ImageArg: A Multi-modal Tweet Dataset for Image Persuasiveness Mining

Zhexiong Liu*, Meiqi Guo*, Yue Dai*, Diane Litman
Department of Computer Science
University of Pittsburgh, Pittsburgh, Pennsylvania, 15260
{zhexiong.liu,meiqi.guo,yud42,dlitman}@pitt.edu

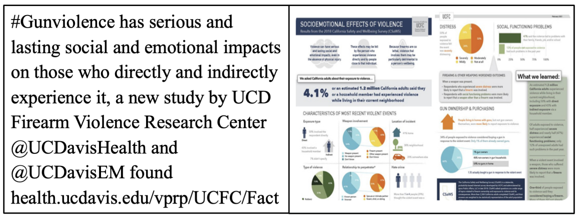
Abstract

The growing interest in developing corpora of persuasive texts has promoted applications in automated systems, e.g., debating and essay scoring systems; however, there is little prior work mining image persuasiveness from an argumentative perspective. To expand persuasiveness mining into a multi-modal realm, we present a multi-modal dataset, *ImageArg*, consisting of annotations of image persuasiveness in tweets. The annotations are based on a persuasion taxonomy we developed to explore image functionalities and the means of persuasion. We benchmark image persuasiveness tasks on *ImageArg* using widely-used multi-modal learning methods. The experimental results show that our dataset offers a useful resource for this rich and challenging topic, and there is ample room for modeling improvement.

1 Introduction

Argumentation mining (AM) aims to analyze authors' argumentative stance by automatically identifying argumentative structures and their relationships (Green et al., 2014). As a fundamental component in AM, computational persuasiveness analysis has gained considerable momentum due to growing resources and downstream applications (Chatterjee and Agrawal, 2006; Park et al., 2014; Wei et al., 2016; Lukin et al., 2017; Chakrabarty et al., 2017; Lytos et al., 2019). Aiming at automatically evaluating how well one party can change another party's opinions or behaviors, computational persuasiveness tasks are critical yet challenging.

Recent work in AM has brought attention to mining persuasiveness in essays. Stab and Gurevych (2014) and Habernal and Gurevych (2017) developed the Argument Annotated Essays Corpus (AAEC) where stance, argument components, and argumentative relations were annotated. Carlile et al. (2018) extended AAEC annotations with persuasiveness scores, as well as with argumentative attributes that potentially impact persuasiveness (Eloquence, Specificity, Relevance, and Evidence) and the means of persuasion (Ethos, Pathos, or Logos). These are all text-based annotations, however, missing the opportunity to leverage other modalities (e.g., images) that potentially enhance the persuasiveness of the argument. For example, the image showing statistic charts in Fig. 1 makes the tweet text more convincing. To address the gap that image persuasivness has rarely been explored in the AM community, we create a new multi-modal dataset, *ImageArg*, that annotates image persuasiveness in tweets and extends persuasiveness mining to a multi-modal realm.



#Gunviolence has serious and lasting social and emotional impacts on those who directly and indirectly experience it, a new study by UCD Firearm Violence Research Center @UCDavisHealth and @UCDavisEM found health.ucdavis.edu/vprp/UCFC/Fact

SOCIOEMOTIONAL EFFECTS OF VIOLENCE
Results from the 2018 California Safety and Wellbeing Survey (CSaWS)
UCFC
DISTRESS
SOCIAL FUNCTIONING PROBLEMS
4.1%
or an estimated 1.2 million California adults said they or a household member had experienced violence while living in their current neighborhood
CHARACTERISTICS OF MOST RECENT VIOLENT EVENTS
FIREARMS & OTHER WEAPONS WORSENED OUTCOMES
GUN OWNERSHIP & PURCHASING
What we learned:
CSaWS

(a) A posted tweet text

(b) An associated posted tweet image

Figure 1: (a) The tweet text uses gun violence to argue for *gun control*. (b) The image makes the argument more persuasive by providing supplementary statistics relating violence to gun ownership in California.

Regarding *ImageArg* construction, we first extend annotation schemes that are previously developed to capture the persuasive strength of text arguments in AAEC (Duthie et al., 2016; Wachsmuth et al., 2018; Carlile et al., 2018) to a new modality of image. Specifically, we develop a novel strategy (Sec. 3.2) to annotate multi-modal persuasiveness gains that measure if the persuasivness of a tweet's text increases after adding a visual image. Second, we devise a taxonomy to annotate image content (Sec. 3.3) that explicitly identifies image functionalities from a persuasive perspective. Furthermore,

[*] These authors contributed equally to this work.

Proceedings of the 9th Workshop on Argument Mining, co-located with the 29th International Conference on Computational Linguistics, pages 1–18
October 17, 2022.

we adapt existing text attributes used in Carlile et al. (2018) to annotate image persuasion modes (Sec. 3.4) by exploring different annotation strategies (Sec. 4.2). We evaluate the inter-rater agreement on our proposed annotation schemes as well as the quality of the annotated samples.

With *ImageArg*, we first report the basic statistics of the dataset and conduct a thorough analysis between different annotation dimensions (Sec. 4.3). We observe a strong correlation between human political ideology (i.e. stance towards a social topic) and the argumentative features in their posted tweets, as well as mutual influences between image content and persuasion mode. In addition, we benchmark model performance on multiple argumentative classification tasks annotated in *ImageArg* (Sec. 5.2). Specifically, we employ multi-modal learning methods to classify stance, image persuasiveness, image content, and image persuasion mode. Our benchmark results highlight the challenge of these tasks and indicate there is ample room for model improvement. We demonstrate the limitation of these general multi-modal methods and discuss possible future work. We further conduct a qualitative study on a real-world application, retrieving the most persuasive images given a tweet text, by using our trained classifiers (Sec. 5.3), which offers a starting point for developing an intelligent tool that recommends persuasive images to users based on their textual inputs. Our code and data is publicly available at: `https://github.com/MeiqiGuo/ArgMining2022-ImageArg`.

2 Related Work

Computational Persuasiveness While classical AM focuses on identifying argumentative components and their relations (Stab et al., 2014, 2018; Lawrence and Reed, 2020), recent work has developed interest in persuasiveness related tasks (Chatterjee et al., 2014; Park et al., 2014; Lukin et al., 2017; Carlile et al., 2018; Chakrabarty et al., 2019). In addition, Riley (1954), O'keefe (2015), and Wei et al. (2016) investigate ranking debate arguments on the same topic based on their persuasiveness, but they failed to investigate the factors that make arguments persuasive. Lukin et al. (2017) and Persing and Ng (2017) examine how audience variables (e.g., personality) influence persuasiveness through different argument styles (e.g., factual vs. emotional arguments), but only focus on the text modality. Higgins and Walker (2012) and Carlile et al. (2018) study the persuasion strategies, i.e., Ethos (credibility), Logos (reason), and Pathos (emotion), in the scope of reports or student essays. We follow their work developed for text corpora and extend the annotation schemes to the image modality. Although Park et al. (2014), Joo et al. (2014), and Huang and Kovashka (2016) utilize facial expressions and bodily gestures to analyze persuasiveness in social multimedia, their work is limited to the human portrait and fails to generalize to diverse image domains. Some prior work study persuasive advertisements in a multi-modal way (Hussain et al., 2017; Guo et al., 2021). Different from our argumentative mining goal, they focus on the sentiment, intent reasoning and persuasive strategies that are narrowly designed for ads. Thus, annotating a multi-modal tweet dataset focusing on image persuasiveness is under-explored in existing work, and has ample value for social science.

Multi-modal Learning The ability to process and understand multi-modal input for AI models has recently received much attention since the multi-modal signals are generally complementary for real-world applications (Aytar et al., 2016; Zhang et al., 2018; Alwassel et al., 2020). In the area of vision-language, tasks are mainly designed for evaluating models' ability to understand visual information as well as expressing the reasoning in language (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019). In addition to the main stream, a few works study the relationship between image and text: Alikhani et al. (2019) annotates the discourse relations between text and accompanying imagery in recipe instructions; and Kruk et al. (2019) investigates the multi-modal document intent in instagram posts. However, multi-modal learning for AM has been under-explored due to a lack of multi-modal corpora. This drives us to build *ImageArg* and to analyze the effectiveness of multi-modal learning on AM tasks. With respect to modeling, researchers focus on learning good representation of each modality and developing effective fusion methods (Tsai et al., 2018; Hu et al., 2019; Tan and Bansal, 2019; Lu et al., 2020). In this work, we establish a benchmark performance for *ImageArg* by using fundamental and common encoders and fusion methods.

3 Annotation Scheme

We propose an annotation scheme to capture an image's impact on the persuasiveness of multi-modal

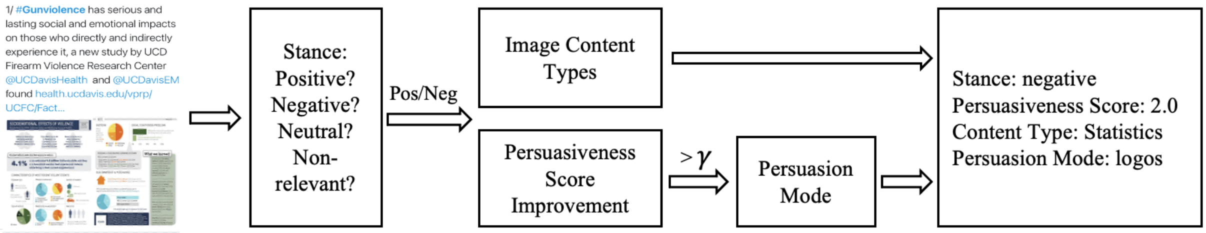

1/ #Gunviolence has serious and lasting social and emotional impacts on those who directly and indirectly experience it, a new study by UCD Firearm Violence Research Center @UCDavisHealth and @UCDavisEM found health.ucdavis.edu/vprp/UCFC/Fact...
Stance: Positive? Negative? Neutral? Non-relevant?
Pos/Neg
Image Content Types
Persuasiveness Score Improvement
$>\gamma$
Persuasion Mode
Stance: negative
Persuasiveness Score: 2.0
Content Type: Statistics
Persuasion Mode: logos

Figure 2: The overview of our annotation pipeline. Annotators start by annotating the argumentative stance of input tweets. Afterwards, tweets with either positive or negative stances are annotated for image content types and persuasiveness score improvement. The persuasion mode is further annotated if persuasiveness score improvement exceeds a given threshold γ. We use $\gamma = 0.5$ when we annotate data and test with different γ values for persuasiveness classification task (Table 6).

THE HOUSE JUST PASSED H.R. 8, A BILL TO REQUIRE BACKGROUND CHECKS ON ALL GUN SALES, AND H.R. 1446, A BILL TO ADDRESS THE CHARLESTON LOOPHOLE.
NOW THE SENATE MUST ACT.
EVERYTOWN.ORG/CHECKS
GUN CONTROL! The THEORY that becoming a VICTIM is somehow Morally Superior to DEFENDING your Family!

Support: ANOTHER BIG WIN FOR GUN SAFETY! Just after passing a bill to require background checks on all gun sales #HR8 the U.S House just passed #HR1446, a bill that would address the deadly Charleston loophole. This bill now heads to the Senate.

Oppose: #GunControl The THEORY that becoming a VICTIM is somehow morally superior to DEFENDING yourself or family! We support #SelfDefence #Legalgunownership #SafeCitizen

Figure 3: Examples of positive (support) and negative (oppose) tweets.

tweets. We build a corpus of Twitter posts on a social topic (e.g., *gun control*), then annotate the image within each post along four dimensions. The annotation pipeline is shown in Fig. 2. First, we determine **(1)** the **stance** of the entire tweet (Sec. 3.1). Specifically, we assume one tweet holds a consistent stance in its text and image since the author would intend to deliver a consistent argument. For those tweets annotated with a positive or negative stance, we also annotate **(2)** the **persuasiveness scores** of the tweet image (Sec. 3.2) and **(3)** the image **content type**. The content types identify image roles from an argumentative perspective (Sec. 3.3). Finally, we **(4)** identify the **persuasion mode** of an image that is annotated as persuasive. The persuasion mode indicates how the images persuade audiences (Sec. 3.4). Note that with this annotation pipeline, all tweets will first be annotated for stance. Then, only tweets with a clear stance will be annotated for content type and persuasiveness scores. Finally, only tweets where the images are persuasive will be annotated for persuasion mode.

3.1 Stance

We use existing methods (Mohammad et al., 2017) to verify if the image holds a clear stance on a given topic. Specifically, given a tweet (including text and images), we ask annotators to select among four stances that are extended from Mohammad et al. (2017): positive (i.e., support), negative (i.e., oppose), neutral, or irrelevant to the topic. We continue with the next annotation steps only if a tweet holds a positive or negative stance. Otherwise, it is discarded for our persuasion study. We show examples in Fig. 3.

3.2 Image Persuasiveness Scores

For a tweet that holds a positive or negative stance, we study the impact of its image by computing an image persuasiveness score improvment. We adopt five levels of text persuasiveness scores proposed in Carlile et al. (2018) in the annotation process: (L0) no persuasiveness (score = 0): the annotated target fails to convince the audience at all. (L1) medium persuasiveness (score = 1): the annotated target partially convinces the audience. (L2) persuasive (score = 2): the annotated target is convincing to the audience. (L3) high persuasiveness (score = 3): the annotated target is very convincing to the audience. (L4) extreme persuasiveness (score = 4): the annotated target is compelling to the audience.

Different from Carlile et al. (2018) that annotates the persuasiveness score directly, we propose a novel method to compute the image persuasiveness score. In particular, we calculate the differences with/without images to quantify image persuasiveness scores. We first ask annotators to choose one of 5 persuasiveness levels based on pure text from the tweet. Next, we ask annotators to give a second choice based on both text and image from the tweet. Suppose each sample has three annotations and each annotation has two persuasiveness scores: one for the text-only (s_t), the other for the image-text (s_{it}). We compute persuasiveness score differ-

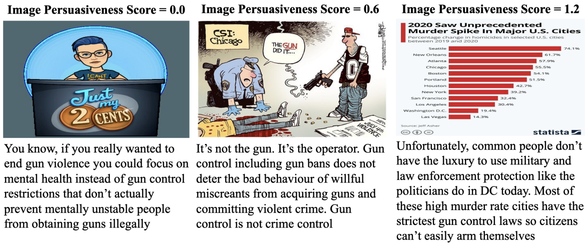


Image Persuasiveness Score = 0.0

You know, if you really wanted to end gun violence you could focus on mental health instead of gun control restrictions that don't actually prevent mentally unstable people from obtaining guns illegally

Image Persuasiveness Score = 0.6

It's not the gun. It's the operator. Gun control including gun bans does not deter the bad behaviour of willful miscreants from acquiring guns and committing violent crime. Gun control is not crime control

Image Persuasiveness Score = 1.2

Unfortunately, common people don't have the luxury to use military and law enforcement protection like the politicians do in DC today. Most of these high murder rate cities have the strictest gun control laws so citizens can't easily arm themselves

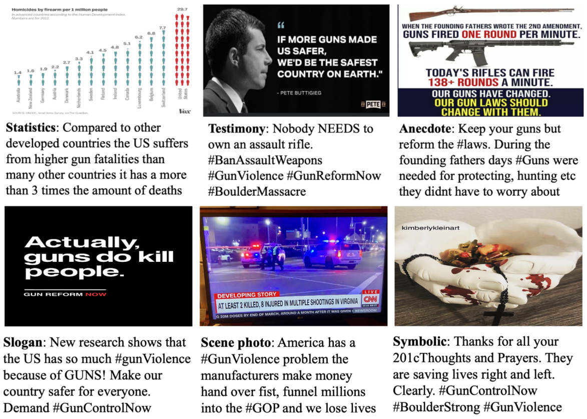
Figure 4: Examples of tweets with 0, 0.6, and 1.2 image persuasiveness scores.

ence $\Delta s_i = max(s_{it} - s_t, 0)$ for each annotation, as the persuasiveness gain from the image. Then, we compute the average of the three annotations (Δs_i) as the final image persuasiveness score. To interpret image persuasiveness, we use a threshold (γ) that encodes the score into a binary label (i.e., persuasiveness or not). If Δs_i is higher than the threshold (γ), it indicates that adding an image improves tweet persuasiveness, thus the image is considered as persuasive. We show examples with different image persuasiveness scores in Fig. 4.

3.3 Image Content Types

For persuasive samples, we investigate their image argumentative roles. In particular, we annotate the image content types from an argumentative perspective to describe what kind of evidence images provide to improve tweet persuasiveness (e.g., supportive data, authorized photos, etc.). We leverage Al Khatib et al. (2016)'s definition of argumentative roles of evidence to categorize image content: Statistics, Testimony, and Anecdote. However, we notice that the categories fail to capture all the image contents that frequently appear in tweet posts, for example, photographs. To this end, we propose a Slogan category highlighting text in images, and also propose Scene photo and Symbolic photo categories regarding image content in the visual modality. More details are specified as follows:

- **Statistics**: Images provide evidence by stating or quoting quantitative information, such as a chart or diagram showing data, that is related to the tweet text. In Fig. 5, the image provides quantitative statistics on gun fatalities.
- **Testimony**: Images quote statements or conclusions from an authority, such as a piece of articles or claims from an official document, that is related to the tweet text. For example, in Fig. 5, the testimony image cites a statement given by the transportation secretary.


IF MORE GUNS MADE US SAFER, WE'D BE THE SAFEST COUNTRY ON EARTH."
- PETE BUTTIGIEG
WHEN THE FOUNDING FATHERS WROTE THE 2ND AMENDMENT, GUNS FIRED ONE ROUND PER MINUTE.
TODAY'S RIFLES CAN FIRE 138+ ROUNDS A MINUTE.
OUR GUNS HAVE CHANGED. OUR GUN LAWS SHOULD CHANGE WITH THEM.
Actually, guns do kill people.
GUN REFORM NOW
DEVELOPING STORY
AT LEAST 2 KILLED, 8 INJURED IN MULTIPLE SHOOTINGS IN VIRGINIA
CNN
kimberlykleinart

Statistics: Compared to other developed countries the US suffers from higher gun fatalities than many other countries it has a more than 3 times the amount of deaths

Testimony: Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre

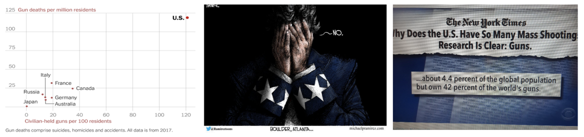
Anecdote: Keep your guns but reform the #laws. During the founding fathers days #Guns were needed for protecting, hunting etc they didnt have to worry about

Slogan: New research shows that the US has so much #gunViolence because of GUNS! Make our country safer for everyone. Demand #GunControlNow

Scene photo: America has a #GunViolence problem the manufacturers make money hand over fist, funnel millions into the #GOP and we lose lives

Symbolic: Thanks for all your 201cThoughts and Prayers. They are saving lives right and left. Clearly. #GunControlNow #BoulderStrong #GunViolence

Figure 5: Examples of image content types in tweets: statistics, testimony, anecdote, slogan, scene photo, and symbolic.


Gun deaths per million residents
U.S.
Italy
France
Canada
Russia
Germany
Japan
Australia
Civilian-held guns per 100 residents
NO.
The New York Times
Why Does the U.S. Have So Many Mass Shootings? Research Is Clear: Guns.
...about 4.4 percent of the global population but own 42 percent of the world's guns

Logos: Gun deaths and gun ownership by population - by country. Hmmm. Well, this doesn't take much effort to figure out why we've got such #GunViolence…

Pathos: A personal narrative - Dr. Sonya Lewis "We must reject helplessness and complacency and we must allow ourselves to feel the raw, sick"

Ethos: The US has 4.4 % of the world's population but 42% of gun violence. #guncontrol #gunviolence

Figure 6: Examples of persuasion mode in tweet: logos, pathos, and ethos.

- **Anecdote**: Images provide information based on the author's personal experience, such as facts/personal stories, that are related to the tweet text. In Fig. 5, the anecdote image shows the fact that guns are developed since the period of the 2nd amendment, and therefore the laws for guns should be developed as well.
- **Slogan**: Images embed pieces of advertising/slogan text. In Fig. 5, the slogan image presents a phrase "Actually guns do kill people. Gun Reform Now".
- **Scene photo**: Images show a real scene or photograph that is related to the tweet text. In Fig. 5, the image shows a photo of a gun violence scene reported by CNN news.
- **Symbolic photo**: Images show a symbol/art that expresses the author's viewpoints in a non-literal way. In Fig. 5, the symbolic photo shows a pair of artificial bloodied hands holding bullets and a cross which symbolically reveals the brutality of gun violence.

3.4 Image Persuasion Modes

To investigate how images convince an audience (e.g., by providing strong logic, touching audiences emotionally, etc.), we annotate the persuasion modes of images by leveraging the definitions in Braet (1992) for Logos, Pathos, and Ethos. The modes form the rhetorical triangle, and both the textual and visual modalities follow these dimensions in the persuasiveness perspective. Fig. 6 shows examples, details are specified below:

- **Logos**: The image appeals to logic and reasoning, which persuades audiences with reasoning from a fact/statistics/study case/scientific evidence. In Fig. 6, the Logos image provides a chart that shows the high gun deaths and the high gun ownership by the population of the US, which implies a logical relationship between gun death and gun ownership.
- **Pathos**: The image appeals to emotion, i.e., evokes emotional impact that leads to higher persuasiveness. In Fig. 6, the Pathos image provides art that shows the grieved "Uncle Sam" saying "no" with helplessness, which evokes the desire to *gun control*.
- **Ethos**: The image appeals to ethics, which enhances credibility and trustworthiness. In Fig. 6, the Ethos image takes a screenshot of the source of a report from New York Times, which increases credibility.

4 Corpus Creation

4.1 Data Collection

We collect raw tweets containing both image and text across 3 topics (*gun control, immigration* and *abortion*) used in Mochales and Moens (2011) and Stab et al. (2018). Specifically, we retrieve tweets with images that contain pre-defined keywords[1] through TwitterAPI[2]. The raw data (286k tweets) are collected in a two-year window from 3/29/2019 to 3/29/2021. We retain tweets whose texts tend to be argumentative, with an argument confidence score larger than 0.9 by using ArgumentText Classify API[3]. 99.48% of tweets are discarded for having an argument confidence score below 0.9. These filtering processes ensure our annotation data has high argumentation-confidence and topic-relevance.

4.2 Annotation Strategies

We develop annotation strategies based on several rounds of pilot annotations. To ensure the annotation quality, we provide coding manual and examples for annotators (see the Appendix A for details). We employ qualified workers who passed a qualitative test that evaluates the workers' understanding on our annotation manual.

Task	Alpha	Count
Stance	64.5	87
Content type	71.1	38
Persuasion mode	19.9	38

Table 1: First pilot annotation inter-agreement on *gun control* topic. Persuasion modes are annotated as single choices from logos, pathos, and ethos.

Task	Alpha	Count
Stance	76.1	1003
Persuasiveness*	/	1003
Content type	64.6	1003
Logos	55.3	259
Pathos	51.0	259
Ethos	57.8	259

Table 2: Inter-agreement rate of each annotation task in our final corpus on *gun control* topic, and the number of samples with the corresponding annotation. (*) We only show numbers of persuasiveness since they are annotated with average persuasiveness scores from annotators rather than labels.

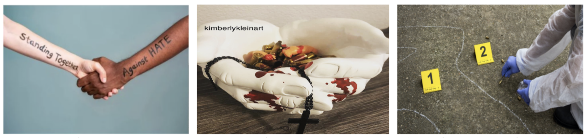
We start with the topic of *gun control*. In the first-round, we distribute 87 samples to two random annotators on MTurk. Table 1 shows Krippendorff's alpha (Krippendorff, 2011) score for interrater agreement[4]. Based on the interpretation of alpha scores in Landis and Koch (1977); Hartling et al. (2012), we conclude that stance and content type have a substantial inter-agreement but persuasion mode inter-agreement is slight. To investigate this issue, we modify our annotation guideline for persuasion mode. Instead of using three-class annotation (i.e., choosing one persuasion mode from 3 options), we move to three-label annotation that asks a binary question for each mode for each sample (i.e., annotating yes/no for each persuasion mode, individually). Moreover, the annotators are required to justify their choices by giving short comments. The improved results (on the final corpus from Sec. 4.3) are shown in Table 2, although the persuasion mode agreement (i.e., Logos, Pathos, and Ethos) is still lower than stance and content type. This is likely because annotators have different emotional reasoning (i.e., some annotators are easily evoked by images while others are not).

[1] We use keywords provided in Guo et al. (2020)'s work.

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

[3] https://api.argumentsearch.com

[4] Note that the availability of annotation questions is based on the answer to the prior questions (Fig. 2) therefore each task has different sample numbers.



A: Togetherness; B: No A: Sadness; B: No A: Sadness & Anxiety; B: No

Figure 7: Annotator A annotates the above images as Pathos because these examples express emotions, while annotator B disagrees and marks as not Pathos.

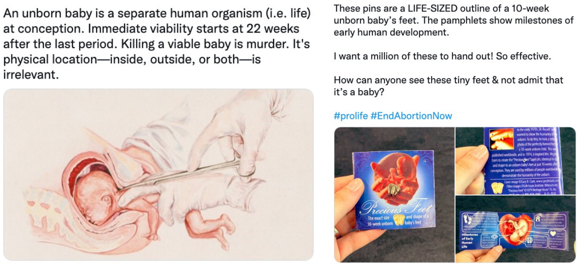
For example, one annotator recognized strong emotional impact (e.g., togetherness, sadness, anxiety, etc.), while the other not as shown in Fig. 7.

We further perform pilot annotations for the topics of *immigration* and *abortion*, with the best annotation strategies that we developed for annotating *gun control*. We randomly choose 100 or 200 tweets respectively on *immigration* or *abortion* for the pilot study, and make a topic-specific instruction for the stance annotation that provides some topic-specific examples. The Inter-rater Agreement for both topics is shown in Table 3. We observe high Inter-rater Agreements on the stance annotation, which demonstrates the utility of our topic-specific instructions. The agreement on the content type is generally good, however, *abortion* has relatively lower agreement than the other two topics. One main reason is that authors prefer using photos to support their arguments. Such photos lead to ambiguity between scene photos and symbolic photos, as examples shown in Fig. 8. Moreover, we notice that the agreements on the persuasion modes are not satisfying. For *immigration*, Ethos has the lowest agreement, and one explanation is that there are few authentic resources that provide credible and trustworthy arguments on this topic; for *abortion*, the agreement on all three persuasion modes are relatively low, in particular, Logos surprisingly gets the lowest agreement.

These studies indicate that the inter-rater agreement on annotating persuasion mode is topic-dependent, and the relationship between topics and persuasion modes needs further investigation. We thus create the first version of *ImageArg* data using only the *gun control* topic, and leave the other two topics for future work.

Task	Immigration		Abortion	
	Alpha	Count	Alpha	Count
Stance	61.5	100	68.7	200
Content type	65.8	53	56.6	76
Logos	56.7	23	25.0	48
Pathos	46.0	23	37.5	48
Ethos	30.8	23	28.2	48

Table 3: Inter-agreement rate of each annotation task on the topic *immigration* and *abortion*. The count represents the number of samples after filtering from previous questions.



An unborn baby is a separate human organism (i.e. life) at conception. Immediate viability starts at 22 weeks after the last period. Killing a viable baby is murder. It's physical location—inside, outside, or both—is irrelevant.

These pins are a LIFE-SIZED outline of a 10-week unborn baby's feet. The pamphlets show milestones of early human development.

I want a million of these to hand out! So effective.

How can anyone see these tiny feet & not admit that it's a baby?

#prolife #EndAbortionNow

A: Scene Photo
B: Symbolic Photo

A: Scene Photo
B: Symbolic Photo

Figure 8: Samples of disagreed on the content type in the topic *abortion*.

4.3 Corpus Statistics and Analysis

We annotate 1003 samples that hold a support or oppose stance on *gun control* topic. 36% of data is discarded for not having an agreed support/oppose stance. We report the distribution of each annotation scheme in Fig. 9, and the inter-rater agreement evaluation in Table 2. The results reveal that the annotators have substantial agreement on the stance and content types, and moderate agreement on the image persuasion mode. Specifically, the stance annotations are balanced distributed as shown in Fig. 9 (a): 46.3% support and 54.7% oppose. As for image persuasiveness annotations, Table 4 shows sample distributions in different persuasiveness score intervals. We use a threshold γ to discretize numerical persuasiveness scores to binary labels (i.e., persuasiveness or not). The γ is set to 0.5 in our annotations since the persuasiveness score is an average of three annotators, thus γ greater than 0.5 suggests that there is at least two annotators annotating images persuasiveness with L1 or higher ($\geq$ 1) scores (as defined in Sec. 3.2) or at least one annotator annotating L2 or higher scores ($\geq$ 2). In terms of image content types, its distribution is shown in Fig. 9 (b): Symbolic photo (23.43%), Scene photo (21.93%), Anecdote (19.84%), Slogan (14.76%), Testimony (10.87%), Statistics (7.28%), Other (1.89%). We observe that images (i.e., symbolic photo/scene photo) occupy a high proportion of the samples, in contrast, data evidence (i.e., statistics) takes the relatively low ratio. One potential reason is that social media contents like tweets are generally short and informal, which prefers relatively simple evidence. Note that there are 19

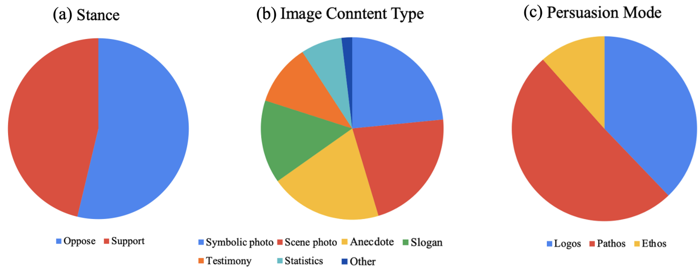


Figure 9: Distributions of (a) stance, (b) image content type, and (c) persuasion mode in our corpus on *gun control* topic.

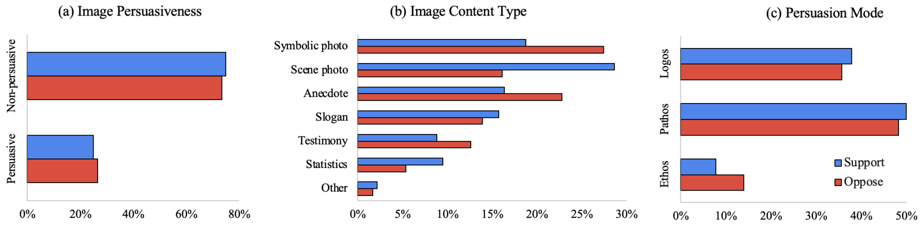


Figure 10: Distributions of (a) image persuasiveness, (b) content type and (c) persuasion mode regarding stances (support in blue and oppose in red) in our corpus on *gun control* topic.

Persuasiveness Score	Count	Percentage
0.0 - 0.1	336	33.50%
0.1 - 0.3	232	23.13%
0.3 - 0.5	176	17.55%
0.5 - 0.7	118	11.76%
0.7 - 0.9	66	6.58%
≥ 0.9	75	7.48%

Table 4: The annotated image persuasiveness score distribution on *gun control* topic in *ImageArg*.

"other" out of 1003 annotations that annotators were confused about; however, it does suggest that our image content type scheme works very well as only 1.89% are out of our defined labels. In terms of image persuasion mode, we only annotate images with persuasiveness score γ greater than 0.5, which produces 259 samples. As shown in Fig. 9 (c), we have 37.85% Logos, 50.60% Pathos, and 11.55% Ethos.

Additionally, we show how the stance impacts image persuasiveness, content type, and persuasion mode. In Fig. 10 (a), supporting and opposing *gun control* stance are almost evenly distributed with respect to persuasiveness and non-persuasiveness, which suggests that images generally support both positive and negative arguments. For the image content type in Fig. 10 (b), opposing *gun control* stance uses significantly more images with respect to Symbolic photos, Anecdote, and Testimony; however, supporting stance prefers images in the content of Scene photos and Statistics. Regarding persuasion mode in Fig. 10 (c), images in supporting *gun control* stance uses more Logos and Pathos but less Ethos than those in the opposing stance.

To further study the relevance between image content type and persuasion mode, we report their correlated distributions in charts. Fig. 11 (a) shows that most Logos samples use Statistics and Anecdote evidence. It meets the intuition that the logical reasoning can usually be clarified by introducing anecdotes and justified by providing supportive statistics. In terms of Pathos in Fig. 11 (b), the majority of samples utilize Scene and Symbolic photos. This is also reasonable since images generally promote emotional impression by presenting visual information. Regarding Ethos, Fig. 11 (c) shows Testimony takes the most ratio because statements from authorities can enhance trustworthiness. These correlations imply mutual influences between different annotation dimensions and raise demands for further study.

5 Experiments

5.1 Models and Tasks

We evaluate our corpus on *gun control* topic with binary classification tasks for Stance, Persuasiveness, Logos, Pathos, and Ethos and multi-class classification task for Image Content. Since data size is relatively small, we use pretrained image encoder ResNet50 (He et al., 2016) and text encoder BERT (Devlin et al., 2019) to fine-tune linear classifiers.

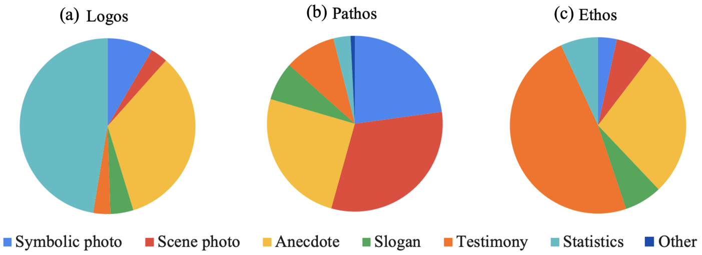


Figure 11: Distributions of image content type in different persuasion mode (a) Logos, (b) Pathos, and (c) Ethos in our corpus on *gun control* topic.

For fair comparison, we project both image and text embeddings into 1024 dimension before feeding into classification layers. We compare task performance on Text Modality (T-M), Image Modality (I-M), and Image-Text Multi-modality (M-M) that concatenates T-M and I-M. As for baseline (BASE), we report the performance when all samples are predicted as positive for binary classification, or predicted as the majority label for multi-class classification. We don't use the majority baseline for the binary classification task because the recall and F1 scores are always 0 if the majority label is negative, which is not interesting to compare with.

In the implementation, we follow the annotation strategy (Sec. 4.2) that uses threshold γ equal to 0.5 to encode persuasiveness scores into binaries. We remove Emoji, URLs, Mentions, and Hashtags in tweet texts, and discard 19 samples labeled with "Other" for the image content classification task. All images are resized to 224×224 dimension, and augmented (i.e., horizontal-flipped) only in training. Our models are implemented with Pytorch, and trained on a GeForce RTX 3080 GPU. We freeze BERT and ResNet50 encoders while training classifiers, and optimize the networks using Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$. The learning rate is 0.001 and the batch size is 16. We conduct 5-fold cross-validation (80% data in train; 20% data in test). We report 5-fold average Precision, Recall, F1, and AUC scores for binary classification and macro Precision, Recall, and F1 scores for multi-class classification on the test set.

5.2 Quantitative Results Analysis

Table 5 shows the classification benchmark results with standard deviation on *gun control* topic in *ImageArg* corpus.

Task-Stance Regarding stance, T-M has the highest performance in terms of AUC scores. It reveals that the image information is redundant to the text for identifying the stance; moreover, the image might introduce disturbing noise due to limited training samples.

Task	Model	Precision	Recall	F1	AUC
Stance (binary)	BASE	$0.470_{\pm 0.02}$	$\mathbf{1.000}_{\pm \mathbf{0.00}}$	$\mathbf{0.639}_{\pm \mathbf{0.02}}$	/
	T-M	$\mathbf{0.501}_{\pm \mathbf{0.05}}$	$0.740_{\pm 0.03}$	$0.596_{\pm 0.04}$	$\mathbf{0.527}_{\pm \mathbf{0.04}}$
	I-M	$0.443_{\pm 0.08}$	$0.147_{\pm 0.03}$	$0.218_{\pm 0.04}$	$0.472_{\pm 0.05}$
	M-M	$0.414_{\pm 0.04}$	$0.369_{\pm 0.06}$	$0.390_{\pm 0.05}$	$0.417_{\pm 0.03}$
Persua. (binary)	BASE	$0.257_{\pm 0.03}$	$\mathbf{1.000}_{\pm \mathbf{0.00}}$	$\mathbf{0.408}_{\pm \mathbf{0.04}}$	/
	T-M	$0.260_{\pm 0.01}$	$0.725_{\pm 0.11}$	$0.380_{\pm 0.01}$	$0.502_{\pm 0.03}$
	I-M	$\mathbf{0.313}_{\pm \mathbf{0.02}}$	$0.196_{\pm 0.05}$	$0.238_{\pm 0.05}$	$0.528_{\pm 0.03}$
	M-M	$0.296_{\pm 0.05}$	$0.486_{\pm 0.05}$	$0.364_{\pm 0.03}$	$\mathbf{0.534}_{\pm \mathbf{0.04}}$
Content (6-class)	BASE	$0.041_{\pm 0.00}$	$0.167_{\pm 0.00}$	$0.066_{\pm 0.00}$	/
	T-M	$0.198_{\pm 0.08}$	$0.201_{\pm 0.03}$	$\mathbf{0.165}_{\pm \mathbf{0.03}}$	/
	I-M	$\mathbf{0.235}_{\pm \mathbf{0.09}}$	$\mathbf{0.204}_{\pm \mathbf{0.02}}$	$0.151_{\pm 0.02}$	/
	M-M	$0.200_{\pm 0.02}$	$0.179_{\pm 0.01}$	$\mathbf{0.165}_{\pm \mathbf{0.01}}$	/
Logos (binary)	BASE	$\mathbf{0.405}_{\pm \mathbf{0.05}}$	$\mathbf{1.000}_{\pm \mathbf{0.00}}$	$\mathbf{0.575}_{\pm \mathbf{0.05}}$	/
	T-M	$0.364_{\pm 0.08}$	$0.613_{\pm 0.13}$	$0.456_{\pm 0.10}$	$0.439_{\pm 0.08}$
	I-M	$0.351_{\pm 0.22}$	$0.097_{\pm 0.07}$	$0.144_{\pm 0.10}$	$0.406_{\pm 0.08}$
	M-M	$0.262_{\pm 0.27}$	$0.047_{\pm 0.05}$	$0.077_{\pm 0.08}$	$\mathbf{0.508}_{\pm \mathbf{0.06}}$
Pathos (binary)	BASE	$0.554_{\pm 0.04}$	$\mathbf{1.000}_{\pm \mathbf{0.00}}$	$\mathbf{0.712}_{\pm \mathbf{0.04}}$	/
	T-M	$0.613_{\pm 0.11}$	$0.714_{\pm 0.08}$	$0.658_{\pm 0.09}$	$0.582_{\pm 0.10}$
	I-M	$\mathbf{0.666}_{\pm \mathbf{0.09}}$	$0.184_{\pm 0.07}$	$0.280_{\pm 0.07}$	$\mathbf{0.593}_{\pm \mathbf{0.09}}$
	M-M	$0.471_{\pm 0.42}$	$0.071_{\pm 0.10}$	$0.114_{\pm 0.15}$	$0.507_{\pm 0.12}$
Ethos (binary)	BASE	$0.128_{\pm 0.04}$	$\mathbf{1.000}_{\pm \mathbf{0.00}}$	$0.226_{\pm 0.06}$	/
	T-M	$0.168_{\pm 0.05}$	$0.817_{\pm 0.15}$	$\mathbf{0.272}_{\pm \mathbf{0.06}}$	$\mathbf{0.580}_{\pm \mathbf{0.09}}$
	I-M	$\mathbf{0.244}_{\pm \mathbf{0.16}}$	$0.233_{\pm 0.16}$	$0.221_{\pm 0.13}$	$0.459_{\pm 0.18}$
	M-M	$0.124_{\pm 0.15}$	$0.083_{\pm 0.11}$	$0.098_{\pm 0.12}$	$0.450_{\pm 0.09}$

Table 5: Classification benchmark results with standard deviation on *gun control* topic in *ImageArg* corpus. Note that the reported Persuasiveness results use threshold γ equal to 0.5. The Stance, Persuasiveness, and Image Content tasks use 1003 annotations; The Logos, Pathos, and Ethos use 259 annotations.

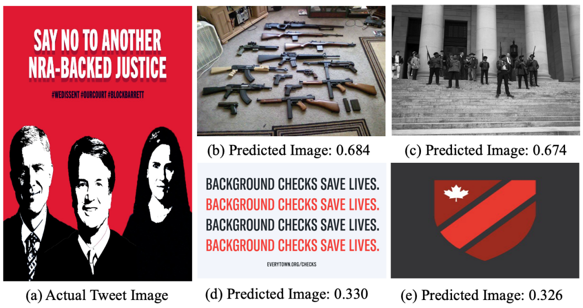
Task-Persuasiveness As for persuasiveness task, we observe that M-M performs slightly poorer than T-M regarding F1 score but relatively better in AUC score. This is because persuasiveness (positive/negative) labels are unbalanced if we use $\gamma = 0.5$ (as shown in Table 4). We show F1 scores drop with respect to threshold increases from 0.1 to 0.9 in Table 6.

Task-Content In terms of 6-class classification for image content, although all modalities outperform the baseline, the task is shown to be very challenging. It is surprising that the performance with I-M is lower than T-M. The reason might be that visual argumentative tasks demand more specific image encoders that learn sufficient knowledge on

Threshold (γ)	Pos. Ratio	F1 Score		
		T-M	I-M	M-M
0.1	66.50%	$\mathbf{0.681}_{\pm 0.02}$	$\mathbf{0.265}_{\pm 0.05}$	$\mathbf{0.536}_{\pm 0.03}$
0.3	43.37%	$0.538_{\pm 0.03}$	$0.251_{\pm 0.04}$	$0.459_{\pm 0.05}$
0.5	25.8%	$0.380_{\pm 0.01}$	$0.238_{\pm 0.05}$	$0.364_{\pm 0.03}$
0.7	14.1%	$0.246_{\pm 0.02}$	$0.168_{\pm 0.04}$	$0.233_{\pm 0.01}$
0.9	7.48%	$0.138_{\pm 0.03}$	$0.084_{\pm 0.03}$	$0.115_{\pm 0.01}$

Table 6: F1 scores with standard deviation and positive label ratio for Persuasiveness classification with respect to different threshold (γ).

Every day, 100 people are shot and killed. 40,000 are killed each year. But Judge Barrett would rule to gut both our gun safety laws and our access to healthcare. We don't need a reckless, undemocratic nominee who threatens our livelihoods! #WeDissent #BlockBarrett


SAY NO TO ANOTHER NRA-BACKED JUSTICE
#WEDISSENT #OURCOURT #BLOCKBARRETT
BACKGROUND CHECKS SAVE LIVES.
BACKGROUND CHECKS SAVE LIVES.
BACKGROUND CHECKS SAVE LIVES.
BACKGROUND CHECKS SAVE LIVES.
EVERYTOWN.ORG/CHECKS

(a) Actual Tweet Image (b) Predicted Image: 0.684 (c) Predicted Image: 0.674 (d) Predicted Image: 0.330 (e) Predicted Image: 0.326

Figure 12: (a) the actual tweet image annotated with persuasiveness score 0 in *ImageArg*; (b) and (c) with top predicted persuasiveness scores; (d) and (e) with lowest predicted persuasiveness scores while retrieving images given the same tweet text.

persuasiveness and social science; however, the used image encoder is pretrained on a general object detection task on the ImageNet (Krizhevsky et al., 2012), thus our model is unable to learn well for this argumentative task with very limited training data.

Task-Logos Regarding logos, we observe that M-M gains the best AUC score but I-M has lower AUC than T-M. The reason might be that logos images usually contain statistic charts, as shown in Fig. 11 (a), that are relatively more difficult to encode than normal images (e.g., images with explicit objects), but multi-modal models might learn these patterns directly from textual inputs.

Task-Pathos As for pathos, I-M has the best performance in terms of AUC score, and T-M is quite close to I-M while M-M has the lowest. This suggests that the multi-modal representation fusion method we used might be too weak to conduct complex reasoning on the pathos task.

Task-Ethos The best performance in ethos is from T-M. It is intuitive because the image encoder pre-trained on object detection is unable to recognize the optical characters on the image, while this kind of images are common in ethos, e.g., testimony images in Fig. 11 (c).

5.3 Qualitative Results Analysis

We conduct qualitative analysis by retrieving the most persuasive images given a text. Specifically, we run the multi-modality (M-M) model, trained for the persuasiveness task, on the test set in each fold (out of 5 folds). The inputs are image-text pairs of which all candidate images are paired with the same text, and the outputs are image persuasiveness scores. Fig. 12 shows the actual, top, and bottom images with the highest and lowest persuasiveness scores, respectively. It is interesting to find that images with specific objects or scenes (image (b), and (c) in Fig. 12) boost the persuasiveness scores; however, images with slogans or symbolism have lower scores (image (d), and (e) in Fig. 12). This suggests that our image encoder is capable of capturing object information but not optical characters on images (e.g., slogans); therefore, our retrieved images with best persuasion scores are mostly related to gun-object images. Thus, learning an image encoder pre-trained on slogans and visual symbolism is a promising future direction to improve the performance. In the meanwhile, extracting text information from images by OCR tools and use it as an auxiliary modality may help models learn the context.

6 Conclusion and Future Work

We create a brand-new multi-modal persuasiveness dataset *ImageArg* that focuses on image functionality and persuasion mode for persuasive arguments. We extend the argumentative annotation scheme from text to vision, and demonstrate its feasibility. We then establish a benchmark on our defined tasks using computational models, with multiple input modalities. Our experimental results reveal that image persuasiveness mining is challenging and that there is ample room for model improvement. We identify the image encoder as a key modeling bottleneck through a series of qualitative and quantitative analysis, which offers a good starting point for further exploration on this rich and challenging topic. The first version of *ImageArg* has 1003 annotations on the *gun control* topic. In the future work, we will work on constructing datasets on the topics of *immigration* and *abortion*, and scaling up the annotations.

References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. Cite: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770.

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.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29.

Antoine C Braet. 1992. Ethos, pathos and logos in aristotle's rhetoric: A re-examination. *Argumentation*, 6(3):307–320.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Abhisek Chakrabarty, Onkar Arun Pandit, and Utpal Garain. 2017. Context sensitive lemmatization using two successive bidirectional gated recurrent networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1481–1491.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen Mckeown, and Alyssa Hwang. 2019. Ampersand: Argument mining for persuasive online discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2933–2943.

Moitreya Chatterjee, Sunghyun Park, Han Suk Shim, Kenji Sagae, and Louis-Philippe Morency. 2014. Verbal behaviors and persuasiveness in online multimedia content. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 50–58.

Niladri Chatterjee and Saumya Agrawal. 2006. Word alignment in english-hindi parallel corpus using recency-vector approach: some studies. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 649–656.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rory Duthie, Katarzyna Budzynska, and Chris Reed. 2016. Mining ethos in political debate. In *Computational Models of Argument: Proceedings from the Sixth International Conference on Computational Models of Argument (COMMA)*, pages 299–310. IOS Press.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Nancy Green, Kevin D Ashley, Diane Litman, Chris Reed, and Vern Walker. 2014. Proceedings of the first workshop on argumentation mining. In *Proceedings of the First Workshop on Argumentation Mining*.

Meiqi Guo, Rebecca Hwa, and Adriana Kovashka. 2021. Detecting persuasive atypicality by modeling contextual compatibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 972–982.

Meiqi Guo, Rebecca Hwa, Yu-Ru Lin, and Wen-Ting Chung. 2020. Inflating topic relevance with ideology: A case study of political ideology bias in social topic detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4873–4885.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Lisa Hartling, Michele Hamm, Andrea Milne, Ben Vandermeer, P Lina Santaguida, Mohammed Ansari, Alexander Tsertsvadze, Susanne Hempel, Paul Shekelle, and Donna M Dryden. 2012. Validity and inter-rater reliability testing of quality assessment instruments.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Colin Higgins and Robyn Walker. 2012. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In *Accounting forum*, volume 36, pages 194–208. Elsevier.

Di Hu, Chengze Wang, Feiping Nie, and Xuelong Li. 2019. Dense multimodal fusion for hierarchically joint representation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3941–3945. IEEE.

Xinyue Huang and Adriana Kovashka. 2016. Inferring visual persuasion via body language, setting, and deep features. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 778–784.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. 2014. Visual persuasion: Inferring communicative intents of images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–223.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753.

Anastasios Lytos, Thomas Lagkas, Panagiotis Sarigiannidis, and Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management*, 56(6):102055.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Daniel J O'keefe. 2015. *Persuasion: Theory and research*. Sage Publications.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4082–4088.

Matilda White Riley. 1954. Communication and persuasion: psychological studies of opinion change.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *ArgNLP*, pages 21–25.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. In *International Conference on Learning Representations*.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200.

Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. 2018. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*.

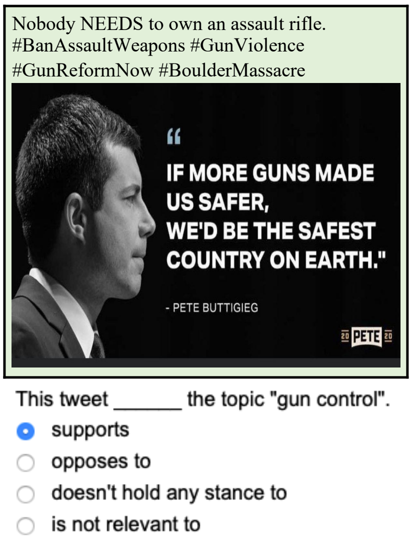
A Coding Manual

A.1 Stance

We setup different instructions for stance annotations on different topics since we would like to provide detailed instructions and examples for different topics separately.

A.1.1 Stance: Gun Control

We aim to study the topic and the stance of tweets. Given a tweet accompanying with an image, you need to answer the stance of the tweet towards a given topic, as depicted in Figure 13. Please make sure that you have the basic knowledge about that social topic and you understand the key message that the tweet (i.e. both the text and the image) sends. Just skip the HIT if you are not sure.



Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre

"IF MORE GUNS MADE US SAFER, WE'D BE THE SAFEST COUNTRY ON EARTH."

- PETE BUTTIGIEG

This tweet ______ the topic "gun control".

● supports

○ opposes to

○ doesn't hold any stance to

○ is not relevant to

Figure 13: Example of stance annotation on gun control

The question is about the stance. You need to decide whether the tweet is relevant or not to the social topic gun control. If it is relevant, then you need to annotate the stance: supports/opposes to/doesn't hold any stance.

A tweet is considered as relevant if it talks about anything that has to do with, but not limited to, the following issue categories: the Second Amendment, Gun control laws, etc. Tweets which contain the following hashtags are probably relevant to gun control: #NoBillNoBreak, #WearOrange, #EndGunViolence, #DisarmHate, #molonlabe, etc.

A tweet should be considered as irrelevant if it mentions a gun death event or a gun violence news, but the context is not necessarily about gun control.

Some examples for relevant tweets and their stance (we only show the text here, but you need to answer this question from both the text and image):

- *"Standing up for the second amendment and carrying a firearm for self defense."* This tweet asks the audience to stand up for the 2nd amendment, which opposes to gun control;
- *"I don't understand why we can't ban assault weapons. We all know they are only used for hunting people. #PrayForOrlando #guncontrolplease."* This tweet talks about banning weapons and contains the hashtag "#guncontrolplease", which supports gun control;
- *A common way to reduce violence in schools is to implement stronger security measures, such as surveillance cameras, security systems, campus guards and metal detectors. #violence #domesticviolence #gun #gunviolence #abuse #people #world #person #workplace."* This tweet is relevant to the topic, but we are not sure about its stance.

Some examples for non-relevant tweets (we only show the text here, but you need to answer this question from both the text and image):

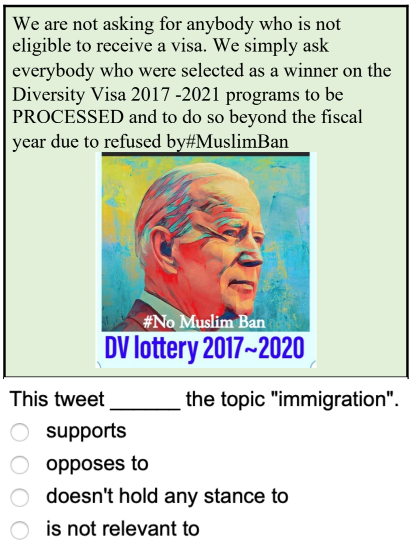
- *"Love will always conquer hate. #PrayForOrlando #OrlandoShooting."* This tweet talks about gun violence but not about gun control;
- *"#Gunviolence has serious and lasting social and emotional impacts on those who directly and indirectly experience it."* This tweet points out the impact of gun violence but not about gun control.

A.1.2 Stance: Immigration

We aim to study the topic and the stance of tweets. Given a tweet accompanying with an image, you need to answer the stance of the tweet towards a given topic, as depicted in Figure 14. Please make sure that you have the basic knowledge about that social topic and you understand the key message that the tweet (i.e. both the text and the image) sends. Just skip the HIT if you are not sure.

The question is about the stance. You need to decide whether the tweet is relevant or not to the social topic immigration. If it is relevant, then you need to annotate the stance: supports/opposes to/doesn't hold any stance.

A tweet is considered as relevant if it talks about anything that has to do with, but not limited to, the following issue categories: Borders,



We are not asking for anybody who is not eligible to receive a visa. We simply ask everybody who were selected as a winner on the Diversity Visa 2017 -2021 programs to be PROCESSED and to do so beyond the fiscal year due to refused by#MuslimBan

This tweet ______ the topic "immigration".

○ supports
○ opposes to
○ doesn't hold any stance to
○ is not relevant to

Figure 14: Example of stance annotation on immigration

Birthright citizenship, Immigrant Crime, DACA and the DREAM Act, Deportation debate, Economic impact, Immigration quotas, Immigrants' rights and access to services, Labor Market - American workers and employers, Law enforcement, Refugees, etc.

A tweet should be considered as irrelevant if it mentions a group of immigrant people such as Muslim, Syrian refugees but doesn't explicitly talk about immigration issues.

Some examples for relevant tweets and their stance (we only show the text here, but you need to answer this question from both the text and image):

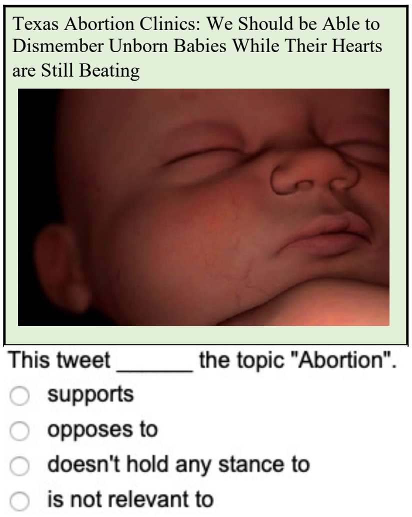
- *"Man feels bad for new immigrant driver in Brampton that crashed into his truck, causing $6K worth of damages - he had no licence or insurance"*. This tweet is related to the topic of immigration under the category of Immigrant Crime, and it opposes to immigration.
- *"House Bill 3438 will finally give our immigrant students some desperately needed resources! Thank you State Representative Maura Hirschauer for introducing this bill! Now, let's make sure this bill becomes law!"* This tweet is related to the topic of immigration under the category of DREAM Act, and it supports immigration.
- *"I'm a woman that supports Trump to fix economy, immigration, school, military more. #MAGA3X"* We consider a tweet as relevant even if it mentions several topics in addition to immigration, and it opposes to immigration.

Some examples for non-relevant tweets (we only show the text here, but you need to answer this question from both the text and image):

- *"'Will I die, miss?' Terrified Syrian boy suffers suspected gas attack."* This tweet talks about a Syrian boy suffering a gas attack, which may be pointing to a war or terrorist event in Syria, not necessarily directly about an immigration issue.
- *"Virtual tour of Steinbach, in partnership with MANSO, Welcome Place, Eastman Immigrant Services and the Steinbach LIP, coming up March 9th, 2021. It's free so don't miss out!"* This tweet mentions Immigrant Services, but does not talk about any immigration issue.
- *"I called on [USERNAME] for increased vaccine access for South Philadelphia seniors and for members of our immigrant communities. We can't let physical distance and language barriers keep people from this lifesaving vaccine."* This tweet talks about vaccine access for the immigrant community but it doesn't hold any stance towards any immigration policy.

A.1.3 Stance: Abortion

We aim to study the topic and the stance of tweets. Given a tweet accompanying with an image, you need to answer the stance of the tweet towards a given topic, as depicted in Figure 15. Please make sure that you have the basic knowledge about that social topic and you understand the key message that the tweet (i.e. both the text and the image) sends. Just skip the HIT if you are not sure.



Texas Abortion Clinics: We Should be Able to Dismember Unborn Babies While Their Hearts are Still Beating

This tweet ______ the topic "Abortion".

○ supports
○ opposes to
○ doesn't hold any stance to
○ is not relevant to

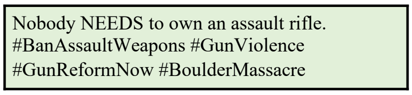
Figure 15: Example of stance annotation on abortion

The question is about the stance. You need to decide whether the tweet is relevant or not to the social topic abortion. If it is relevant, then you need to annotate the stance: supports/opposes to/doesn't hold any stance.

A tweet is considered as relevant if it talks about anything that discusses whether the abortion should be a legal option. If the arguments in the tweet text and image support that the abortion should be a legal option, then please choose "supports"; if arguments oppose to legal abortion, then choose "opposes to"; if arguments doesn't hold any stance for the topic then choose "doesn't hold any stance". Notice that a tweet is considered as irrelevant if it doesn't directly discuss whether the abortion should be a legal option or not, even though it may talk about related topics such as babies born alive after an abortion, birth control, etc.

A.2 Persuasiveness level and image content

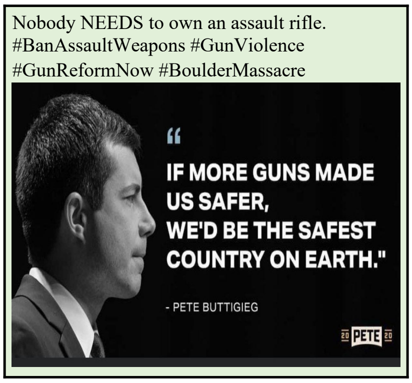
We aim to study the persuasiveness level of images in tweets as well as their content. Given a tweet text shown as Figure 16, you need to give a persuasiveness score of it.



Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre

Figure 16: Example of a text only tweet

Then given a tweet accompanying an image shown as Figure 17, you need to give a persuasiveness score again.



Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre

"
IF MORE GUNS MADE US SAFER, WE'D BE THE SAFEST COUNTRY ON EARTH."
- PETE BUTTIGIEG
20 PETE 20

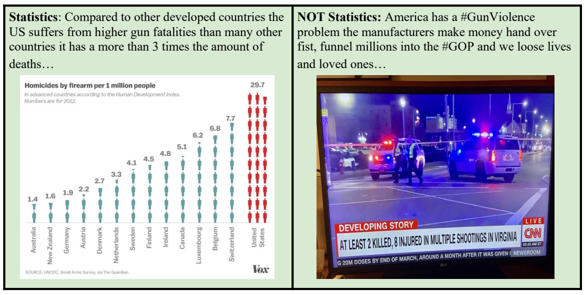
Figure 17: Example of a tweet accompanying an image

Finally, you need to select the content type of the image. The content type of an image represents what type of the information the image mainly carries. Specifically, you need to pick one out of six types below for each image.

Statistics: the image provides evidence by **stating or quoting quantitative information**, such as a chart/data analysis, that is related to the tweet text.

An image could be considered statistics if: 1) It carries quantitative information (number/statistics/etc). 2) The key purpose of the image is to deliver this quantitative information, in the case there are multiple content types involved.

For the examples shown in Figure 18, in the statistics example, the image mainly shows a chart and delivers quantitative information (homicides by firearm per 1 million people). In contrast, in the NOT statistics example, though there are numbers in the image, the main information is a news title and the shooting scene, but not these numbers.



Statistics: Compared to other developed countries the US suffers from higher gun fatalities than many other countries it has a more than 3 times the amount of deaths…

NOT Statistics: America has a #GunViolence problem the manufacturers make money hand over fist, funnel millions into the #GOP and we loose lives and loved ones…

Homicides by firearm per 1 million people
In advanced countries according to the Human Development Index. Numbers are for 2012.
1.4
1.6
1.9
2.2
2.7
3.3
4.1
4.5
4.8
5.1
6.2
6.8
7.7
29.7
Australia
New Zealand
Germany
Austria
Denmark
Netherlands
Sweden
Finland
Ireland
Canada
Luxembourg
Belgium
Switzerland
United States
SOURCE: UNODC, Small Arms Survey, via The Guardian.
Vox
DEVELOPING STORY
AT LEAST 2 KILLED, 8 INJURED IN MULTIPLE SHOOTINGS IN VIRGINIA
LIVE
CNN
10:15 AM ET
G 20M DOSES BY END OF MARCH, AROUND A MONTH AFTER IT WAS GIVEN E NEWSROOM

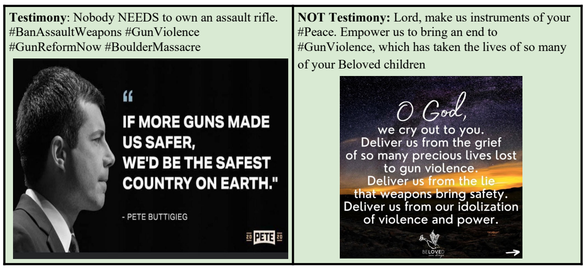
Figure 18: Example of tweets with statistics image and a non-statistics image.

Testimony: the image **quotes statements or conclusions from an authority**, such as a piece of an article/claim from an official document, that is related to the tweet text.

The image can be considered as testimony if: 1) The content contains texts such as statements/conclusions/pieces of article. 2) These texts are original from other resources such as news/celebrities/official documents/etc. 3) The key purpose of the image is to quote the authorized statement, in the case there are multiple content types involved.

For the examples shown in Figure 19, in the Testimony tweet example, the image mainly cites a statement given by the transportation secretary. However, in the NOT Testimony tweet example, though it contains a piece of texts, these texts are not cited from an authority, therefore, it is not testimony.

Anecdote: the image provides information based on the **author's personal experience**, such as facts/personal stories, that are related to the tweet text.



Testimony: Nobody NEEDS to own an assault rifle. #BanAssaultWeapons #GunViolence #GunReformNow #BoulderMassacre

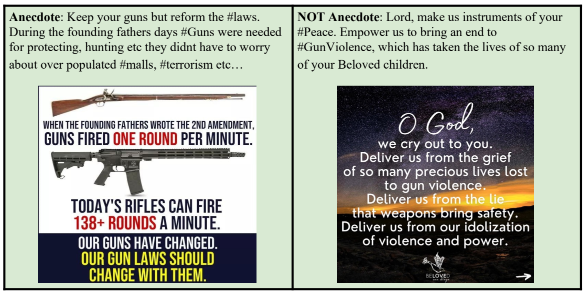
NOT Testimony: Lord, make us instruments of your #Peace. Empower us to bring an end to #GunViolence, which has taken the lives of so many of your Beloved children

"
IF MORE GUNS MADE US SAFER, WE'D BE THE SAFEST COUNTRY ON EARTH."
- PETE BUTTIGIEG
PETE
O God,
we cry out to you.
Deliver us from the grief
of so many precious lives lost
to gun violence.
Deliver us from the lie
that weapons bring safety.
Deliver us from our idolization
of violence and power.

Figure 19: Example of tweets with testimony image and a non-testimony image.

An image can be considered as an anecdote if: 1) It delivers a personal experience, Or 2) it shows a fact/experience that comes from personal view/known by the author. 3) The key purpose of the image is to deliver personal experience, in the case there are multiple content types involved.

For the examples shown in Figure 20, the anecdote image shows the personal view on the fact that guns have been developed since the period of the 2nd amendment, and therefore the laws for guns should be developed as well. However, in the NOT anecdote example, though it comes from a personal statement, it does not describe any fact/experience/stories.



Anecdote: Keep your guns but reform the #laws. During the founding fathers days #Guns were needed for protecting, hunting etc they didnt have to worry about over populated #malls, #terrorism etc…

NOT Anecdote: Lord, make us instruments of your #Peace. Empower us to bring an end to #GunViolence, which has taken the lives of so many of your Beloved children.

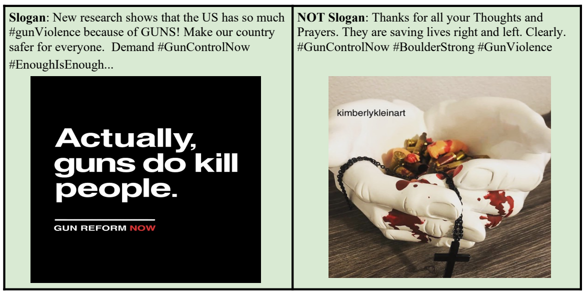
WHEN THE FOUNDING FATHERS WROTE THE 2ND AMENDMENT,
GUNS FIRED ONE ROUND PER MINUTE.
TODAY'S RIFLES CAN FIRE
138+ ROUNDS A MINUTE.
OUR GUNS HAVE CHANGED.
OUR GUN LAWS SHOULD CHANGE WITH THEM.
O God,
we cry out to you.
Deliver us from the grief
of so many precious lives lost
to gun violence.
Deliver us from the lie
that weapons bring safety.
Deliver us from our idolization
of violence and power.

Figure 20: Example of tweets with anecdote image and a non-anecdote image.

Slogan: the image expresses a piece of **advertising phrase**.

An image can be considered as a slogan if: 1) It mainly delivers a piece of text as slogan; 2) The text is for advertising purposes as an advertising phrase/claim/statement. 3) The key purpose of the image is to deliver the piece of text, in the case there are multiple content types involved.

For the examples shown in Figure 21, the slogan image presents a phrase "Actually guns do kill people. Gun Reform Now", therefore it is a slogan. However, For the example of NOT Slogan, though the image is for advertising, it does not contain a phrase for that, therefore it is not a slogan.



Slogan: New research shows that the US has so much #gunViolence because of GUNS! Make our country safer for everyone. Demand #GunControlNow #EnoughIsEnough...

NOT Slogan: Thanks for all your Thoughts and Prayers. They are saving lives right and left. Clearly. #GunControlNow #BoulderStrong #GunViolence

Figure 21: Example of tweets with slogan image and a non-slogan image.

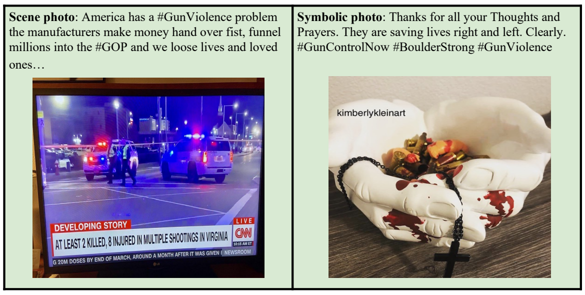
Scene photo: the image shows a **literal scene/photograph** that is related to the tweet text.

An image can be considered as a scene photo if: 1) It shows a literal photograph/scene. 2) The image is directly related to the text. 3) The key purpose of the image is to deliver the image content but not the text within, in the case there are multiple content types involved.

Symbolic photo: the image shows a **symbol/art** that expresses the author's viewpoints in a **non-literal** way.

An image can be considered as a symbolic photo if: 1) It shows a symbol/art. 2) It expresses the viewpoint from the author in an implicit way. 3) The key purpose of the image is to deliver the image content but not the text within, in the case there are multiple content types involved.

For example, in Figure 22, the scene photo image shows a real photograph of a gun violence scene reported by CNN news. In the Symbolic photo, though relevant to the text, it shows a photo/image that is related to the text in a non-literal way (blood signifies gun-killing and the hand posture signifies praying), therefore it is not a scene photo but a symbolic photo.



Scene photo: America has a #GunViolence problem the manufacturers make money hand over fist, funnel millions into the #GOP and we loose lives and loved ones…

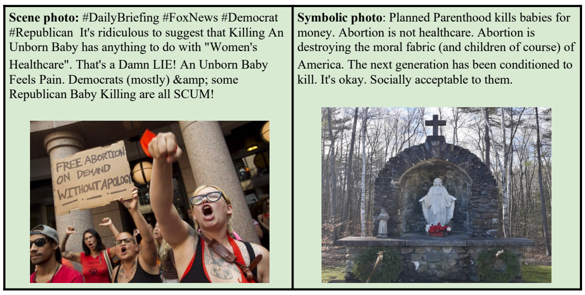
Symbolic photo: Thanks for all your Thoughts and Prayers. They are saving lives right and left. Clearly. #GunControlNow #BoulderStrong #GunViolence

DEVELOPING STORY
AT LEAST 2 KILLED, 8 INJURED IN MULTIPLE SHOOTINGS IN VIRGINIA
LIVE
CNN
kimberlykleinart

Figure 22: Example of tweets with scene photo image and a symbolic photo image.

The key difference between the Scene photo and Symbolic photo is **whether the photograph**

sends a message literally or symbolically. For a scene photo, the image directly expresses/supports the author's view without any rhetoric; for a symbolic photo, the image may have several possible interpretations and the audience can understand its symbolic meaning after considering the tweet text. Consider the example shown in Figure 23: for the scene photo, it directly shows a protest scene and the author opposes to the abortion by considering it as a lie. In the symbolic photo, the author shows a photo of Notre Dame as a symbol of anti-abortion. The photo is not directly related to abortion, but audience can understand its symbolic meaning after reading the text.



Scene photo: #DailyBriefing #FoxNews #Democrat #Republican It's ridiculous to suggest that Killing An Unborn Baby has anything to do with "Women's Healthcare". That's a Damn LIE! An Unborn Baby Feels Pain. Democrats (mostly) &amp; some Republican Baby Killing are all SCUM!

Symbolic photo: Planned Parenthood kills babies for money. Abortion is not healthcare. Abortion is destroying the moral fabric (and children of course) of America. The next generation has been conditioned to kill. It's okay. Socially acceptable to them.

Figure 23: Another example of tweets with scene photo image and a symbolic photo image.

In the case there are multiple content types involved: You need to first identify the key purpose of the image (i.e. what is the most important information in the image). Then please select the content type of the key purpose. Table 7 shows the summary of content types for each key purpose.

Table 7: Summary of content types for each key purpose

Key Purpose		**Content Type**
Quantitative information in the image		Statistics
Textual information in the image	Statements or conclusions from an authority	Testimony
	Personal experiences/views	Anecdote
	Advertising phrases	Slogan
Graphical information in the image	Literal photograph	Scene Photo
	Non-literal/rhetorical photograph	Symbolic Photo

A.3 Persuasion Mode

We aim to study the **argumentative roles of images** in tweets. Given a tweet accompanying an image, we would ask you to choose the persuasion mode of the image. The persuasion mode of an image represents how the image convinces the audience. Specifically, we will ask you whether the image appeals to logic/emotion/credibility. Additionally, we will ask you why you make the choices.

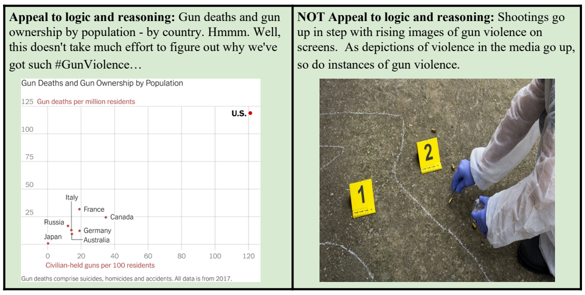
Q1: Does the image make the tweet more persuasive by appealing to **logic and reasoning**?

The image appeals to logic and reasoning if it persuades audiences with reasoning from a fact/statistics/study case/scientific evidence. Specifically, if: 1) the image **contains information for logic and reasoning**; 2) the image **presents logic and reasoning**.

Also, we will ask you why you made the choice. i.e. Describing the logic/reasoning brought by the image. Such as following, by filling the blank in the textbox:

The logic/reasoning of the image is [the correlation between gun deaths and gun ownership by population].

For example shown in Figure 24, the left image provides a chart that shows the high gun deaths and the high gun ownership by the population of the US, which implies [a correlation between gun death and gun ownership which demonstrates that there will be less gun deaths with gun control.]. On the contrary, the right image shows the scene of the shooting but does not provide any reasoning or logic.



Appeal to logic and reasoning: Gun deaths and gun ownership by population - by country. Hmmm. Well, this doesn't take much effort to figure out why we've got such #GunViolence…

NOT Appeal to logic and reasoning: Shootings go up in step with rising images of gun violence on screens. As depictions of violence in the media go up, so do instances of gun violence.

Gun Deaths and Gun Ownership by Population
Gun deaths per million residents
125
100
75
50
25
U.S.
Italy
France
Canada
Russia
Germany
Japan
Australia
0
20
40
60
80
100
120
Civilian-held guns per 100 residents
Gun deaths comprise suicides, homicides and accidents. All data is from 2017.
1
2

Figure 24: Example of tweets with logos image and non-logos image.

Q2: Does image make the tweet more persuasive by appealing to **emotion**?

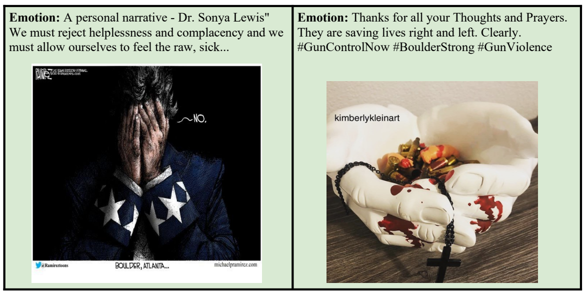
The image appeals to **emotion**, if it puts audiences in a certain frame of mind by stimulating them to identify/empathize/sympathize with the arguments.

Specifically, if : 1) the image **invokes the audience with strong emotion**, such as sadness, happiness, compassion, worriness; 2) the image **makes the audience identify/empathize/sympathize** with the author/arguments.

Also, we will ask you why you made the choice. i.e. Describing the emotion(such as anger/amusement/sad/etc.) or impulsion(desire to do something) brought by the image. Such as following, by filling the blank within the [bracket]:

The image evokes my emotion/impulse of *[anger]*.

For example shown in Figure 25, the left image shows the grieved ”Uncle Sam” saying ”no” with helplessness, which evokes the [desire for gun control]. The right image provides an item that can revoke [compassion and forgiveness].



Emotion: A personal narrative - Dr. Sonya Lewis" We must reject helplessness and complacency and we must allow ourselves to feel the raw, sick...

Emotion: Thanks for all your Thoughts and Prayers. They are saving lives right and left. Clearly. #GunControlNow #BoulderStrong #GunViolence

Figure 25: Example of tweets with pathos images.

Q3: Does image make the tweet more persuasive by **enhancing credibility and trustworthiness**?

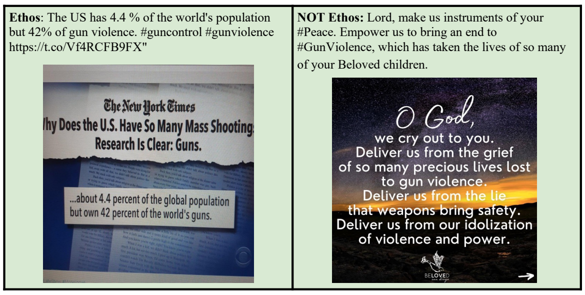
The image **enhances credibility and trustworthiness**, if it makes people trust something more via authorized/trusted expertise/title/reputation.

Specifically, if 1) The image **cites reliable sources** of the event/story/opinion/stance, that can make the contents trustworthy. Reliable sources include news, research reports, celebrated dictum, etc. Sources which are not proved/well-known by the audience (.e.g. an organization logo) are not considered as reliable. 2) the image **shows authorities** that can convince the audience to believe the arguments.

Also, we will ask you why you made the choice. i.e. Describing the resources of the citation that enhances the credibility. Such as following, by filling the blank within the [bracket]:

The credibility is enhanced by [a citation to political report]

For example shown in Figure 26, the left image takes a screenshot of the source of a report from [New York Times], which increases credibility. The NOT Ethos right image shows the views but are not quoted sentences that do not provide the credibility to enhance the argument.



Ethos: The US has 4.4 % of the world's population but 42% of gun violence. #guncontrol #gunviolence https://t.co/Vf4RCFB9FX"

The New York Times
Why Does the U.S. Have So Many Mass Shootings
Research Is Clear: Guns.

...about 4.4 percent of the global population but own 42 percent of the world's guns.

NOT Ethos: Lord, make us instruments of your #Peace. Empower us to bring an end to #GunViolence, which has taken the lives of so many of your Beloved children.

O God,
we cry out to you.
Deliver us from the grief
of so many precious lives lost
to gun violence.
Deliver us from the lie
that weapons bring safety.
Deliver us from our idolization
of violence and power.

Figure 26: Example of tweets with ethos image and non-ethos image.