challenges of multimodal digital forensics research, such as understanding the diverse set of text-image relationships present in social media or solving specific subtasks in the verification process such as geolocation.

To my family and friends.

Contents

Acknowledgments

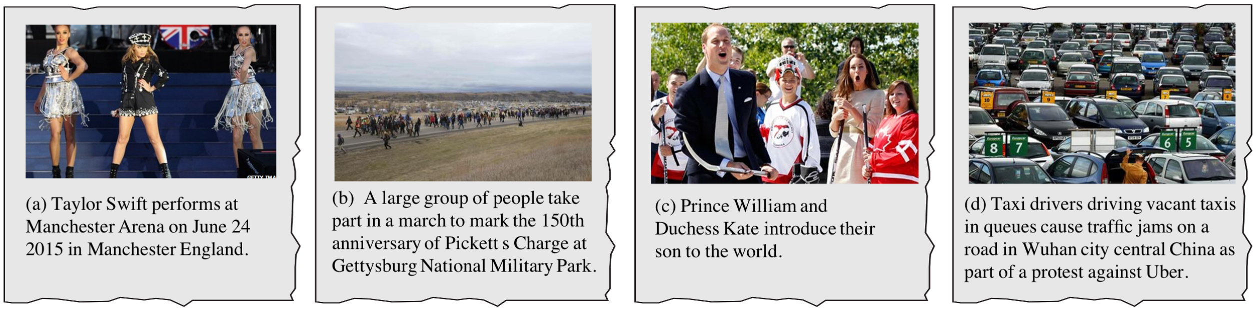
None of this would have been possible without my mentor and collaborator Dr. Anna Rohrbach, whose advice and guidance has shaped the works in this report and my growth as a researcher. I would like to thank my advisor, Professor Trevor Darrell, for supporting and advocating for me during my research journey, and Professor Hany Farid for reviewing this thesis. I would also like to thank Giscard Biamby, Lisa Dunlap, and Sanjay Subramanian for making the vision-and-language bay lively and always being available to brainstorm. I am grateful for my family and their love and support in my ambitions. Finally, I would like to appreciate Nima, Rosanna, and Carolyn and many others for the happiness, laughter, and memories from this year. Go bears!

Chapter 1

NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media

Online misinformation is a prevalent societal issue, with adversaries relying on tools ranging from cheap fakes to sophisticated deep fakes. We are motivated by the threat scenario where an image is used out of context to support a certain narrative. While some prior datasets for detecting image-text inconsistency generate samples via text manipulation, we propose a dataset where both image and text are unmanipulated but *mismatched*. We introduce several strategies for automatically retrieving convincing images for a given caption, capturing cases with inconsistent entities or semantic context. Our large-scale automatically generated NewsCLIPpings Dataset: (1) demonstrates that machine-driven image repurposing is now a realistic threat, and (2) provides samples that represent challenging instances of mismatch between text and image in news that are able to mislead humans. We benchmark several state-of-the-art multimodal models on our dataset and analyze their performance across different pretraining domains and visual backbones.[1]

[1] This chapter is based on joint work with Anna Rohrbach and Trevor Darrell presented at EMNLP 2021 [16].



(a) Taylor Swift performs at Manchester Arena on June 24 2015 in Manchester England.

(b) A large group of people take part in a march to mark the 150th anniversary of Pickett s Charge at Gettysburg National Military Park.

(c) Prince William and Duchess Kate introduce their son to the world.

(d) Taxi drivers driving vacant taxis in queues cause traffic jams on a road in Wuhan city central China as part of a protest against Uber.

Figure 1.1: Consider the following examples and guess whether these are pristine news or automatically matched image-caption pairs. The solution and more discussion are given in text.

1.1 Introduction

Misinformation has reached new heights as sophisticated AI-based tools have come into the spotlight. For instance, it has become easy to generate images of people who "do not exist"[2] and create realistic deepfakes of existing people [28]. Recent language models have become better at fooling people into believing that generated texts are from real people [8]. However, simple and cheap image repurposing remains a widespread and effective form of misinformation [6]. Specifically, real images of people and events get reappropriated and used out of context to illustrate false events and misleading narratives by misrepresenting *who* is in the image, what is the *context* in which they appear, or *where* the event takes place. This method is effective since augmenting a story with an image has been shown to increase user engagement and make false stories seem true [7]. Here, we explore whether such a threat can be *automated*. We show that real world images can be automatically matched to captions to generate false but compelling news stories, a threat scenario that may lead to larger-scale image repurposing.

While synthetic media conceptually could be detected by doing unimodal analysis (e.g. a detector for GAN-generated images), in our case both the text and the image are real. Thus, determining whether an image-caption pair is pristine or falsified requires joint multimodal analysis of the image and text (consider Figure 1.1 and make your guess).

Prior work has proposed several datasets related to our problem statement. One line of work obtains out-of-context image-text pairs by manipulating the named entities within the text [17, 22]. We find that in practice this may lead to linguistic inconsistencies, providing sufficient signal for a text-only model to distinguish between pristine and falsified descriptions without looking at the images. One recent work on detecting out-of-context images [1] focuses on a scenario where an image is accompanied by two captions (from two distinct news sources), and one has to establish whether the two captions are consistent. Here, we do not manipulate textual descriptions as we aim to minimize unimodal bias in

[2] https://thispersondoesnotexist.com

our task. We do not assume that two captions are available per image, rather we focus on classifying each image-caption pair as pristine or falsified.

Specifically, we propose a large-scale automatically constructed dataset with real and out-of-context news based on the VisualNews [15] corpus. We consider several threat scenarios, designing matches based on: (a) *caption-image similarity*, (b) *caption-caption similarity*, where we retrieve an image with similar semantics to a given caption while the named entities between the source and the target are disjoint, (c) *person match*, where we retrieve an image that depicts a person mentioned in the source caption but pictured in a different context, and (d) *scene match*, where we retrieve an image that has the same scene type as the source image but depicts a different event [3]. We use the recent powerful multimodal model CLIP [20] and other image and text models to construct the **NewsCLIPpings Dataset**. To make our dataset more challenging, we introduce an adversarial filtering technique based on CLIP.

We benchmark several state-of-the-art multimodal models and analyze their performance on the NewsCLIPpings Dataset. We investigate the impact of the pretraining domains and various visual backbones. We conduct a human evaluation that shows humans find it challenging to distinguish between pristine and falsified samples from our dataset. We also perform a qualitative analysis with the help of visual salience to shed light onto the useful cues discovered by the models trained on our dataset. Our dataset is publicly available here: https://github.com/g-luo/news_clippings [4].

1.2 Related Work

We review several most relevant datasets in detail. Some earlier proposed datasets for detecting multimodal misinformation are MultimodAl Information Manipulation dataset (MAIM) [12] and Multimodal Entity Image Repurposing (MEIR) [22]. MAIM naively matches images to captions from other random images to create their falsified versions. MEIR introduces swaps over named entities for people, organizations and locations. One of their assumptions is that for each image-caption "package" there is an unmanipulated related package (geographically near and semantically similar) in the reference set. This allows verifying the integrity of the query package by first retrieving a related package and then comparing the two. This problem statement is different from ours, as we do not assume availability of a perfect reference set.

A more recent work has proposed TamperedNews [17], a dataset where named entities specific to people, locations and events are swapped to other random[5] named entities within

[3] This answers the question in Figure 1.1, i.e., examples a), b), c), d) correspond to these four threat scenarios in our dataset, so all four are falsified.

[4] Specifically, we provide pristine and falsified matches for captions/images, i.e. their identifiers within the VisualNews dataset. The copyright and usage rights of the data are subject to that of [15].

[5] With some constraints, such as individuals of the same country and gender or locations within the same region.

the article body. We show that such text manipulations lead to significant linguistic biases and the corresponding tasks can be solved without looking at the images (See Section 2.4 for more details). Another recent work [1] aims to detect when images are used out of context, somewhat similar to MEIR above. They collect a dataset where each image appears in two distinct news sources and thus is associated with two captions. Most of the collected data is not labeled, but a small subset has been manually annotated as in- or out-of-context. Their problem statement (analyzing image and two captions) is again different from ours. One other work [25] tackles Neural News generation by replacing real articles with Grover [30] generated text and real captions with synthetic ones. They do not mismatch the images, which remain relevant to the article's content. The impact of image analysis on this task is rather limited, while analyzing the captions and the article body is key to the best detection performance. Finally, some work focuses on human-made fake news detection, such as FakenewsNet [24] and Fakeddit [18], etc. While these datasets contain important real world examples of fake news, our focus is on exploring an *automated* threat scenario, where an image is automatically retrieved to match a given caption.

1.3 The NewsCLIPpings Dataset

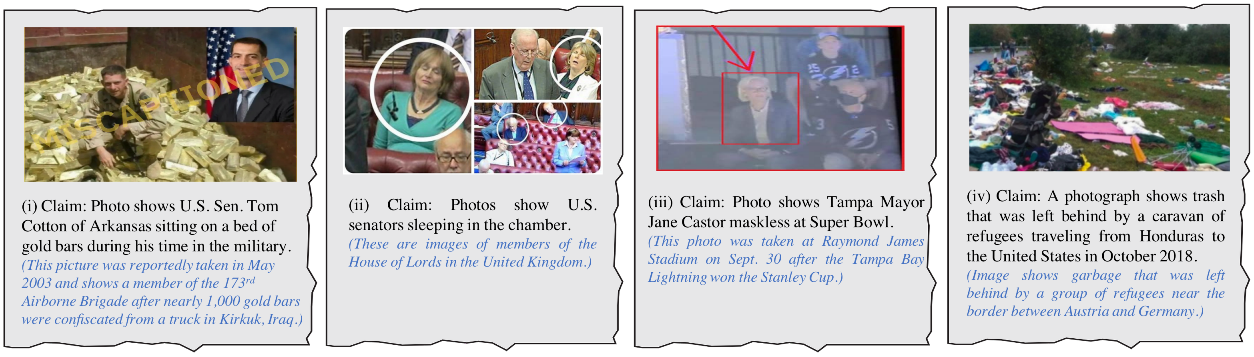
The objective of this work is to explore techniques for creating challenging, non-random image-caption matches that require fine-grained semantic and entity knowledge. As seen in Figure 1.2[6], misinformation in the wild is often extremely subtle and much more difficult than the random matches provided in prior synthetic datasets. In fact, general models that were not specifically trained or finetuned on the news domain can "solve" random news matches. We found that CLIP was able to achieve 97.39% Top-1 accuracy on a caption-image retrieval task with news images[7]. For comparison, a recent method TRIP [26] reports a Top-1 accuracy of 73.78% on a similar task. As a result, we construct several splits that model specific threat scenarios seen in the real world, and we use CLIP ViT-B/32 *off-the-shelf* to filter out the less challenging samples.

In the following, we assume we have a pristine query pair (img_1, cap_1) and retrieve another pair (img_2, cap_2) to form a falsified pair (img_2, cap_1).

Preprocessing Our dataset is derived from VisualNews [15], a large-scale corpus which contains image-caption pairs from four news agencies (The Guardian, BBC, USA Today, and The Washington Post). We use spaCy NER [9] to label named entities in captions and the Radboud Entity Linker (REL) [11] to link them to their Wikipedia 2019 entries. We compute text embeddings using SBERT-WK [29] and CLIP [20]. We compute image embeddings with Faster R-CNN [21] and CLIP. We use a ResNet50 classifier trained on the Places365 dataset [31] to get scene embeddings from images. We ensure that all matched

[6] Examples found on `https://www.snopes.com` and `https://www.politifact.com`.

[7] We ran this on a random 40k subset of VisualNews and counted how often CLIP selected the true image vs. four random negative images.



(i) Claim: Photo shows U.S. Sen. Tom Cotton of Arkansas sitting on a bed of gold bars during his time in the military. *(This picture was reportedly taken in May 2003 and shows a member of the 173rd Airborne Brigade after nearly 1,000 gold bars were confiscated from a truck in Kirkuk, Iraq.)*

(ii) Claim: Photos show U.S. senators sleeping in the chamber. *(These are images of members of the House of Lords in the United Kingdom.)*

(iii) Claim: Photo shows Tampa Mayor Jane Castor maskless at Super Bowl. *(This photo was taken at Raymond James Stadium on Sept. 30 after the Tampa Bay Lightning won the Stanley Cup.)*

(iv) Claim: A photograph shows trash that was left behind by a caravan of refugees traveling from Honduras to the United States in October 2018. *(Image shows garbage that was left behind by a group of refugees near the border between Austria and Germany.)*

Figure 1.2: We are motivated by the real-world examples of images used out-of-context. Here we include *real* misinformation examples found online[6] which closely resemble the four threat scenarios in our dataset.

samples are at least 30 days apart and that (cap_1, cap_2) have no overlapping named entities identified by spaCy and REL to prevent true matches, with the exception of the Person split, where we expect at least one "PERSON" entity to match.

Query for Semantics Our first split models a threat scenario that queries for specific semantic content, with the intent to portray the subjects of the image as certain other named entities, see Figure 1.2 (i, ii). We consider two ways of getting the matches. **(a) CLIP Text-Image**: We rely on the state-of-the-art CLIP representation to retrieve samples with the highest CLIP text-image similarity between (img_2, cap_1). **(b) CLIP Text-Text**: We match samples with the highest CLIP text-text similarity between (cap_1, cap_2) and retrieve the corresponding img_2. See examples (a) and (b) in Figure 1.1.

Query for Person This split models a threat scenario that queries for a specific person, with the intent to portray them in a false context, as in Figure 1.2 (iii). We ensure that the person of interest is pictured: all considered samples must have "PERSON" entities in their captions and a person related Faster-RCNN bounding box detected in the image. To avoid cases where the query person is mentioned but unlikely to be pictured, we filter captions where the person is in the possessive form, the object of the sentence, or modify a noun as determined by spaCy's dependency parser. We ensure that the context is distinct: the Places365 ResNet similarity must be less than 0.9. Finally, we found that there were a number of unsolvable falsified samples where the caption could be plausibly matched with any image of the person of interest. We minimize the number of such "generic" captions: we finetune a BERT [5] model on a small labelled subset of our training data to filter these captions from our matching process.**(c) SBERT-WK Text-Text**: We match samples that mention the query person based on the lowest semantic similarity measured by their SBERT-WK score, a text-only sentence embedding. See example (c) in Figure 1.1.

Query for Scene This split models a threat scenario that queries for a specific scene, with the intent to mislabel the event, see Figure 1.2 (iv). All samples must have no "PERSON"

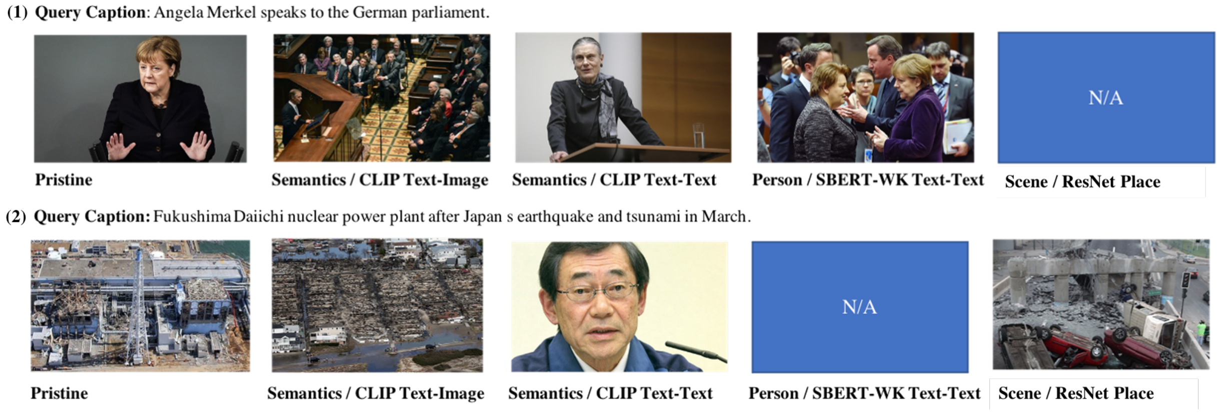

(1) Query Caption: Angela Merkel speaks to the German parliament.
N/A
Pristine
Semantics / CLIP Text-Image
Semantics / CLIP Text-Text
Person / SBERT-WK Text-Text
Scene / ResNet Place
(2) Query Caption: Fukushima Daiichi nuclear power plant after Japan s earthquake and tsunami in March.
N/A
Pristine
Semantics / CLIP Text-Image
Semantics / CLIP Text-Text
Person / SBERT-WK Text-Text
Scene / ResNet Place

Figure 1.3: Comparison of the retrieved matches for the same query caption obtained within our four splits.

named entities in the captions. This aims to filter headshots and other images with little scene information. **(d) ResNet Place:** We match samples with the highest Places365 image similarity, as determined by the dot product of their ResNet embeddings. See example (d) in Figure 1.1.

Merged Split This split mixes samples from all the splits to model a more realistic case where a variety of methods are used to generate out-of-context images, i.e. all types of mismatch may be encountered at test time. We merge the splits such that there is an equal number of samples from every split, and the captions and images across splits are disjoint.

Adversarial CLIP Filtering In the initial version of our dataset, we observed a distributional shift of CLIP Text-Image scores between the pristine (img_1, cap_1) and falsified (img_2, cap_1) samples. This makes sense, since it is not always possible to find a falsified image that is more convincing than the original. To reduce the difference between the two distributions, we use CLIP Text-Image similarity to adversarially filter our splits. For each pristine sample (img_1, cap_1) with CLIP Text-Image similarity CTI_p we have two options: (1) There may exist a set of falsified candidates (img_2, cap_1), where their score $CTI_f \geq CTI_p$, ordered for each of our splits: using (a) CLIP Text-Image, (b) CLIP Text-Text, (c) SBERT-WK Text-Text, (d) ResNet Place, respectively. (2) There exists a set of candidates where their score $CTI_f < CTI_p$, ordered in the same way. We select the top scoring sample from set (1), else we select the top sample from (2) if set (1) is empty. Finally, we remove the sample with $max(CTI_p - CTI_f)$ until we get a 50-50 ratio of samples from sets (1) and (2) since the larger the delta $CTI_p - CTI_f$ the more likely the falsified sample is of low quality. As a result, on a ranking task where CLIP off-the-shelf is given a caption and two images, it correctly chooses the pristine image 50% of the time by design.

NewsCLIPpings Dataset Statistics The detailed statistics for the proposed NewsCLIPpings Dataset are reported in Table 1.1. Each caption appears twice, once in a pristine sample

Table 1.1: NewsCLIPpings Dataset Statistics.

Split	Train	Val	Test
(a) Semantics/CLIP Text-Image	453,128	47,248	47,288
(b) Semantics/CLIP Text-Text	516,072	53,876	54,164
(c) Person/SBERT-WK Text-Text	17,768	1,756	1,816
(d) Scene/ResNet Place	124,860	13,588	13,636
Total/Sum	1,111,828	116,468	116,904
Total/Unique	816,922	85,609	85,752
Merged/Balanced	71,072	7,024	7,264

then again in a falsified sample. Thus exactly half of the samples are pristine and half are falsified, and there is no unimodal text bias in the dataset. We report the total number of samples across splits *including any duplicates* as Total/Sum, and the number of *unique* text-image pairings as Total/Unique in Table 1.1.

Table 1.2 provides a comparison to the most related prior datasets, highlighting the key differences, such as the image-text mismatch procedure used in each dataset.

Table 1.2: Comparison to prior related datasets. Size is the total number of unique samples across all splits.

Dataset	Data	Source	Mismatch	Size
MAIM [12]	Caption, Image	Flickr	Random	239k
MEIR [22]	Caption, Image, GPS	Flickr	Text entity manipulation	57k
TamperedNews [17]	Article, Image	BreakingNews	Text entity manipulation	776k
COSMOS [1]	Caption, Image	News Outlets	Two sources (3k labeled)	453k
NewsCLIPpings (Ours)	Caption, Image	VisualNews	Automatic retrieval	988k

Dataset Examples Here, we provide a few samples from the NewsCLIPpings Dataset. Figure 1.3 compares the matches from each split for the same query caption and pristine image. Our diverse methods of computing similarity result in different weightings for concepts, displaying the realm of plausible images for a given caption. In (1), CLIP Text-Image matches "parliament" to Tennessee's governor speaking to a General Assembly, CLIP Text-Text matches "Angela Merkel" to Ingeborg Berggreen-Merkel speaking, and SBERT-WK Text-Text finds a match of Angela Merkel at a summit. In (2), CLIP Text-Image matches "tsunami" to a flooding in New York, CLIP Text-Text matches "Japan" to the president of a Japanese company, and ResNet Place matches "earthquake" to a destroyed highway after an earthquake in Chile[8].

1.4 Experiments

We start by describing our experimental setup and then present the results of our benchmarking study.

Experimental Setup

Model Architectures For our base models we rely on CLIP [20] and VisualBERT [13]. We include VisualBERT as it is a representative recent model and is an appropriate baseline for addressing the semantic mismatch tasks.

CLIP passes image and text through separate encoders that are trained to generate similar representations for related concepts. The model is pretrained on a web-based corpus of 400M image-text pairs using a contrastive loss, in which the cosine similarity of true image-text pairs is maximized.

VisualBERT passes image and text through a shared series of transformer layers to align them into one embedding space. For its bounding box features, we use a Faster-RCNN model [21] trained on Visual Genome with a ResNeXT-152 backbone. For pretraining, we only use the Masked Token Loss reported by Li et al. [13], which masks each text token with probability 0.15. We pretrain VisualBERT either on the 3M image-caption pairs from Conceptual Captions [23], based on alt-texts from web images stripped of all named entities, or on the 1M pairs from the VisualNews [15], based on captions from the news images.

Implementation Details Our task is to *classify* each image-caption pair as pristine or falsified. We fine-tune both models as we train the classifiers. When finetuning, we use a learning rate of 5e-5 for the classifier and 5e-7 for other layers. We train with a batch size of 32 for 88k steps for the Semantics splits and 44k steps for the Person and Scene splits. We report *classification accuracy* over all samples (All) and separately for the Pristine and Falsified samples. We also report model performance at varying false alarm rates via ROC curves.

Experimental Results

In this section, we benchmark several methods on our proposed dataset to assess its difficulty. First, we compare the performance of unimodal vs. multimodal models to ensure that methods cannot exploit unimodal biases. Next, since we leverage CLIP ViT/B-32 to make our dataset challenging, we explore whether our task could be solved by a different model specifically pretrained on the news domain, leveraging a different backbone, or with more

[8] The Person split and Scene splits have no shared pristine samples since all matches either do or do not have 'PERSON" named entities depending on the split.

Table 1.3: Classification performance on the test set for the following models: (I) Image-only CLIP (w/ ViT-B/32), (II) Multimodal CLIP (w/ ViT-B/32), (III) VisualBERT-CC pretrained on the Conceptual Captions dataset, (IV) VisualBERT-VN pretrained on the Visual News.

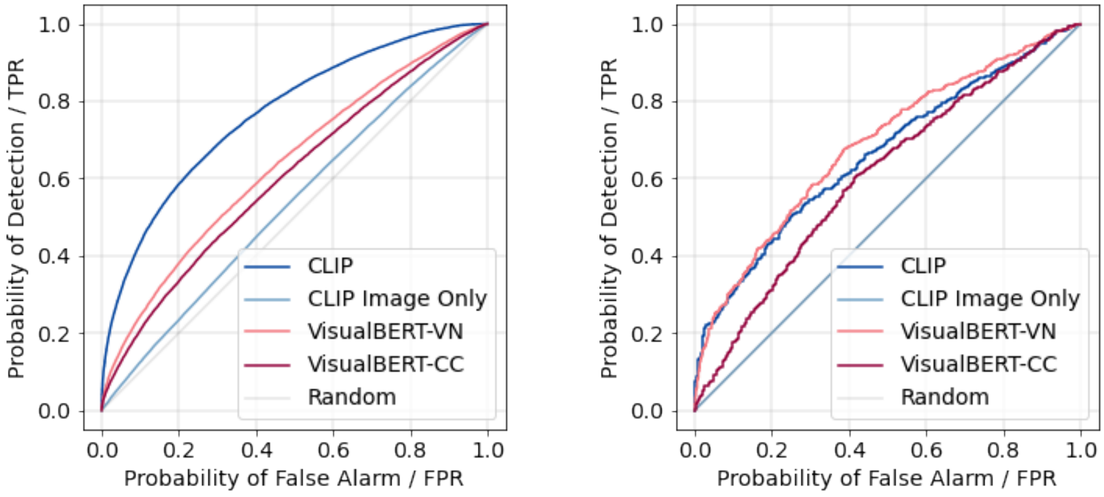
	(I) CLIP Image-Only	(II) CLIP			(III) VisualBERT-CC	(IV) VisualBERT-VN		
Split	All	All	Pristine	Falsified	All	All	Pristine	Falsified
(a) Semantics/CLIP Text-Image	0.5471	0.6698	0.7543	0.5853	0.5413	0.5774	0.6770	0.4778
(b) Semantics/CLIP Text-Text	0.5247	0.6939	0.7409	0.6469	0.5714	0.5949	0.6591	0.5307
(c) Person/SBERT-WK Text-Text	0.5000	0.6101	0.6178	0.6024	0.5947	0.6333	0.7247	0.5419
(d) Scene/ResNet Place	0.5391	0.6821	0.7835	0.5807	0.5636	0.6112	0.6693	0.5532
Merged/Balanced	0.5288	0.6023	0.7007	0.5039	0.5482	0.5863	0.7841	0.3885

model parameters. In our final experiment, we train a single model on the union of all splits (Total/Sum in Table 2.3), while all the other experiments report the performance of the *distinct models trained on each split individually.* All tables in this section evaluate on the same test set per split.

Unimodal Model Performance One motivation for this work is that several prior works rely on automatic text manipulation to generate mismatched media. We argue that entity manipulation can introduce linguistic biases. We trained a text-only BERT model [5] on *just the named entities* of the TamperedNews dataset (rather than the full articles) and achieved comparable results to the original paper's image-and-text based system [17]. For their "Document Verification" task, where the goal is to select one out of two articles given an image, we were able to achieve 90% versus their 93% on the Persons Country Gender (PsCG) split. For the Outdoor Places City Region split, GCD(25, 200), we were able to achieve 96% versus their 76%. This suggests that text manipulation can introduce biases that make the use of images unnecessary.

To avoid unimodal biases, our dataset is *balanced* with respect to its captions (every caption is used once in a pristine sample and again in a falsified sample). Since we do not have such constraint on our images, we ran an image-only CLIP model (i.e. zeroing out the text inputs to CLIP) to verify that there is minimal visual bias. Based on our findings and due to the smaller size of the Person split (c), we additionally balance this particular split with respect to images, which means any image-only model is expected to achieve exactly 50% accuracy on this split. As shown in Table 1.3 (I), overall the image-only CLIP model obtains slightly above chance performance, significantly lower than the full image-text model, Table 1.3 (II).

Multimodal Model Performance We report results for the multimodal CLIP-based classifiers in Table 1.3 (II). (Again, we repeat that here we train distinct classifiers for each split individually.) CLIP tends to "over-predict" pristine labels, indicating that many falsified



(a) Semantics/CLIP Text-Text ROC Curve

(b) Person/SBERT-WK Text-Text ROC Curve

samples are highly realistic and plausible. The Person split appears the most challenging, which could be partly explained by having the least number of samples. The Merged split, which contains an equal proportion of all four splits, is as difficult as its most difficult subsplit, seen by how CLIP classifies correctly 60% of the time compared with 61% for the Person split.

On the other hand, VisualBERT-CC (pretrained on Conceptual Captions) in Table 1.3 (III) performs the best on the Person split with a performance of 59% that approaches CLIP's. This indicates that the Person split primarily requires semantic understanding, and that a model with no knowledge of named entities can compete with a model that is strong at recognizing celebrities and other named entities. As expected, on all other splits that test entity understanding VisualBERT-CC performs on average 10% worse than CLIP.

Pretraining VisualBERT on News Domain We also compare the performance of VisualBERT pretrained on Conceptual Captions (VisualBERT-CC) vs. VisualNews (VisualBERT-VN). In Table 1.3 (III vs. IV) we observe that in-domain data, including named entities, provides a 3-5% boost uniformly across all splits. Even more, with a training corpus less than 1% the size of CLIP's training data, VisualBERT-VN is able to exceed CLIP performance on the Person split and approach CLIP performance on the Merged split. In fact, the largest gap between VisualBERT-VN and CLIP remains in the Semantics splits, where named entity understanding is crucial. Hence, through these results we can observe that VisualBERT-VN is strongest at semantic reasoning while CLIP is strongest at named entity recognition, which makes sense given their architectures (more deeply interactive VisualBERT-VN vs. more shallow CLIP).

ROC Curves We also include the ROC curves for the softmaxed logits produced by the

Table 1.4: Comparing different CLIP backbones, classification performance (test set).

Split	Model	All	Pristine	Falsified
(a) Sem/CLIP T-I	ViT-B/32	0.6698	0.7543	0.5853
	RN50	**0.6824**	**0.7461**	**0.6188**
	RN101	0.6765	0.7444	0.6085
(b) Sem/CLIP T-T	ViT-B/32	0.6939	0.7409	0.6469
	RN50	0.7182	0.7486	0.6878
	RN101	**0.7244**	**0.7442**	**0.7046**
(c) Per/SB-WK T-T	ViT-B/32	0.6101	0.6178	0.6024
	RN50	0.6123	0.7357	0.4890
	RN101	**0.6393**	**0.7004**	**0.5782**
(d) Scene/RN Place	ViT-B/32	0.6821	0.7835	0.5807
	RN50	0.7004	0.7765	0.6244
	RN101	**0.7137**	**0.7712**	**0.6562**
Merged/Balanced	ViT-B/32	0.6023	0.7007	0.5039
	RN50	0.6162	0.6836	0.5487
	RN101	**0.6597**	**0.6768**	**0.6426**

models, see Figure 1.4a, 1.4b. We see that the trends for these curves are consistent with the model rankings recorded in Table 1.3, with CLIP outperforming the other models by a wide margin on Semantics/CLIP Text-Text in Figure 1.4a across all false alarm rates. For Person/SBERT-WK Text-Text in Figure 1.4b, VisualBERT-VN has virtually identical performance as CLIP at low false alarm rates. However, for systems that can tolerate more false alarms VisualBERT-VN shows a small advantage.

Comparing CLIP Models Recall that we used CLIP ViT-B/32 to construct our dataset. Here, we investigate whether our dataset could be "solved" by an existing CLIP model with a different backbone (ViT-B/32 vs. RN50) or a bigger model (RN50 vs. RN101). In Table 1.4, we observe that RN50 performs slightly better than ViT-B/32 across the board, with at most 2% performance difference between the two. We also see that while RN101 has more parameters than RN50, it only provides a small 1-2% boost on most splits. The split where RN101 achieves the largest improvement (4%) is the Merged/Balanced split, which aligns with the need for a model to capture more complex patterns to classify samples from multiple generation methods. Although we used a specific CLIP model architecture during dataset generation, we see that our dataset is still challenging for models with different architectures and more parameters.

Evaluating A Single Unified Model Finally, we explore whether it is beneficial to combine various splits during training. Unlike Tables 1.3, 1.4 which evaluate *separate* models trained on each individual split, here we evaluate a *single* model trained on all the splits

Table 1.5: CLIP (ViT/B-32) test set classification performance when training a single model with all the available training samples, i.e. Total / Sum in Table 1.1.

Split	All	Pristine	Falsified
(a) Semantics/CLIP Text-Image	0.6651	0.7582	0.5720
(b) Semantics/CLIP Text-Text	0.6457	0.7563	0.5351
(c) Person/SBERT-WK Text-Text	0.6399	0.7434	0.5363
(d) Scene/ResNet Place	0.6824	0.7778	0.5870
Merged/Balanced	0.6611	0.7574	0.5647

jointly, see Table 1.5. The Total/Sum set (introduced in Table 1.1) combines the samples from all the splits, so that it is balanced with respect to pristine and falsified labels but has different proportions of each type, e.g. around 87% of samples are from the Semantics splits (a,b).

Comparing Table 1.3 (II) with Table 1.5, we note that the Person split experiences a 2% boost in performance even though it represents only 1% of the training data. Clearly, it benefits from the other sample types. We also note the 5% degradation in performance for the Semantics/CLIP Text-Text, likely due to the challenges in learning to address several mismatch types at once[9]. Finally, we see a boost of almost 6% for the Merged/Balanced set, showing the benefit of training in a unified setting for this more realistic split. One other trend we notice is that the Pristine accuracy seems to overall benefit more than the Falsified accuracy.

1.5 Additional Analysis

In this section, we gain further insights into the quality of our dataset via human evaluation and saliency map analysis. With the human evaluation, we assess whether our dataset could fool humans and pose a realistic threat. We also assess whether our dataset may have "unsolvable" true matches that in fact do not misrepresent anything. With our qualitative saliency map analysis, we investigate if the automatic models are learning to leverage high level semantic or entity cues after training on our dataset.

Human Performance Here, we estimate the difficulty of the proposed task for *humans*, aiming to assess how convincing our automatically matched images and captions are. We randomly select a set of 200 samples from the Merged/Balanced split, with an equal number of samples from all types (50 from each split, where 25 are pristine and 25 are falsified). We

[9] We hypothesize that this may be due to the joint training with the Person samples – if a model does not know who the pictured individual is, then the mismatches in Semantics/CLIP Text-Text may look similar to those in the Person split, as they both are matched using only textual information.

Table 1.6: Human Performance on 200-sample subset of Merged/Balanced. “Optimistic” accuracy is defined as at least 1 worker gave the correct answer.

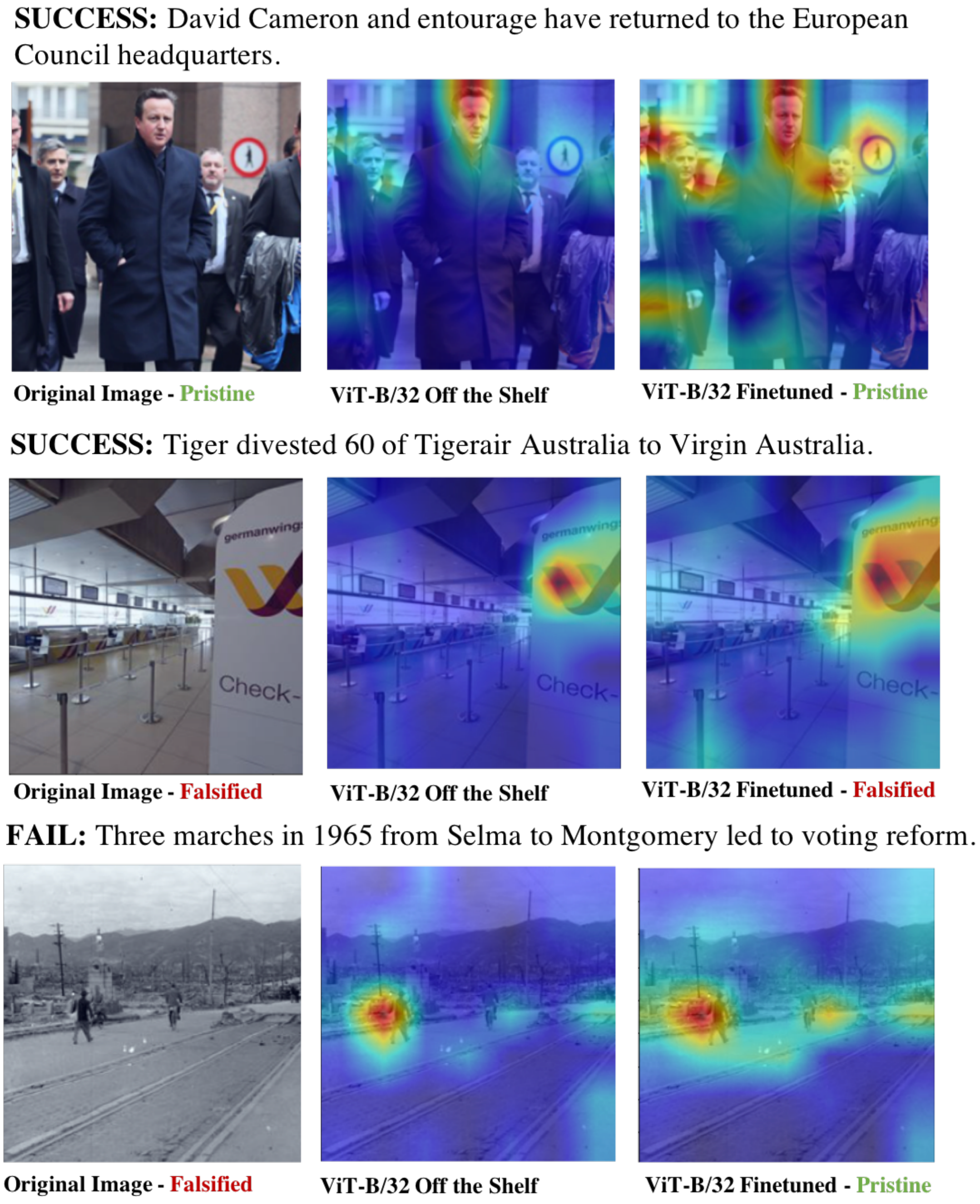
	All	Pristine	Falsified
Average	0.656	0.962	0.350
Optimistic	0.845	1.000	0.690

conduct our evaluation on Amazon Mechanical Turk[10]. For each image-caption pair we ask 5 workers the following three questions: (a) “Could this image belong to the given caption?” (Yes/No), (b) “How confident are you in your answer?” (1: Very, 2: Somewhat, 3: Not at all), (c) “Would it help to use a search engine to be more confident?” (Yes/No). Note, that we specifically instruct the workers **not** to use search engines, to prevent them from discovering the original news articles on the Web. The key takeaways from the evaluation are as follows. (1) The average accuracy over all samples is 0.656, while the most “optimistic” accuracy (at least 1 worker gave the correct answer) is 0.845. This clearly shows that the task is not easy for humans. For reference, our CLIP model trained on the Total/Sum set (Table 1.5) achieves 0.6650 on these 200 samples, essentially *matching human performance.* (2) Humans are much better in recognizing pristine than falsified samples, with an average accuracy of 0.962 and 0.350 respectively. This shows that they are often misled by our falsified matches. (3) The “optimistic” accuracy for falsified samples is 0.690, meaning that majority are still solvable with just the prior knowledge of those workers. Among the 31 samples that all workers classified incorrectly, we estimate that 67% are answerable with additional knowledge of person identity and other context cues. (4) While the average confidence score is 1.755, the confidence on the correctly vs. incorrectly predicted samples is 1.658 and 1.940 respectively (lower is better), i.e. humans were more confident on samples they predicted correctly. (5) The average accuracy when there is a reported “need to use a search engine” is 0.589 vs. 0.760 otherwise. This shows that the humans do better when they encounter familiar concepts vs. less familiar ones. Hence, additional search is likely to boost the results, as we have observed in our own internal analysis. (6) Across the four types of mismatch, the easiest for humans is Scene, followed by Semantics/CLIP Text-Text, Semantics/CLIP Text-Image, and finally the Person split. Interestingly, this overall aligns with the trends observed for the automatic methods.

Qualitative Analysis Finally, we analyze CLIP ViT saliency maps and prediction using the method presented in Chefer, Gur, and Wolf [4]. We select examples from the 200 samples used in our human evaluation.

As seen in Figure 1.5, finetuning on our dataset often forces CLIP to focus on salient

[10] www.mturk.com



SUCCESS: David Cameron and entourage have returned to the European Council headquarters.

Original Image - Pristine
ViT-B/32 Off the Shelf
ViT-B/32 Finetuned - Pristine

SUCCESS: Tiger divested 60 of Tigerair Australia to Virgin Australia.

germanwings
Check-
Original Image - Falsified
ViT-B/32 Off the Shelf
ViT-B/32 Finetuned - Falsified

FAIL: Three marches in 1965 from Selma to Montgomery led to voting reform.

Original Image - Falsified
ViT-B/32 Off the Shelf
ViT-B/32 Finetuned - Pristine

Figure 1.5: Qualitative examples of CLIP ViT-B/32 success and failure cases with salience visualization.

objects mentioned in the caption beyond the person of interest, for example expanding its attention from David Cameron to “entourage.“ CLIP also has a number of capabilities off-the-shelf that require minimal finetuning, for example sign reading and logo recognition in the case of distinguishing “Tigerair Australia” from “Germanwings.” We also present a failure case where CLIP focuses on the two people in the foreground when the caption talks about “marches.“ Evidently the model does find one point of support – the photo looks like it could have been taken in 1965 – but it fails to identify the absence of a crowd which would have been a “red flag” for a human. Similar failure cases, such as a falsified caption that mentions Mitt Romney and Rand Paul at a rally but only pictures Romney, highlight how our dataset can be particularly challenging because pristine news is an ambiguous domain that often mentions entities but does not picture them.

1.6 Conclusion

We introduced **NewsCLIPpings**, a large-scale automatically constructed dataset for classifying news image-caption pairs as real or out-of-context. We found CLIP to be effective for the dataset construction and recognition of mismatches. By design, unimodal models cannot solve our task, while multimodal ones require named entity and semantic knowledge to do well on our diagnostic splits. Our Merged set aims to model the more realistic diversity of image-caption mismatches in the wild.

From our experimental results, we find that the ResNet backbone offers a modest performance boost compared to a ViT-B/32 model. We find that a CLIP ViT model is able to match human performance on a small subset of our Merged / Balanced split, and that our task is generally difficult with an average 66% human accuracy.

Our training data could be used to augment and increase the training data size of human-made falsified news, which often lack ground truth labels or sufficient scale. Overall, we show that it is possible to automatically match plausible images for given input captions, and we present a challenging benchmark to foster the development of defenses against large-scale image repurposing.

Chapter 2

Twitter-COMMs: Detecting Climate, COVID, and Military Multimodal Misinformation

Detecting out-of-context media, such as "miscaptioned" images on Twitter, is a relevant problem, especially in domains of high public significance. In this work we aim to develop defenses against such misinformation for the topics of Climate Change, COVID-19, and Military Vehicles. We first present a large-scale *multimodal* dataset with over 884k tweets relevant to these topics. Next, we propose a detection method, based on the state-of-the-art CLIP model, that leverages automatically generated hard image-text mismatches. While this approach works well on our automatically constructed out-of-context tweets, we aim to validate its usefulness on data representative of the real world. Thus, we test it on a set of human-generated fakes, created by mimicking in-the-wild misinformation. We achieve an 11% detection improvement in a high precision regime over a strong baseline. Finally, we share insights about our best model design and analyze the challenges of this emerging threat.[1]

[1] This chapter is based on joint work with Giscard Biamby, Anna Rohrbach, and Trevor Darrell presented at NAACL 2022 [3]. Giscard Biamby also led the paper.

2.1 Introduction

Out-of-context images are a popular form of misinformation where an image is miscaptioned to support a false claim [6]. Such image repurposing is extremely cheap yet can be as damaging as more sophisticated fake media. In this work we focus on domains important for society and national security, where implications of inexpensive yet effective misinformation can be immense.

Specifically, we analyze multimodal Twitter posts that are of significant public interest, related to topics of COVID-19, Climate Change and Military Vehicles. Our goal is to learn to categorize such image-text posts as pristine or falsified (out-of-context) by means of detecting semantic inconsistencies between images and text. To that end, we first collect a large-scale dataset of *multimodal* tweets, Twitter-COMMs, with over 884k tweets. In our approach, we fuse input image and text embeddings generated by CLIP [20] via an element-wise product, and train a classifier to distinguish real tweets from automatically constructed random and hard mismatches. To validate this approach and demonstrate the usefulness of the Twitter-COMMs dataset, we report results on human-generated test data, created to mimic real-world misinformation. We discuss the results and model ablations, and provide additional insights into the challenges of this task. Our dataset is publicly available at: https://github.com/GiscardBiamby/Twitter-COMMs.

2.2 Related Work

There exist a number of large-scale Twitter datasets concentrated on topics such as COVID-19 [2] or Climate Change [14]. However, it remains difficult to collect labeled misinformation. Researchers have collected COVID-19 misconceptions on social media via manual annotation [10] or by linking to fact checking articles [19]. Not only are these datasets small (a few thousand samples), but they focus on false claims rather than multimodal inconsistency. Here, we curate social media posts that are topical and multimodal, and we demonstrate an application to misinformation detection of human-generated fakes.

Recent work has developed approaches for multimodal fact checking, e.g., Jaiswal et al. [12] and Müller-Budack et al. [17], who query an external knowledge base. Similar to Luo, Darrell, and Rohrbach [16] in the news domain, we use a large pretrained model that does not require an external reference set.

2.3 Twitter-COMMs Dataset

Here, we describe the data collection strategies behind Twitter-COMMs, which consists of multimodal tweets covering the topics of COVID-19, Climate Change, and Military Vehicles.

Data Collection: We collected data using Twitter API v2[2] in three stages for COVID-

[2] https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

Table 2.1: Twitter-COMMs breakdown. “Collected“ denotes all unique samples collected via the Twitter API. “Pristine“ and “Falsified“ denote all samples in our automatically generated Training set. To ensure the balanced Training set, we “repeat” Pristine samples such that there is an equal number of Pristine and Falsified samples.

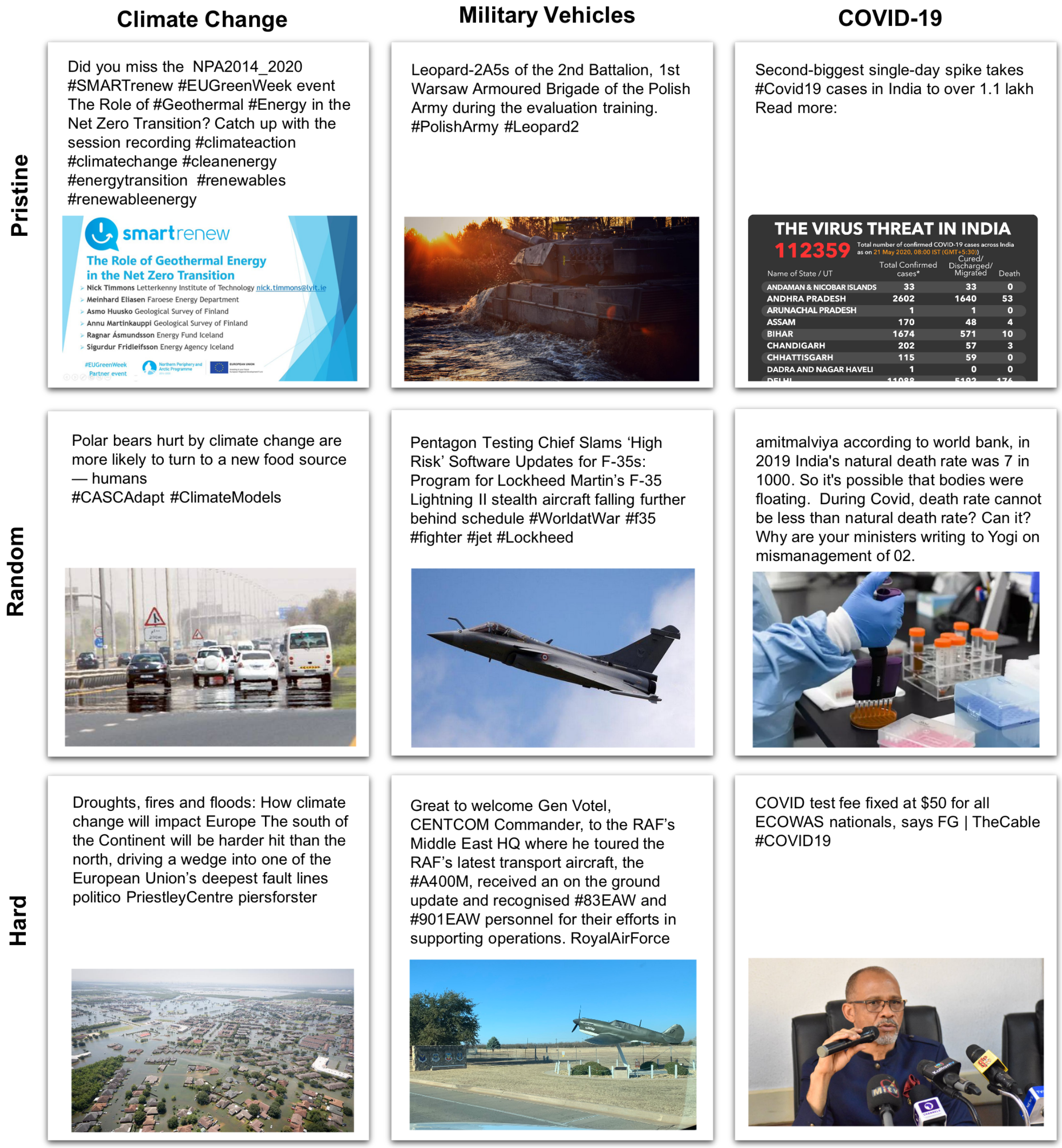
Topic / Samples	Collected	Pristine	Falsified	
			Random	Hard
Climate Change	212,665	298,809	84,432	214,377
COVID-19	569,982	736,539	162,410	574,129
Military Vehicles	101,684	139,213	35,376	103,837
Cross Topic	-	59,735	59,735	-
Total	884,331	2,468,592		

19 and Climate Change, and two stages for Military Vehicles, refining the filters at each stage to acquire more relevant tweets. COVID-19 and Climate Change stages progressed from simple high level keywords towards more specific ones in stage two and tweets authored by news organizations in the final stage. For Military Vehicles the first stage used high level search terms such as “military”, “aircraft”, “tank”, which resulted in noisy data, so the second stage used a large number of highly specific terms related to vehicle models. We employed the following global filters for all topics: (1) language=English, (2) has at least one image, and (3) not a retweet.

In total, we have collected 884, 331 tweets, each having at least one image (composed of 24% Climate Change, 64.5% COVID-19, and 11.5% Military Vehicles tweets), see Table 2.1. Tweets for Climate Change and Military Vehicles were collected starting from June 2016 and for COVID-19 starting from February 2020, all ending in September 2021.

Falsified Samples: In addition to the pristine samples, we automatically generate falsified samples where there is some inconsistency between image and text. We create random negatives (denoted as “Random”) by selecting an image for a given caption at random. We also create hard negatives (denoted as “Hard”) by retrieving the image of the sample with the greatest textual similarity for a given caption (following the “Semantics / CLIP Text-Text” split from Luo, Darrell, and Rohrbach [16]). We mainly generate mismatches *within* each topic (COVID-19, Climate Change, Military Vehicles), except for a small set of random mismatches *across* topics (denoted as “Cross Topic”). Our dataset is balanced with respect to labels, where half of the samples are pristine and half are falsified. Table 2.1 presents summary statistics for the training samples. We detail our development set and other data used for evaluation in the next section.

Qualitative Analysis: We present random examples from our training set in Figure 2.1. Overall, we see that the collected Twitter samples tend to be “on topic” and the amount


Climate Change
Military Vehicles
COVID-19
Pristine
Random
Hard
smartrenew
The Role of Geothermal Energy in the Net Zero Transition
THE VIRUS THREAT IN INDIA
112359

Did you miss the NPA2014_2020 #SMARTrenew #EUGreenWeek event The Role of #Geothermal #Energy in the Net Zero Transition? Catch up with the session recording #climateaction #climatechange #cleanenergy #energytransition #renewables #renewableenergy

Leopard-2A5s of the 2nd Battalion, 1st Warsaw Armoured Brigade of the Polish Army during the evaluation training. #PolishArmy #Leopard2

Second-biggest single-day spike takes #Covid19 cases in India to over 1.1 lakh Read more:

Polar bears hurt by climate change are more likely to turn to a new food source — humans #CASCAdapt #ClimateModels

Pentagon Testing Chief Slams 'High Risk' Software Updates for F-35s: Program for Lockheed Martin's F-35 Lightning II stealth aircraft falling further behind schedule #WorldatWar #f35 #fighter #jet #Lockheed

amitmalviya according to world bank, in 2019 India's natural death rate was 7 in 1000. So it's possible that bodies were floating. During Covid, death rate cannot be less than natural death rate? Can it? Why are your ministers writing to Yogi on mismanagement of 02.

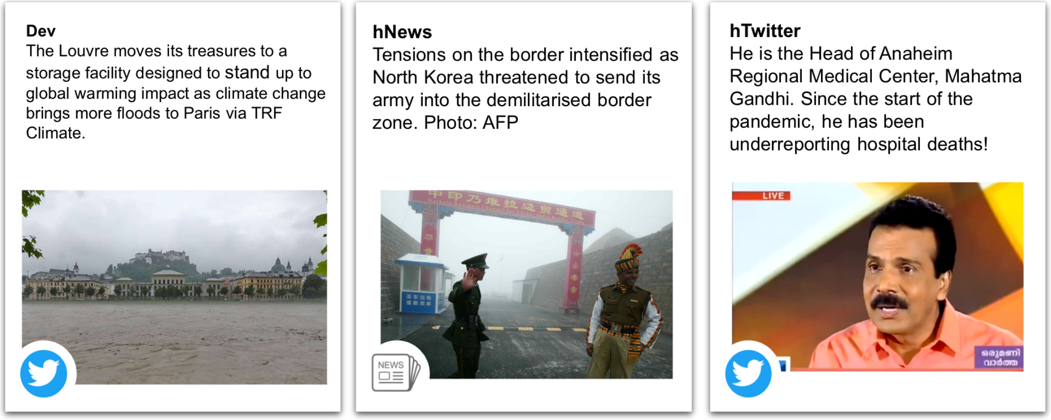
Droughts, fires and floods: How climate change will impact Europe The south of the Continent will be harder hit than the north, driving a wedge into one of the European Union's deepest fault lines politico PriestleyCentre piersforster

Great to welcome Gen Votel, CENTCOM Commander, to the RAF's Middle East HQ where he toured the RAF's latest transport aircraft, the #A400M, received an on the ground update and recognised #83EAW and #901EAW personnel for their efforts in supporting operations. RoyalAirForce

COVID test fee fixed at $50 for all ECOWAS nationals, says FG | TheCable #COVID19

Figure 2.1: Twitter-COMMs examples of Pristine and Falsified (Random / Hard) samples by topic.

of noise is low. Hard negatives are often visually grounded, while random negatives contain image/text pairs that are only weakly related, since they pertain to the same topic. The Climate Change hard negative depicts an image of flooded homes to represent "droughts, fires and floods" while the random negative depicts an image of cars relevant to climate but inconsistent with "polar bears". The COVID-19 hard negative uses an image of a Nigerian



Dev
The Louvre moves its treasures to a storage facility designed to stand up to global warming impact as climate change brings more floods to Paris via TRF Climate.

hNews
Tensions on the border intensified as North Korea threatened to send its army into the demilitarised border zone. Photo: AFP

hTwitter
He is the Head of Anaheim Regional Medical Center, Mahatma Gandhi. Since the start of the pandemic, he has been underreporting hospital deaths!

Figure 2.2: Examples of the falsified samples from the evaluation sets. Dev example is our automatically constructed hard negative sample. hNews and hTwitter samples are manually curated. Note, for hNews/hTwitter we do not show the actual samples but create similar examples for illustrative purpose, as the data is not yet publicly available.

spokesman to depict news pertaining to "ECOWAS[3]" while the random one uses a stock photo of lab testing to represent Covid. These entity-level, rather than topic-level, alignments more closely resemble real-world out-of-context images that often reference and misrepresent visually depicted entities. Note the diversity of images and text in our training set, where there exist both natural images and info-graphics, and language varies from organizational announcements and news headlines to personal opinions.

2.4 Experiments

Next, we discuss the data used for evaluation, present our approach and ablate various design choices, report results on our evaluation sets, and provide additional analysis of the task difficulty.

Evaluation Sets

We report results on three evaluation sets. (a) We validate our approach on samples synthetically generated using the same procedure as our training set (denoted Dev), where all topics and falsification methods are equally represented (i.e., the ratio of random vs. hard negatives is 50-50). We also evaluate on *human-curated* samples from the DARPA Semantic Forensics (SemaFor) Program[4] derived from (b) news images and captions (denoted hNews) and (c) Twitter (denoted hTwitter). To generate this data, humans manually introduced

[3] Economic Community of West African States

[4] Dedicated to defense against misinformation and falsified media: https://www.darpa.mil/program/semantic-forensics

Table 2.2: Evaluation samples breakdown.

	Domain	Pristine	Falsified	Total
Dev	Social Media	13,276	13,276	26,552
hNews	News	1,112	256	1,368
hTwitter	Social Media	114	122	236

inconsistencies to pristine image-caption pairs.[5] While hNews/hTwitter data is not *real* misinformation, it is *in-the-wild* w.r.t. our synthetic training data and much more representative of real-world human-generated misinformation. All three evaluation sets contain a mixture of samples relevant to the topics of COVID-19, Climate Change, and Military Vehicles (Figure 2.2). Table 2.2 provides the number of samples in each set. While the hNews set is available to us, the hTwitter set is hidden.

Approach and Design Choices

For our approach we fine-tune CLIP [20], a large pretrained multimodal model that maps images and text into a joint embedding space via contrastive learning. Our model generates CLIP embeddings using the RN50x16 backbone, multiplies the image and text embeddings, and passes the result to a classifier that scores the pair as pristine or falsified. We use a learning rate of 5e-08 for CLIP and 5e-05 for the classifier and train for 16 epochs. For our baseline CLIP Zero Shot model, we generate CLIP embeddings of-the-shelf and compute a dot product, which is used to score the pair.

We report metrics for varying thresholds over the predicted scores; in most tables we report balanced classification accuracy at equal error rate (Acc @ EER). We also report falsified class accuracy at two thresholds (pD @ 0.1 FAR and pD @ EER).

Multimodal Fusion: First, we compare different multimodal fusion techniques, see Table 2.3. We try three fusion methods: concatenating the CLIP image and text embeddings (Concat), concatenating the embeddings and their dot product (Concat + Dot), and multiplying the embeddings element-wise (Multiply). Inspired by how CLIP was trained to maximize the dot product of normalized image-text pairs, Concat + Dot and Multiply incentivize the classifier to stay faithful to the pre-initialized joint embedding space. These architecture choices yield on average a 7% performance improvement over simple concatenation. For future experiments we choose the Multiply method to minimize trainable parameters and maintain a simple approach.

Percentage of Hard Negatives: Next, we analyze the importance of using hard negatives in our training data. Specifically, we measure the impact of different percentages of hard negative samples, where the rest are random negatives. Table 2.4 presents the results. More hard negatives in training generally improves the performance on hard negatives in our

[5] We thank PAR Tech, Syracuse University, and the University of Colorado, Denver for creating the evaluation data.

Table 2.3: Balanced binary classification accuracy at EER by fusion method, Dev set.

	Climate Change		COVID-19		Military Vehicles	
	Random	Hard	Random	Hard	Random	Hard
Concat	0.8712	0.6810	0.8797	0.6882	0.9111	0.6775
Concat+Dot	0.9305	**0.8038**	0.9191	**0.7848**	**0.9485**	**0.7472**
Multiply	**0.9344**	0.7968	**0.9247**	0.7807	0.9440	0.7467

development set, but there is also a trade-off in performance on random negatives. Given that we care about samples that more closely mimic challenging real-world misinformation but also want to avoid degrading performance on easy samples, we opt for a ratio of 75% hard and 25% random negatives for future experiments.

Table 2.4: Balanced binary classification accuracy at EER by percentage of hard negatives, Dev set.

	Climate Change		COVID-19		Military Vehicles	
	Random	Hard	Random	Hard	Random	Hard
0%	0.9352	0.7714	0.9188	0.7600	0.9405	0.7236
50%	0.9344	0.7968	**0.9247**	0.7807	**0.9440**	0.7467
75%	**0.9356**	0.7979	0.9241	0.7809	0.9410	**0.7470**
100%	0.9311	**0.8004**	0.9227	**0.7834**	0.9425	0.7457

Results and Analysis

Results on hNews, hTwitter Sets: Our final model was directly fine-tuned on the entire training set of over 2M training samples, with a ratio of 75% hard and 25% random negatives. We report results in Table 2.5, comparing to CLIP Zero Shot. We improve by 11% in pD @ 0.1FAR, meaning that our method is able to detect more falsified samples with minimal false alarms. At equal error rate we improve by 5% in both detection and accuracy. We emphasize that the hTwitter data is unseen to us.

Next, we analyze the performance of our final model w.r.t. several characteristics on our Dev set.

OCR Coverage: Given that text present in images can often be used to corroborate captions, we break down model performance by the amount of text detected by an English OCR model[6]. In Table 2.6 (top), we report results broken down by the % of the image covered by text (the area of the union of text detections divided by the image size). Each bucket roughly corresponds to natural images, natural images with scene text, graphics, and

[6] https://github.com/JaidedAI/EasyOCR

Table 2.5: Balanced binary classification accuracy at varying thresholds on Dev, hNews and hTwitter sets. We report based on Probability of Detection (pD), False Alarm Rate (FAR), and Equal Error Rate (EER).

		pD @ 0.1 FAR	pD @ EER	Acc @ EER
Dev	Zero Shot	0.7396	0.8287	0.8286
	Ours	**0.8044**	**0.8546**	**0.8546**
hNews	Zero Shot	0.2852	0.6133	0.6133
	Ours	**0.4219**	**0.6836**	**0.6840**
hTwitter	Zero Shot	0.7623	0.8279	0.8306
	Ours	**0.8771**	**0.8771**	**0.8771**

screenshots of text. The presence of any text yields more than a 6% improvement for pD @ 0.1FAR and performance peaks at 10-50% coverage.

Table 2.6: Balanced binary classification accuracy at varying thresholds on Dev set broken down by: % of image covered by text (top), various text-image relationships (middle) and within- vs. cross-cluster status of the hard falsifications (bottom). The latter results are obtained on the subset of hard falsified samples and their corresponding pristine samples.

	pD @ 0.1 FAR	pD @ EER	Acc @ EER
OCR Coverage			
=0%	0.7588	0.8329	0.8329
0-10%	0.8192	0.8575	0.8575
10-50%	0.8367	0.8709	0.8710
¿50%	0.8412	0.8588	0.8588
Text-Image Relationship			
Image does not add	0.7908	0.8471	0.8470
Image adds	0.8308	0.8675	0.8674
Text not represented	0.7696	0.8401	0.8401
Text represented	0.8518	0.8745	0.8745
Tweet Text Clustering			
Climate Change			
Cross-cluster	0.7214	0.8268	0.8268
Within-cluster	0.6571	0.8055	0.8055
COVID-19			
Cross-cluster	0.6837	0.8099	0.8103
Within-cluster	0.6013	0.7758	0.7753
Military Vehicles			
Cross-cluster	0.7826	0.8634	0.8618
Within-cluster	0.6000	0.7539	0.7545

Text-Image Relationship: Within social media, there exist more complex interactions than the direct relationships seen in formats like image alt-text. As such, we trained a CLIP model on the dataset presented by [27] to characterize these relationships: classifying if the image content adds additional meaning (image adds / does not add) or if there is semantic overlap between the text and image (text represented / not represented).[7] As observed in Table 2.6 (middle), for samples with *text represented* model performance improves by 8% and for samples where *image adds* performance improves by 4% for detection in a high precision regime (pD @ 0.1FAR). Although the text-image relationship model has somewhat noisy classifications for the text task, the *text represented* class generally contains samples with a shared entity between image and text, which would make fine-grained misinformation detection easier. The *image adds* class mostly contains info-graphics, likely due to training data bias, which aligns with the OCR coverage experiments above.

Tweet Text Clustering: Finally, we analyze the sub-topics obtained as a result of clustering Tweets within each topic. This allows us to tease out clusters, e.g., *vaccination* for COVID-19, *floods* for Climate Change or *drones* for Military Vehicles. Recall that our model performs the best on Climate Change and the worst on the Military Vehicles (Table 2.4). Possible factors include the smaller amount of training data and visual similarity of different vehicle types. We also observe that among the hard negatives for Military Vehicles, only 39% are cross-cluster (while Climate Change and COVID-19 have 51% and 58% respectively), indicating the Military Vehicles set contains a larger proportion of harder fakes. These factors may explain the larger difference between cross/within cluster performance for this topic (Table 2.6, bottom).

2.5 Conclusion

In this work we tackle a real-world challenge of detecting out-of-context image-text tweets on COVID-19, Climate Change, Military Vehicles topics. To approach it, we collect Twitter-COMMs, a large-scale topical dataset with *multimodal* tweets, and construct corresponding hard mismatches. We design our approach based on the CLIP model with several important design choices, e.g. multiplying the embeddings for multimodal fusion and increasing the percentage of hard negatives in our training data. This approach substantially improves over a powerful baseline, an off-the-shelf CLIP model, when evaluated on human-curated in-the-wild mismatches. We hope our work and insights will benefit multimedia forensics practitioners.

[7] Our model achieves 86% and 62% on the image and text binary classification tasks respectively, which is 5% and 4% higher than the best models presented in the original paper.

Bibliography

[1] Shivangi Aneja, Christoph Bregler, and Matthias Nießner. "Catching Out-of-Context Misinformation with Self-supervised Learning". In: *arXiv:2101.06278* (2021).

[2] Juan M. Banda et al. "A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration". In: *Epidemiologia* 2.3 (2021), pp. 315–324.

[3] Giscard Biamby et al. "Twitter-COMMs: Detecting Climate, COVID, and Military Multimodal Misinformation". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (2022).

[4] Hila Chefer, Shir Gur, and Lior Wolf. "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2019.

[6] Lisa Fazio. *Out-of-context photos are a powerful low-tech form of misinformation.* `https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959/`. 2020.

[7] Elise Fenn et al. "Nonprobative photos increase truth, like, and share judgments in a simulated social media environment". In: *JARMAC* 8.2 (2019), pp. 131–138.

[8] Karen Hao. *A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it.* `https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/`. 2020.

[9] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: (2017).

[10] Tamanna Hossain et al. "COVIDLies: Detecting COVID-19 Misinformation on Social Media". In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. 2020.

[11] Johannes M. van Hulst et al. "REL: An Entity Linker Standing on the Shoulders of Giants". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

[12] Ayush Jaiswal et al. "Multimedia semantic integrity assessment using joint embedding of images and text". In: *Proceedings of the ACM international conference on Multimedia (MM)*. 2017.

[13] Liunian Harold Li et al. "VisualBERT: A simple and performant baseline for vision and language". In: *arXiv:1908.03557* (2019).

[14] Justin Littman and Laura Wrubel. *Climate Change Tweets Ids*. 2019. URL: `https://doi.org/10.7910/DVN/5QCCUU`.

[15] Fuxiao Liu et al. "VisualNews: A Large Multi-source News Image Dataset". In: (2021).

[16] Grace Luo, Trevor Darrell, and Anna Rohrbach. "NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).

[17] Eric Müller-Budack et al. "Multimodal analytics for real-world news using measures of cross-modal entity consistency". In: *ACM International Conference on Multimedia Retrieval (ICMR)*. 2020.

[18] Kai Nakamura, Sharon Levy, and William Yang Wang. "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2019.

[19] Parth Patwa et al. "Fighting an Infodemic: COVID-19 Fake News Dataset". In: *In Proceedings of the First Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation at AAAI 2021*. 2021.

[20] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *arXiv:2103.00020* (2021).

[21] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2015.

[22] Ekraam Sabir et al. "Deep multimodal image-repurposing detection". In: *Proceedings of the ACM international conference on Multimedia (MM)*. 2018.

[23] Piyush Sharma et al. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018.

[24] Kai Shu et al. "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media". In: *Big Data* 8.3 (2020), pp. 171–188.

[25] Reuben Tan, Kate Saenko, and Bryan A Plummer. “Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[26] Christopher Thomas and Adriana Kovashka. “Preserving Semantic Neighborhoods for Robust Cross-modal Retrieval”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.

[27] Alakananda Vempala and Daniel Preoţiuc-Pietro. “Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

[28] Daniel Victor. *Your Loved Ones, and Eerie Tom Cruise Videos, Reanimate Unease With Deepfakes*. `https://www.nytimes.com/2021/03/10/technology/ancestor-deepfake-tom-cruise.html/`. 2021.

[29] Bin Wang and C-C Jay Kuo. “SBERT-WK: A sentence embedding method by dissecting BERT-based word models”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2146–2157.

[30] Rowan Zellers et al. “Defending against neural fake news”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[31] Bolei Zhou et al. “Places: A 10 million Image Database for Scene Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).