SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images

Dimitar Dimitrov,[1] Bishr Bin Ali,[2] Shaden Shaar,[3] Firoj Alam,[3] Fabrizio Silvestri,[4] Hamed Firooz,[5] Preslav Nakov, [3] and Giovanni Da San Martino[6]
[1] Sofia University "St. Kliment Ohridski", Bulgaria, [2] King's College London, UK,
[3] Qatar Computing Research Institute, HBKU, Qatar
[4] Sapienza University of Rome, Italy, [5] Facebook AI, USA, [6] University of Padova, Italy
mitko.bg.ss@gmail.com, bishrkc@gmail.com
{sshaar, fialam, pnakov}@hbku.edu.qa, mhfirooz@fb.com
fsilvestri@diag.uniroma1.it, dasan@math.unipd.it

Abstract

We describe *SemEval-2021 task 6 on Detection of Persuasion Techniques in Texts and Images*: the data, the annotation guidelines, the evaluation setup, the results, and the participating systems. The task focused on memes and had three subtasks: (*i*) detecting the techniques in the text, (*ii*) detecting the text spans where the techniques are used, and (*iii*) detecting techniques in the entire meme, i.e., both in the text and in the image. It was a popular task, attracting 71 registrations, and 22 teams that eventually made an official submission on the test set. The evaluation results for the third subtask confirmed the importance of both modalities, the text and the image. Moreover, some teams reported benefits when not just combining the two modalities, e.g., by using early or late fusion, but rather modeling the interaction between them in a joint model.

1 Introduction

Internet and social media have amplified the impact of disinformation campaigns. Traditionally a monopoly of states and large organizations, now such campaigns have become within the reach of even small organisations and individuals (Da San Martino et al., 2020b).

Such propaganda campaigns are often carried out using posts spread on social media, with the aim to reach very large audience. While the rhetorical and the psychological devices that constitute the basic building blocks of persuasive messages have been thoroughly studied (Miller, 1939; Weston, 2008; Torok, 2015), only few isolated efforts have been made to devise automatic systems to detect them (Habernal et al., 2018; Habernal et al., 2018; Da San Martino et al., 2019b).

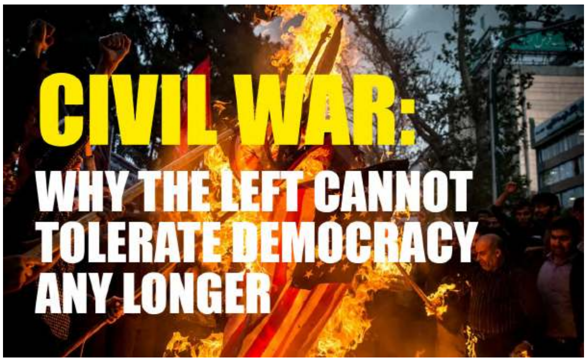


Figure 1: A meme with a civil war threat during the President Trump's impeachment trial. Two persuasion techniques are used: (*i*) *Appeal to Fear* in the image, and (*ii*) *Exaggeration* in the text. ***Source(s):*** Image ; License

Thus, in 2020, we proposed *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles*, with the aim to help bridge this gap (Da San Martino et al., 2020a). The task focused on text only. Yet, some of the most influential posts in social media use memes, as shown in Figure 1,[1] where visual cues are being used, along with text, as a persuasive vehicle to spread disinformation (Shu et al., 2017). During the 2016 US Presidential campaign, malicious users in social media (bots, cyborgs, trolls) used such memes to provoke emotional responses (Guo et al., 2020).

In 2021, we introduced a new SemEval shared task, for which we prepared a multimodal corpus of memes annotated with an extended set of techniques, compared to SemEval-2020 task 11. This time, we annotated both the text of the memes, highlighting the spans in which each technique has been used, as well as the techniques appearing in the visual content of the memes.

WARNING: This paper contains meme examples and wording that might be offensive to some readers.

[1] In order to avoid potential copyright issues, all memes we show in this paper are our own recreation of existing memes, using images with clear copyright.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 70–98
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

Based on our annotations, we offered the following three subtasks:

Subtask 1 (ST1) Given the textual content of a meme, identify which techniques (out of 20 possible ones) are used in it. This is a multilabel classification problem.

Subtask 2 (ST2) Given the textual content of a meme, identify which techniques (out of 20 possible ones) are used in it together with the span(s) of text covered by each technique. This is a multilabel sequence tagging task.

Subtask 3 (ST3) Given a meme, identify which techniques (out of 22 possible ones) are used in the meme, considering both the text and the image. This is a multilabel classification problem.

A total of 71 teams registered for the task, 22 of them made an official submission on the test set and 15 of the participating teams submitted a system description paper.

2 Related Work

Propaganda Detection Previous work on propaganda detection has focused on analyzing textual content (Barrón-Cedeno et al., 2019; Da San Martino et al., 2019b; Rashkin et al., 2017). See (Martino et al., 2020) for a recent survey on computational propaganda detection. Rashkin et al. (2017) developed the `TSHP-17` corpus, which had document-level annotations with four classes: *trusted*, *satire*, *hoax*, and *propaganda*. Note that `TSHP-17` was labeled using distant supervision, i.e., all articles from a given news outlet were assigned the label of that news outlet. The news articles were collected from the English Gigaword corpus (which covers reliable news sources), as well as from seven unreliable news sources, including two propagandistic ones. They trained a model using word n-grams, and reported that it performed well only on articles from sources that the system was trained on, and that the performance degraded quite substantially when evaluated on articles from unseen news sources. Barrón-Cedeno et al. (2019) developed a corpus `QProp` with two labels (propaganda vs. non-propaganda), and experimented with two corpora: `TSHP-17` and `QProp` . They binarized the labels of `TSHP-17` as follows: propaganda *vs.* the other three categories.

They performed massive experiments, investigated writing style and readability level, and trained models using logistic regression and SVMs. Their findings confirmed that using distant supervision, in conjunction with rich representations, might encourage the model to predict the source of the article, rather than to discriminate propaganda from non-propaganda. The study by Habernal et al. (2017, 2018) also proposed a corpus with 1.3k arguments annotated with five fallacies, including *ad hominem*, *red herring*, and *irrelevant authority*, which directly relate to propaganda techniques.

A more fine-grained propaganda analysis was done by Da San Martino et al. (2019b), who developed a corpus of news articles annotated with the spans of use of 18 propaganda techniques, from an invetory they put together. They targeted two tasks: (*i*) binary classification —given a sentence, predict whether any of the techniques was used in it; and (*ii*) multi-label multi-class classification and span detection task —given a raw text, identify both the specific text fragments where a propaganda technique is being used as well as the type of technique. They further proposed a multi-granular gated deep neural network that captures signals from the sentence-level task to improve the performance of the fragment-level classifier and vice versa. Subsequently, an automatic system, `Prta`, was developed and made publicly available (Da San Martino et al., 2020c), which performs fine-grained propaganda analysis of text using these 18 fine-grained propaganda techniques.

Multimodal Content Another line of related research is on analyzing multimodal content, e.g., for predicting misleading information (Volkova et al., 2019), for detecting deception (Glenski et al., 2019), emotions and propaganda (Abd Kadir et al., 2016), hateful memes (Kiela et al., 2020), and propaganda in images (Seo, 2014). Volkova et al. (2019) developed a corpus of 500K Twitter posts consisting of images and labeled with six classes: disinformation, propaganda, hoaxes, conspiracies, clickbait, and satire. Glenski et al. (2019) explored multilingual multimodal content for deception detection. Multimodal hateful memes were the target of the *Hateful Memes Challenge*, which was addressed by fine-tuning state-of-art methods such as ViLBERT (Lu et al., 2019), Multimodal Bitransformers (Kiela et al., 2019), and VisualBERT (Li et al., 2019) to classify hateful vs. not-hateful memes (Kiela et al., 2020).

Related Shared Tasks The present shared task is closely related to *SemEval-2020 task 11 on Detection of Persuasion Techniques in News Articles* (Da San Martino et al., 2020a), which focused on news articles, and asked (*i*) to detect the spans where propaganda techniques are used, as well as (*ii*) to predict which propaganda technique (from an inventory of 14 techniques) is used in a given text span. Another closely related shared task is the *NLP4IF-2019 task on Fine-Grained Propaganda Detection*, which asked to detect the spans of use in news articles of each of 18 propaganda techniques (Da San Martino et al., 2019a). While these tasks focused on the text of news articles, here we target memes and multimodality, and we further use an extended inventory of 22 propaganda techniques.

Other related shared tasks include the FEVER 2018 and 2019 tasks on *Fact Extraction and VERification* (Thorne et al., 2018), the SemEval 2017 and 2019 tasks on predicting the veracity of rumors in Twitter (Derczynski et al., 2017; Gorrell et al., 2019), the SemEval-2019 task on *Fact-Checking in Community Question Answering Forums* (Mihaylova et al., 2019), the NLP4IF-2021 shared task on *Fighting the COVID-19 Infodemic* (Shaar et al., 2021). We should also mention the CLEF 2018–2021 *CheckThat!* lab (Nakov et al., 2018; Elsayed et al., 2019a,b; Barrón-Cedeño et al., 2020; Barrón-Cedeño et al., 2020), which featured tasks on automatic identification (Atanasova et al., 2018, 2019) and verification (Barrón-Cedeño et al., 2018; Hasanain et al., 2019, 2020; Shaar et al., 2020; Nakov et al., 2021) of claims in political debates and social media. While these tasks focused on factuality, check-worthiness, and stance detection, here we target propaganda; moreover, we focus on memes and on multimodality rather than on analyzing the text of tweets, political debates, or community question answering forums.

3 Persuasion Techniques

Scholars have proposed a number of inventories of persuasion techniques of various sizes (Miller, 1939; Torok, 2015; Abd Kadir and Sauffiyan, 2014). Here, we use an inventory of 22 techniques, borrowing from the lists of techniques described in (Da San Martino et al., 2019b), (Shah, 2005) and (Abd Kadir and Sauffiyan, 2014). Among these 22 techniques, the first 20 are applicable to both text and images, while the last two, *Appeal to (Strong) Emotions* and *Transfer*, are reserved for images.

Below, we provide a definition for each of these 22 techniques; more detailed instructions of the annotation process and examples are provided in Appendix A.

1. **Loaded Language:** Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.

2. **Name Calling or Labeling:** Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable, or loves, praises.

3. **Doubt:** Questioning the credibility of someone or something.

4. **Exaggeration or Minimisation:** Either representing something in an excessive manner, e.g., making things larger, better, worse ("*the best of the best*", "*quality guaranteed*"), or making something seem less important or smaller than it really is, e.g., saying that an insult was just a joke.

5. **Appeal to Fear or Prejudices:** Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases, the support is built based on preconceived judgments.

6. **Slogans:** A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

7. **Whataboutism:** A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

8. **Flag-Waving:** Playing on strong national feeling (or positive feelings toward any group, e.g., based on race, gender, political preference) to justify or promote an action or idea.

9. **Misrepresentation of Someone's Position (Straw Man):** When an opponent's proposition is substituted with a similar one, which is then refuted in place of the original proposition.

10. **Causal Oversimplification:** Assuming a single cause or reason, when there are actually multiple causes for an issue. It includes transferring blame to one person or group of people without investigating the actual complexities of the issue.

11. **Appeal to Authority:** Stating that a claim is true because a valid authority or expert on the issue said so, without any other supporting evidence offered. We consider the special case in which the reference is not an authority or an expert as part of this technique, although it is referred to as *Testimonial* in the literature.
12. **Thought-Terminating Cliché:** Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract the attention away from other lines of thought.
13. **Black-and-White Fallacy or Dictatorship:** Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (*Dictatorship*).
14. **Reductio ad Hitlerum:** Persuading an audience to disapprove of an action or an idea by suggesting that the idea is popular with groups that are hated or in contempt by the target audience. It can refer to any person or concept with a negative connotation.
15. **Repetition:** Repeating the same message over and over again, so that the audience will eventually accept it.
16. **Obfuscation, Intentional Vagueness, Confusion:** Using words that are deliberately not clear, so that the audience can have their own interpretations.
17. **Presenting Irrelevant Data (Red Herring):** Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.
18. **Bandwagon** Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action."
19. **Smears:** A smear is an effort to damage or call into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups.
20. **Glittering Generalities (Virtue):** These are words or symbols in the value system of the target audience that produce a positive image when attached to a person or an issue.
21. **Appeal to (Strong) Emotions:** Using images with strong positive/negative emotional implications to influence an audience.
22. **Transfer:** Also known as *Association*, this is a technique that evokes an emotional response by projecting positive or negative qualities (praise or blame) of a person, entity, object, or value onto another one in order to make the latter more acceptable or to discredit it.

4 Dataset

The annotation process is explained in detail in Appendix A, and in this section, we give a just brief summary.

We collected English memes from our personal Facebook accounts over several months in 2020 by following 26 public Facebook groups, which focus on politics, vaccines, COVID-19, and gender equality. We considered a meme to be a "*photograph style image with a short text on top of it*", and we removed examples that did not fit this definition, e.g., cartoon-style memes, memes whose textual content was strongly dominant or non-existent, memes with a single-color background image, etc. Then, we annotated the memes using our 22 persuasion techniques. For each meme, we first annotated its textual content, and then the entire meme. We performed each of these two annotations in two phases: in the first phase, the annotators independently annotated the memes; afterwards, all annotators met together with a consolidator to discuss and to select the final gold label(s).

The final annotated dataset consists of 950 memes: 687 memes for training, 63 for development, and 200 for testing. While the maximum number of sentences in a meme is 13, the average number of sentences per meme is just 1.68, as most memes contain very little text.

Table 1 shows the number of instances of each technique for each of the tasks. Note that *Transfer* and *Appeal to (Strong) Emotions* are not applicable to text, i.e., to Subtasks 1 and 2. For Subtasks 1 and 3, each technique can be present at most once per example, while in Subtask 2, a technique could appear multiple times in the same example. This explains the sizeable differences in the number of instances for some persuasion techniques between Subtasks 1 and 2: some techniques are over-used in memes, with the aim of making the message more persuasive, and thus they contribute higher counts to Subtask 2.

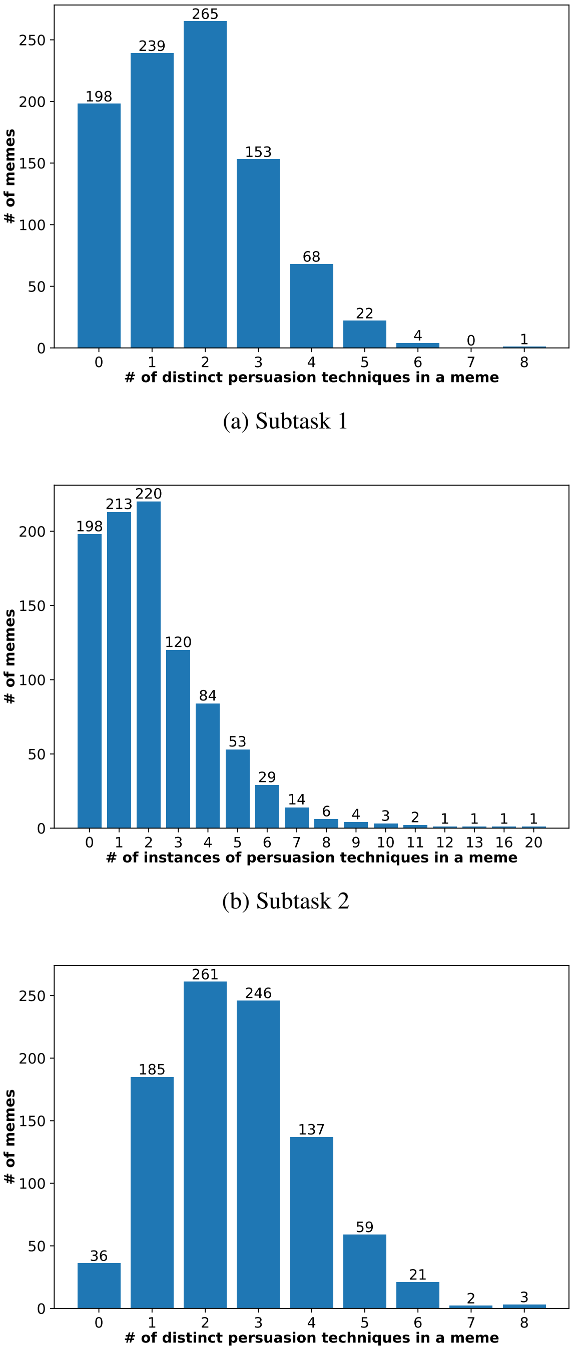
Persuasion Techniques	Subtask 1	Subtask 2		Subtask 3
	#	Len.	#	#
Loaded Language	489	2.41	761	492
Name Calling/Labeling	300	2.62	408	347
Smears	263	17.11	266	602
Doubt	84	13.71	86	111
Exaggeration/Minimisation	78	6.69	85	100
Slogans	66	4.70	72	70
Appeal to Fear/Prejudice	57	10.12	60	91
Whataboutism	54	22.83	54	67
Glittering Generalities (Virtue)	44	14.07	45	112
Flag-Waving	38	5.18	44	55
Repetition	12	1.95	42	14
Causal Oversimplification	31	14.48	33	36
Thought-Terminating Cliché	27	4.07	28	27
Black-and-White Fallacy/Dictatorship	25	11.92	25	26
Straw Man	24	15.96	24	40
Appeal to Authority	22	20.05	22	35
Reductio ad Hitlerum	13	12.69	13	23
Obfuscation, Intentional Vagueness, Confusion	5	9.8	5	7
Presenting Irrelevant Data	5	15.4	5	7
Bandwagon	5	8.4	5	5
Transfer	—	—	—	95
Appeal to (Strong) Emotions	—	—	—	90
Total	**1,642**		**2,119**	**2,488**

Table 1: Statistics about the persuasion techniques. For each technique, we show the average length of its spans (in number of words) and the number of its instances as annotated in the text only vs. in the entire meme.

Note that the number of instances for Subtasks 1 and 3 differs, and in some cases by quite a bit, e.g., for *Smears*, *Doubt*, and *Appeal to Fear/Prejudice*. This shows that many techniques cannot be found in the text, and require the visual content, which motivates the need for multimodal approaches for Subtask 3. Note also that different techniques have different span lengths, e.g., *Loaded Language* and *Name Calling* are about 2–3 words long, e.g., *violence*, *mass shooter*, and *coward*. However, for techniques such as *Whataboutism*, the average span length is 22 words.

Figure 2 shows statistics about the distribution of the number of persuasion techniques per meme. Note the difference for memes without persuasion techniques between Figures 2a and 2c: we can see that the number of memes without any persuasion technique drastically drops for Subtask 3. This is because the visual modality introduces additional context that was not available during the text-only annotation, which further supports the need for multimodal analysis. The visual modality also has an impact on memes that already had persuasion techniques in the text-only phase.

We observe that the number of memes with only one persuasion technique in Subtask 3 is considerably lower compared to Subtask 1, while the number of memes with three or more persuasion techniques has greatly increased for Subtask 3.



(a) Subtask 1

(b) Subtask 2

(c) Subtask 3

Figure 2: Distribution of the number of persuasion techniques per meme. Subfigure (b) reports the **number of instances** of persuasion techniques for a meme. Note that a meme could have multiple instances of the same technique for this subtask. Subfigures (a) and (c) show the number of **distinct** persuasion techniques in a meme.

5 Evaluation Framework

5.1 Evaluation Measures

Subtasks 1 and 3 To measure the performance of the systems, for Subtasks 1 and 3, we use Micro and Macro F_1, as these are multi-class multi-label tasks, where the labels are imbalanced. The official measure for the task is Micro F_1.

Subtask 2 For Subtask 2, the evaluation requires matching the text spans. Hence, we use an evaluation function that gives credit to partial matches between gold and predicted spans.

Let document $\boldsymbol{d}$ be represented as a sequence of characters. The i-th propagandistic text fragment is then represented as a sequence of contiguous characters $t \subseteq \boldsymbol{d}$. A document includes a set of (possibly overlapping) fragments T. Similarly, a learning algorithm produces a set S with fragments $s \subseteq \boldsymbol{d}$, predicted on $\boldsymbol{d}$. A labeling function $l(x) \in \{1, \ldots, 20\}$ associates $t \in T$, $s \in S$ with one of the techniques. An example of (gold) annotation is shown in Figure 3, where an annotation t_1 marks the span *stupid and petty* with the technique *Loaded Language*.

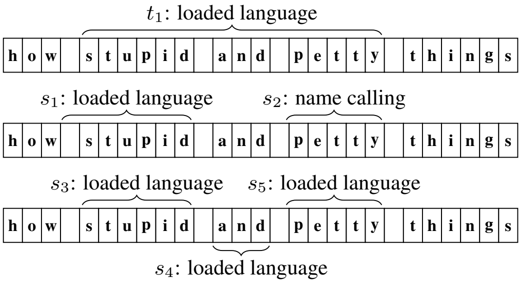

t_1: loaded language
h o w s t u p i d a n d p e t t y t h i n g s
s_1: loaded language
s_2: name calling
h o w s t u p i d a n d p e t t y t h i n g s
s_3: loaded language
s_5: loaded language
h o w s t u p i d a n d p e t t y t h i n g s
s_4: loaded language

Figure 3: Example of gold annotation (top) and the predictions of a supervised model (bottom) in a document represented as a sequence of characters.

We define the following function to handle partial overlaps of fragments with the same labels:

$$C(s,t,h) = \frac{|(s \cap t)|}{h} \delta\left(l(s), l(t)\right), \qquad (1)$$

where h is a normalizing factor and $\delta(a,b) = 1$ if $a = b$, and 0, otherwise. For example, still with reference to Figure 3, $C(t_1, s_1, |t_1|) = \frac{6}{16}$ and $C(t_1, s_2, |t_1|) = 0$.

Given Eq. (1), we now define variants of precision and recall that can account for the imbalance in the corpus:

$$P(S,T) = \frac{1}{|S|} \sum_{\substack{s \in S, \\ t \in T}} C(s,t,|s|), \qquad (2)$$

$$R(S,T) = \frac{1}{|T|} \sum_{\substack{s \in S, \\ t \in T}} C(s,t,|t|), \qquad (3)$$

We define (2) to be zero if $|S| = 0$, and Eq. (3) to be zero if $|T| = 0$. Following Potthast et al. (2010), in (2) and (3) we penalize systems predicting too many or too few instances by dividing by $|S|$ and $|T|$, respectively. Finally, we combine Eqs. (2) and (3) into an F_1-measure, the harmonic mean of precision and recall.

5.2 Task Organization

We ran the shared task in two phases:

Development Phase In the first phase, only training and development data were made available, and no gold labels were provided for the latter. The participants competed against each other to achieve the best performance on the development set. A live leaderboard was made available to keep track of all submissions.

Test Phase In the second phase, the test set was released and the participants were given just a few days to submit their final predictions.

In the *Development Phase*, the participants could make an unlimited number of submissions, and see the outcome in their private space. The best score for each team, regardless of the submission time, was also shown in a public leaderboard. As a result, not only could the participants observe the impact of various modifications in their systems, but they could also compare against the results by other participating teams. In the *Test Phase*, the participants could again submit multiple runs, but they would not get any feedback on their performance. Only the latest submission of each team was considered as official and was used for the final team ranking. The final leaderboard on the test set was made public after the end of the shared task.

In the *Development Phase*, a total of 15, 10 and 13 teams made at least one submission for ST1, ST2 and ST3, respectively. In the *Test Phase* the number of teams who made official submissions was 16, 8, and 15 for ST1, ST2, ST3, respectively.

After the competition was over, we left the submission system open for the development set, and we plan to reopen it on the test set as well. The up-to-date leaderboards can be found on the website of the competition.[2]

[2] http://propaganda.math.unipd.it/semeval2021task6/

Rank. Team	Transformers						Models						Repres.			Misc		
	BERT	RoBERTa	XLNet	ALBERT	DistilBERT	DeBERTa	LSTM	CNN	SVM	Naive Bayes	Random Forest	CRF	Embeddings	Char n-grams	PoS	Ensemble	Data augmentation	Postprocessing
1. MinD	☑	☑	☑	☑		☑							☑	☑		☑		☑
2. Alpha						☑												
3. Volta	✓	☑											☑					
5. AIMH	☑												☑					
6. LeCun		✓				✓	☑						☑				✓	
7. WVOQ															☑			
9. NLyticsFKIE		☑										☑	✓				☑	
12. YNU-HPCC	☑	☑		☑				☑					☑					
13. CSECUDSG					☑											☑		
15. NLP-IITR	✓	☑					✓		✓	✓	✓		☑				☑	

1 (Tian et al., 2021) 6 (Dia et al., 2021) 13 (Hossain et al., 2021)
2 (Feng et al., 2021) 7 (Roele, 2021) 15 (Gupta and Sharma, 2021)
3 (Gupta et al., 2021) 9 (Pritzkau, 2021)
5 (Messina et al., 2021) 12 (Zhu et al., 2021)

Table 2: **ST1:** Overview of the approaches used by the participating systems. ☑=part of the official submission; ✓=considered in internal experiments; *Repres.* stand for Representations. References to system description papers are shown below the table.

6 Participants and Results

Below, we give a general description of the systems that participated in the three subtasks and their results, with focus on those ranked among the top-3. Appendix C gives a description of every system.

6.1 Subtask 1 (Unimodal: Text)

Table 2 gives an overview of the systems that took part in Subtask 1. We can see that transformers were quite popular, and among them, most commonly used was RoBERTa, followed by BERT. Some participants used learning models such as LSTM, CNN, and CRF in their final systems, while internally, Naïve Bayes and Random Forest were also tried. In terms of representation, embeddings clearly dominated. Moreover, techniques such as ensembles, data augmentation, and post-processing were also used in some systems.

The evaluation results are shown in Table 3, which also includes two baselines: (*i*) random, and (*ii*) majority class. The latter always predicts *Loaded Language*, as it is the most frequent technique for Subtask 1 (see Table 1).

The best system **MinD** (Tian et al., 2021) used five transformers: BERT, RoBERTa, XLNet, DeBERTa, and ALBERT. It was fine-tuned on the PTC corpus (Da San Martino et al., 2020a) and then on the training data for Subtask 1.

Rank	Team	F1-Micro	F1-Macro
1	MinD	**.593**	$.290_2$
2	Alpha	.572	$.262_5$
3	Volta	.570	$.266_3$
4	mmm	.548	$\mathbf{.303_1}$
5	AIMH	.539	$.245_6$
6	LeCun	.512	$.227_8$
7	WVOQ	.511	$.227_8$
8	TeamUNCC	.510	$.236_7$
9	NLyticsFKIE	.498	$.140_{13}$
10	TeiAS	.497	$.214_{10}$
11	DAJUST	.497	$.187_{11}$
12	YNUHPCC	.493	$.263_4$
13	CSECUDSG	.489	$.185_{12}$
14	TeamFPAI	.406	$.115_{15}$
15	NLPIITR	.379	$.126_{14}$
	Majority baseline	*.374*	*.033*
16	TriHeadAttention	.184	$.024_{18}$
	Random baseline	*.064*	*.044*

Table 3: Results for Subtask 1. The systems are ordered by the official score: *F1-micro*.

The final prediction for MinD averages the probabilities for these models, and further uses post-processing rules, e.g., each bigram appearing more than three times is flagged as a *Repetition*.

Team **Alpha** (Feng et al., 2021) was ranked second. However, they used features from images, which was not allowed (images were only allowed for Subtask 3).

Team **Volta** (Gupta et al., 2021) was third. They used a combination of transformers with the [CLS] token as an input to a two-layer feed-forward network. They further used example weighting to address class imbalance.

We should also mention team **LeCun**, which used additional corpora such as the PTC corpus (Da San Martino et al., 2020a), and augmented the training data using synonyms, random insertion/deletion, random swapping, and back-translation.

6.2 Subtask 2 (Unimodal: Text)

The approaches for this task varied from modeling it as a question answering (QA) task to performing multi-task learning. Table 4 presents a high-level summary. We can see that BERT dominated, while RoBERTa was much less popular. We further see a couple of systems using data augmentation. Unfortunately, there are too few systems with system description papers for this subtask, and thus it is hard to do a very deep analysis.

Rank. Team	Trans.		Models			Repres.				Misc	
	BERT	**RoBERTa**	**LSTM**	**CNN**	**SVM**	**ELMo**	**PoS**	**Sentiment**	**Rhetorics**	**Ensemble**	**Data augmentation**
1. Volta	✓	☑									
2. HOMADOS	☑										
3. TeamFPAI	☑									☑	☑
5. WVOQ	☑		☑		✓		☑	☑	☑		
6. CSECUDSG	☑						✓				☑
7. YNU-HPCC	☑										

1 (Gupta et al., 2021)
2 (Kaczyński and Przybyła, 2021)
3 (Xiaolong et al., 2021)
5 (Roele, 2021)
6 (Hossain et al., 2021)
7 (Zhu et al., 2021)

Table 4: **ST2:** Overview of the approaches used by the participating systems. ☑=part of the official submission; ✓=considered in internal experiments; *Trans.* is for Transformers; *Repres.* is for Representations. References to system description papers are shown below the table.

Table 5 shows the evaluation results. We report our random baseline, which is based on the random selection of spans with random lengths and a random assignment of labels.

Rank	Team	F1	Precision	Recall
1	Volta	**.482**	$.501_2$	$\mathbf{.464_1}$
2	HOMADOS	.407	$.412_3$	$.403_2$
3	TeamFPAI	.397	$\mathbf{.652_1}$	$.286_5$
4	TeamUNCC	.329	$.285_4$	$.390_3$
5	WVOQ	.268	$.243_5$	$.299_4$
6	CSECUDSG	.120	$.080_8$	$.243_6$
7	YNUHPCC	.091	$.186_6$	$.060_7$
8	TriHeadAttention	.080	$.170_7$	$.052_8$
	Random Baseline	*.010*	*.034*	*.006*

Table 5: Results for Subtask 2. The systems are ordered by the official score: *F1-micro*.

The best model by team **Volta** (Gupta et al., 2021) used various transformer models, such as BERT and RoBERTa, to predict token classes by considering the output of each token embedding. Then, they assigned classes for a given word as the union of the classes predicted for the subwords that make that word (to account for BPEs).

Team **HOMADOS** (Kaczyński and Przybyła, 2021) was second, and they used a multi-task learning (MTL) and additional datasets such as the PTC corpus from SemEval-2020 task 11 (Da San Martino et al., 2020a), and a fake news corpus (Przybyla, 2020). They used BERT, followed by several output layers that perform auxiliary tasks of propaganda detection and credibility assessment in two distinct scenarios: sequential and parallel MTL. Their final submission used the latter.

Team **TeamFPAI (Xiaolong et al., 2021)** formulated the task as a question answering problem using machine reading comprehension, thus improving over the ensemble-based approach of Liu et al. (2018). They further explored data augmentation and loss design techniques, in order to alleviate the problem of data sparseness and data imbalance.

6.3 Subtask 3 (Multimodal: Memes)

Table 6 presents an overview of the approaches used by the systems that participated in Subtask 3. This is a very rich and very interesting table. We can see that transformers were quite popular for text representation, with BERT dominating, but RoBERTa being quite popular as well. For the visual modality, the most common representations were variants of ResNet, but VGG16 and CNNs were also used. We further see a variety of representations and fusion methods, which is to be expected given the multi-modal nature of this subtask.

Rank. Team	Transformers							Models								Representations														Fusion					Misc		
	BERT	RoBERTa	XLNet	ALBERT	FastBERT	GPT-2	DeBERTa	ResNet18	ResNet50	ResNet51	VGG16	LSTM	CNN	SVM	CRF	Embeddings	ELMo	Words/Word n-grams	Char n-grams	PoS	Sentiment	Rhetorics	FR (ResNet34)	MS OCR	YouTube-8M	CLIP	BUTD	ERNIE-VIL	SemVLP	Average	Concat	Attention	MLP	Chained classifier	Ensemble	Data augmentation	Postprocessing
1. Alpha	✓	☑	✓			✓			✓							☑											✓	✓							✓		
2. MinD	☑	☑	☑	☑			☑						☑		☑	☑			☑				☑	☑					☑	☑	✓		✓		☑		☑
3. 1213Li		☑							☑																					☑		☑					
4. AIMH	☑								☑																						☑		☑				
5. Volta	✓	☑											☑			☑															☑				☑		
6. CSECUDSG	☑	☑			☑				☑							☑				✓					☑						☑		☑		☑		
8. LIIR	☑																	☑								☑								☑		☑	
10. WVOQ	☑											☑		✓						☑	☑	☑															
11. YNU-HPCC				☑				☑			☑		☑																		☑		☑				
13. NLyticsFKIE	☑											✓			☑	✓																				☑	
15. LT3-UGent	☑									☑					☑																					☑	

1 (Feng et al., 2021)
2 (Tian et al., 2021)
3 (Peiguang et al., 2021)
4 (Messina et al., 2021)
5 (Gupta et al., 2021)
6 (Hossain et al., 2021)
8 (Ghadery et al., 2021)
10 (Roele, 2021)
11 (Zhu et al., 2021)
13 (Pritzkau, 2021)
15 (Singh and Lefever, 2021)

Table 6: **ST3:** Overview of the approaches used by the participating systems. ☑=part of the official submission; ✓=considered in internal experiments. References to system description papers are shown below the table.

Table 7 shows the performance on the test set for the participating systems for Subtask 3. The two baselines shown in the table are similar to those for Subtask 1, namely a random baseline and a majority class baseline. However, this time the most frequent class baseline always predicts *Smears* (for Subtask 1, it was *Loaded Language*), as this is the most frequent technique for Subtask 3 (as can be seen in Table 1).

Team **Alpha** (Feng et al., 2021) pre-trained a transformer using text with visual features. They extracted grid features using ResNet50, and salient region features using BUTD. They further used these grid features to capture the high-level semantic information in the images. Moreover, they used salient region features to describe objects and to caption the event present in the memes. Finally, they built an ensemble of fine-tuned DeBERTA+ResNet, DeBERTA+BUTD, and ERNIE-VIL systems.

Team **MinD** (Tian et al., 2021) combined a system for Subtask 1 with (*i*) ResNet-34, a face recognition system, (*ii*) OCR-based positional embeddings for text boxes, and (*iii*) Faster R-CNN to extract region-based image features. They used late fusion to combine the textual and the visual representations. Other multimodal fusion strategies they tried were concatenation of the representation and mapping using a multi-layer perceptron.

Team **1213Li** (Peiguang et al., 2021) used RoBERTa and ResNet-50 as feature extractors for texts and images, respectively, and adopted a label embedding layer with a multi-modal attention mechanism to measure the similarity between labels with multi-modal information, and fused features for label prediction.

Rank	Team	F1-Micro	F1-Macro
1	Alpha	**.581**	**.273**$_1$
2	MinD	.566	.244$_3$
3	1213Li	.549	.228$_5$
4	AIMH	.540	.207$_6$
5	Volta	.521	.189$_8$
6	CSECUDSG	.513	.121$_{11}$
7	aircasMM	.511	.200$_7$
8	LIIR	.498	.188$_9$
9	CAU731NLP	.481	.084$_{14}$
10	WVOQ	.478	.240$_4$
11	YNUHPCC	.446	.096$_{13}$
12	TriHeadAttention	.442	.062$_{15}$
13	NLyticsFKIE	.423	.118$_{12}$
	Majority baseline	*.354*	*.036*
14	LT3UGent	.332	.264$_2$
15	TeamUNCC	.224	.124$_{10}$
	Random baseline	*.071*	*.052*

Table 7: Results for Subtask 3. The systems are ordered by the official score: *F1-micro*.

7 Conclusion and Future Work

We presented *SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images*. It was a successful task: a total of 71 teams registered to participate, 22 teams eventually made an official submission on the test set, and 15 teams also submitted a task description paper.

In future work, we plan to increase the data size and to add more propaganda techniques. We further plan to cover several different languages.

Acknowledgments

This research part of the Tanbih mega-project,[3] which is developed at the Qatar Computing Research Institute, HBKU, and aims to limit the impact of "fake news," propaganda, and media bias by making users aware of what they are reading.

Ethics and Broader Impact

User Privacy Our dataset only includes memes and it contains no user information.

Biases Any biases in the dataset are unintentional, and we do not intend to do harm to any group or individual. Note that annotating propaganda techniques can be subjective, and thus it is inevitable that there would be biases in our gold-labeled data or in the label distribution. We address these concerns by collecting examples from a variety of users and groups, and also by following a well-defined schema, which has clear definitions and on which we achieved high inter-annotator agreement.

Moreover, we had a diverse annotation team, which included six members, both female and male, all fluent in English, with qualifications ranging from undergrad to MSc and PhD degrees, including experienced NLP researchers, and covering multiple nationalities. This helped to ensure the quality. No incentives were provided to the annotators.

Misuse Potential We ask researchers to be aware that our dataset can be maliciously used to unfairly moderate memes based on biases that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required in order to ensure this does not occur.

[3] http://tanbih.qcri.org/

References

Shamsiah Abd Kadir, Anitawati Lokman, and T. Tsuchiya. 2016. Emotion and techniques of propaganda in YouTube videos. *Indian Journal of Science and Technology*, Vol (9):1–8.

Shamsiah Abd Kadir and Ahmad Sauffiyan. 2014. A content analysis of propaganda in Harakah newspaper. *Journal of Media and Information Warfare*, 5:73–116.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Ivan Habernal et al. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'18, pages 386–396, New Orleans, LA, USA.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 1: Check-worthiness. In *CLEF 2018 Working Notes*, Avignon, France.

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the CLEF-2019 CheckThat! lab on automatic identification and verification of claims. Task 1: Check-worthiness. In *Working Notes of CLEF 2019*, Lugano, Switzerland.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality. In *CLEF 2018 Working Notes*, Avignon, France.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of CheckThat! 2020 — automatic identification and verification of claims in social media. In *Proceedings of the 11th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, CLEF '2020.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In *Proceedings of the European Conference on Information Retrieval*, ECIR '19, pages 499–507, Lisbon, Portugal.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF '19, pages 162–170, Hong Kong, China.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, SemEval '20, pages 1377–1414, Barcelona, Spain.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. 2020c. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 287–293.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 5636–5646, Hong Kong, China.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 60–67, Vancouver, Canada.

Abujaber Dia, Qarqaz Ahmed, and Abdullah Malak A. 2021. LeCun at SemEval-2021 Task 6: Detecting persuasion techniques in text using ensembled pretrained transformers and data augmentation. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019a. CheckThat! at CLEF 2019: Automatic identification and verification of claims. In *Advances in Information Retrieval*, ECIR '19, pages 309–315, Lugano, Switzerland.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019b. Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, LNCS, pages 301–321.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at SemEval-2021 Tasks 6: Transformer based propaganda classification. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Erfan Ghadery, Damien Sileo, and Marie-Francine Moens. 2021. LIIR at SemEval 2021 Task 6: Detection of persuasion techniques in texts and images using CLIP features. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Maria Glenski, E. Ayton, J. Mendoza, and Svitlana Volkova. 2019. Multilingual multimodal digital deception detection and disinformation spread across social platforms. *ArXiv*, abs/1909.05838.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 845–854, Minneapolis, Minnesota, USA.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53(4).

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 Task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Vansh Gupta and Raksha Sharma. 2021. NLPIITR at SemEval-2021 Task 6: detection of persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 7–12, Copenhagen, Denmark.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC'18, Miyazaki, Japan.

Maram Hasanain, Fatima Haouari, Reem Suwaileh, Zien Sheikh Ali, Bayan Hamdan, Tamer Elsayed, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In *Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum*, CLEF '2020.

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Overview of the CLEF-2019 CheckThat! Lab on Automatic Identification and Verification of Claims. Task 2: Evidence and Factuality. In *CLEF 2019 Working Notes. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, Lugano, Switzerland. CEUR-WS.org.

Tashin Hossain, Jannatun Naim, Fareen Tasneem, Radiathun Tasnia, and Abu Nowshed Chy. 2021. CSECUDSG at SemEval-2021 Task 6: Orchestrating multimodal neural architectures for identifying persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Konrad Kaczyński and Piotr Przybyła. 2021. HOMADOS at SemEval-2021 Task 6: Multi-task learning for propaganda detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. In *Proceedings of the NeurIPS 2019 Workshop on Visually Grounded Interaction and Language*, ViGIL@NeurIPS '19.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NeurIPS '20.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Jiahua Liu, Wan Wei, Maosong Sun, Hao Chen, Yantao Du, and Dekang Lin. 2018. A multi-answer multitask framework for real-world machine reading comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2109–2118, Brussels, Belgium.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. FastBERT: a self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 6035–6044, Online.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems*, NeurIPS '19, pages 13–23, Vancouver, Canada.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832.

Nicola Messina, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. 2021. AIMH at SemEval-2021 Task 6: multimodal classification using an ensemble of transformer models. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 860–869, Minneapolis, Minnesota, USA.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations, Workshop Track*, ICLR '13, Scottsdale, Arizona, USA.

Clyde R. Miller. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the

CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *Proceedings of the International Conference of CLEF*, CLEF '18, pages 372–387, Avignon, France.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval*, ECIR '21, pages 639–649.

Li Peiguang, Li Xuan, and Sun Xian. 2021. 1213Li at SemEval-2021 Task 6: detection of propaganda with multi-modal attention and pre-trained models. In *Proceedings of the Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 997–1005, Beijing, China.

Albert Pritzkau. 2021. NLyticsFKIE at SemEval-2021 Task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Piotr Przybyla. 2020. Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

Cees Roele. 2021. WVOQ at SemEval-2021 Task 6: BART for span detection and classification. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Hyunjin Seo. 2014. Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli–Hamas conflict. *Visual Communication Quarterly*, 21(3):150–161.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, and Preslav Nakov. 2021. Findings of the NLP4IF-2021 shared task on fighting the COVID-19 infodemic. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeño, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In *Working Notes of CLEF 2020—Conference and Labs of the Evaluation Forum*, CLEF '2020.

Anup Shah. 2005. War, propaganda and the media.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Pranaydeep Singh and Els Lefever. 2021. LT3 at SemEval-2021 Task 6: Using multi-modal compact bilinear pooling to combine visual and textual understanding in memes. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 809–819, New Orleans, Louisiana.

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 Task 6: Propaganda detection using transfer learning and multimodal fusion. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In *Proceedings of the Australian Security and Intelligence Conference*, ASIC '15, pages 58–65, Perth, Australia.

Svitlana Volkova, Ellyn Ayton, Dustin L. Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the International Conference on Web and Social Media*, ICWSM '19, pages 659–662, Munich, Germany.

Anthony Weston. 2008. *A rulebook for arguments*. Hackett Publishing.

Hou Xiaolong, Ren Junsong, Rao Gang, Jiang Lianxin, Ruan Zhihao, Yang Mo, and Shen Jianping. 2021. TeamFPAI at SemEval-2020 Task 6: BERT-MRC for propaganda techniques detection. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Xingyu Zhu, Jin Wang, and Xuejie Zhang. 2021. YNU-HPCC at SemEval-2021 Task 6: Combining ALBERT and Text-CNN for persuasion detection in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand.

Appendix

A Data Collection and Annotation

A.1 Data Collection

To collect the data for the dataset, we used Facebook, as it has many public groups with a large number of users, who intentionally or unintentionally share a large number of memes. We used our own private Facebook accounts to crawl the public posts from users and groups. To make sure the resulting feed had a sufficient number of memes, we initially selected some public groups focusing on topics such as politics, vaccines, COVID-19, and gender equality. Then, using the links between groups, we expanded our initial group pool to a total of 26 public groups. We went through each group, and we collected memes from old posts, dating up to three months before the newest post in the group. Out of the 26 groups, 23 were about politics, US and Canadian: left, right, centered, anti-government, and gun control. The other 3 groups were on general topics such as health, COVID-19, pro-vaccines, anti-vaccines, and gender equality. Even though the number of political groups was larger (i.e., 23), the other 3 general groups had a higher number of users and a substantial amount of memes.

A.2 Annotation Process

We annotated the memes using the 22 persuasion techniques from Section 3 in a multi-label setup. Our annotation focused (*i*) on the text only, using 20 techniques, and (*ii*) on the entire meme (text + image), using all 22 techniques.

We could not annotate the visual modality as an independent task because memes have the text as part of the image. Moreover, in many cases, the message in the meme requires both modalities. For example, in Figure 28, the image by itself does not contain any persuasion technique, but together with the text, we can see *Smears* and *Reductio at Hitlerum*.

The annotation team included six members, both female and male, all fluent in English, with qualifications ranging from undergrad to MSc and PhD degrees, including experienced NLP researchers, and covering multiple nationalities. This helped to ensure the quality of the annotation, and our focus was really on having very high-quality annotation. No incentives were given to the annotators.

We used PyBossa[4] as an annotation platform, as it provides the functionality to create a custom annotation interface that we found to be a good fit for our needs in each phase of the annotation process. Figure 4 shows examples of the annotation interface for the five different phases of annotation, which we describe in detail below.

Phase 1: Filtering and Text Editing The first phase of the annotation process is about selecting the memes for our task, followed by extracting and editing the textual contents of each meme. After we collected the memes, we observed that we needed to remove some of them as they did not fit our definition: "*photograph style image with a short text on top of it*." Thus, we asked the annotators to exclude images with the characteristics listed below. During this phase, we filtered out a total of 111 memes.

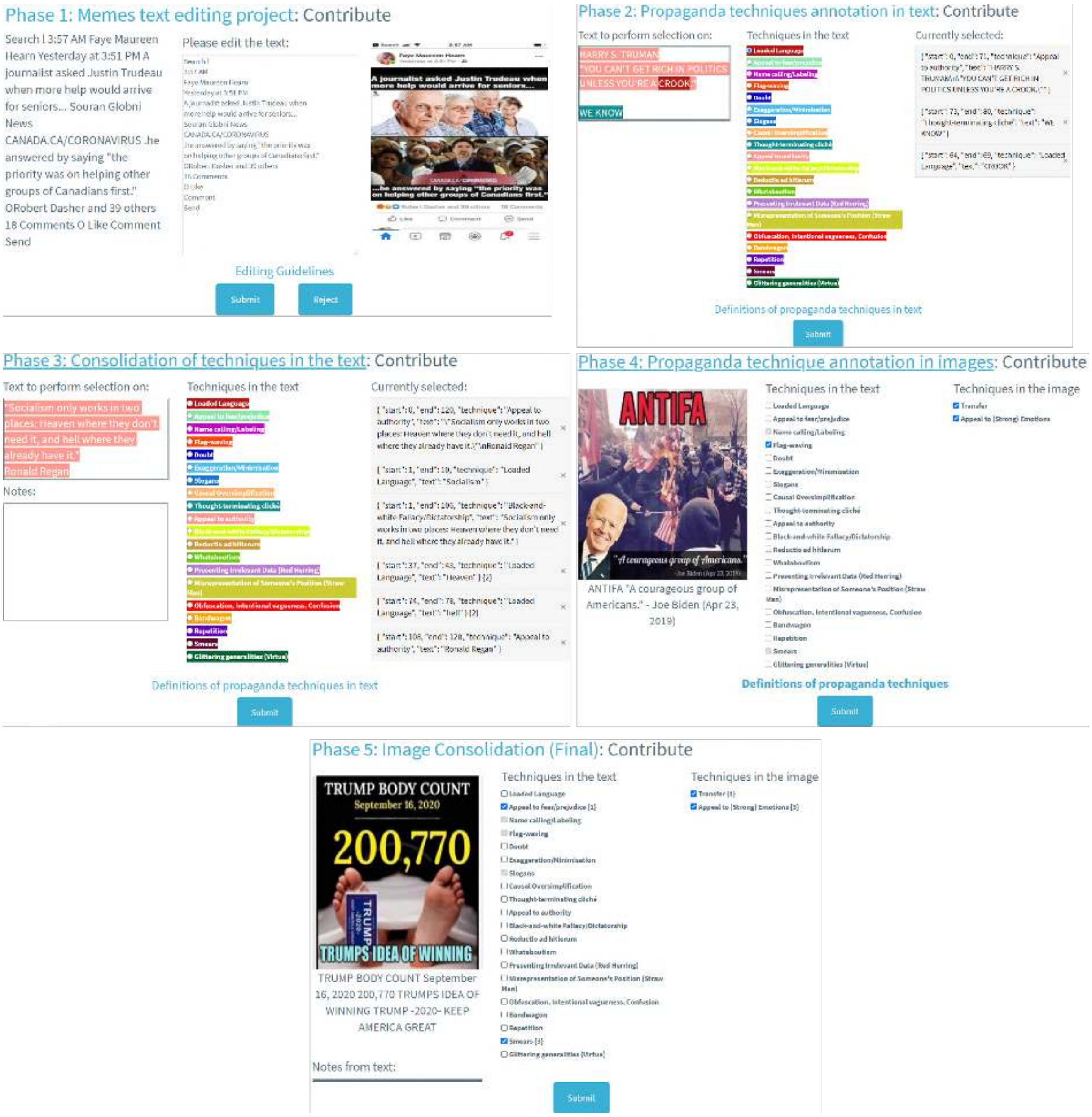
- Images with diagrams/graphs/tables (see Figure 5a).
- Cartoons. (see Figure 5b)
- Memes for which no multi-modal analysis is possible: e.g., only text, only image, etc. (see Figure 5c)

Next, we used the Google Vision API[5] to extract the text from the memes. As the resulting text sometimes contains errors, manual checking was needed to correct it. Thus, we defined several text editing rules, and we asked the annotators to apply them on the memes that passed the filtering rules above.

1. When the meme is a screenshot of a social network account, e.g., WhatsApp, the user name and login can be removed as well as all "Like", "Comment', "Share".
2. Remove the text related to logos that are not part of the main text.
3. Remove all text related to figures and tables.
4. Remove all text that is partially hidden by an image, so that the sentence is almost impossible to read.
5. Remove all text that is not from the meme, but on banners carried on by demonstrators, street advertisements, etc.

[4] https://pybossa.com

[5] http://cloud.google.com/vision



Phase 1: Memes text editing project: Contribute

Search l 3:57 AM Faye Maureen Hearn Yesterday at 3:51 PM A journalist asked Justin Trudeau when more help would arrive for seniors... Souran Globni News CANADA.CA/CORONAVIRUS .he answered by saying "the priority was on helping other groups of Canadians first." ORobert Dasher and 39 others 18 Comments O Like Comment Send

Please edit the text:

Search l
3:57 AM
Faye Maureen Hearn
Yesterday at 3:51 PM
A journalist asked Justin Trudeau when
more help would arrive for seniors...
Souran Globni News
CANADA.CA/CORONAVIRUS
.he answered by saying "the priority was
on helping other groups of Canadians first."
ORobert Dasher and 39 others
18 Comments
0 Like
Comment
Send

Faye Maureen Hearn
A journalist asked Justin Trudeau when more help would arrive for seniors...
...he answered by saying "the priority was on helping other groups of Canadians first."
Like
Comment
Send

Editing Guidelines

Submit Reject

Phase 2: Propaganda techniques annotation in text: Contribute

Text to perform selection on:

HARRY S. TRUMAN
"YOU CAN'T GET RICH IN POLITICS
UNLESS YOU'RE A CROOK."

WE KNOW

Techniques in the text

● Loaded Language
● Appeal to fear/prejudice
● Name calling/Labeling
● Flag-waving
● Doubt
● Exaggeration/Minimisation
● Slogans
● Causal Oversimplification
● Thought-terminating cliché
● Appeal to authority
● Black-and-white Fallacy/Dictatorship
● Reductio ad hitlerum
● Whataboutism
● Presenting Irrelevant Data (Red Herring)
● Misrepresentation of Someone's Position (Straw Man)
● Obfuscation, Intentional vagueness, Confusion
● Bandwagon
● Repetition
● Smears
● Glittering generalities (Virtue)

Currently selected:

{"start": 0, "end": 71, "technique": "Appeal to authority", "text": "HARRY S. TRUMAN\n"YOU CAN'T GET RICH IN POLITICS UNLESS YOU'RE A CROOK.\""} ×

{"start": 73, "end": 80, "technique": "Thought-terminating cliché", "text": "WE KNOW"} ×

{"start": 64, "end": 69, "technique": "Loaded Language", "text": "CROOK"} ×

Definitions of propaganda techniques in text

Submit

Phase 3: Consolidation of techniques in the text: Contribute

Text to perform selection on:

"Socialism only works in two places: Heaven where they don't need it, and hell where they already have it."
Ronald Regan

Notes:

Techniques in the text

● Loaded Language
● Appeal to fear/prejudice
● Name calling/Labeling
● Flag-waving
● Doubt
● Exaggeration/Minimisation
● Slogans
● Causal Oversimplification
● Thought-terminating cliché
● Appeal to authority
● Black-and-white Fallacy/Dictatorship
● Reductio ad hitlerum
● Whataboutism
● Presenting Irrelevant Data (Red Herring)
● Misrepresentation of Someone's Position (Straw Man)
● Obfuscation, Intentional vagueness, Confusion
● Bandwagon
● Repetition
● Smears
● Glittering generalities (Virtue)

Currently selected:

{"start": 0, "end": 120, "technique": "Appeal to authority", "text": "\"Socialism only works in two places: Heaven where they don't need it, and hell where they already have it.\"\nRonald Regan"} ×

{"start": 1, "end": 10, "technique": "Loaded Language", "text": "Socialism"} ×

{"start": 1, "end": 106, "technique": "Black-and-white Fallacy/Dictatorship", "text": "Socialism only works in two places: Heaven where they don't need it, and hell where they already have it."} ×

{"start": 37, "end": 43, "technique": "Loaded Language", "text": "Heaven"} (2) ×

{"start": 74, "end": 78, "technique": "Loaded Language", "text": "hell"} (2) ×

{"start": 108, "end": 120, "technique": "Appeal to authority", "text": "Ronald Regan"} ×

Definitions of propaganda techniques in text

Submit

Phase 4: Propaganda technique annotation in images: Contribute

ANTIFA "A courageous group of Americans." - Joe Biden (Apr 23, 2019)

Techniques in the text

□ Loaded Language
□ Appeal to fear/prejudice
□ Name calling/Labeling
☑ Flag-waving
□ Doubt
□ Exaggeration/Minimisation
□ Slogans
□ Causal Oversimplification
□ Thought-terminating cliché
□ Appeal to authority
□ Black-and-white Fallacy/Dictatorship
□ Reductio ad hitlerum
□ Whataboutism
□ Presenting Irrelevant Data (Red Herring)
□ Misrepresentation of Someone's Position (Straw Man)
□ Obfuscation, Intentional vagueness, Confusion
□ Bandwagon
□ Repetition
□ Smears
□ Glittering generalities (Virtue)

Techniques in the image

☑ Transfer
☑ Appeal to (Strong) Emotions

Definitions of propaganda techniques

Submit

Phase 5: Image Consolidation (Final): Contribute

TRUMP BODY COUNT September 16, 2020 200,770 TRUMPS IDEA OF WINNING TRUMP -2020- KEEP AMERICA GREAT

Notes from text:

Techniques in the text

□ Loaded Language
☑ Appeal to fear/prejudice (1)
□ Name calling/Labeling
□ Flag-waving
□ Doubt
□ Exaggeration/Minimisation
□ Slogans
□ Causal Oversimplification
□ Thought-terminating cliché
□ Appeal to authority
□ Black-and-white Fallacy/Dictatorship
□ Reductio ad hitlerum
□ Whataboutism
□ Presenting Irrelevant Data (Red Herring)
□ Misrepresentation of Someone's Position (Straw Man)
□ Obfuscation, Intentional vagueness, Confusion
□ Bandwagon
□ Repetition
☑ Smears (3)
□ Glittering generalities (Virtue)

Techniques in the image

☑ Transfer (1)
☑ Appeal to (Strong) Emotions (3)

Submit

Figure 4: Examples of the annotation interface for different phases.

6. Remove the author of the meme if it is signed.

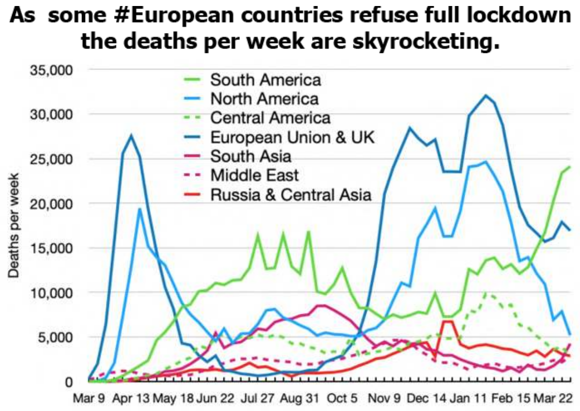
7. If the text is in columns, first put all text from the first column, then all text from the next column, etc.

8. Rearrange the text, so that there is one sentence per line, whenever possible.

9. If there are separate blocks of text in different locations of the image, separate them by a blank line. However, if it is evident that the text blocks are part of a single sentence, keep them together.

Phase 2: Text Annotation The annotations for phase 2 are targeted at Subtasks 1 and 2. Given the list of propaganda techniques for text only annotation, as discussed in Section A.4 (i.e., techniques 1-20), and the textual content of the target meme, the annotators were asked to identify which techniques appear in the text, and also to annotate the span of each instance of a technique use. In this phase, there were three annotators per example.

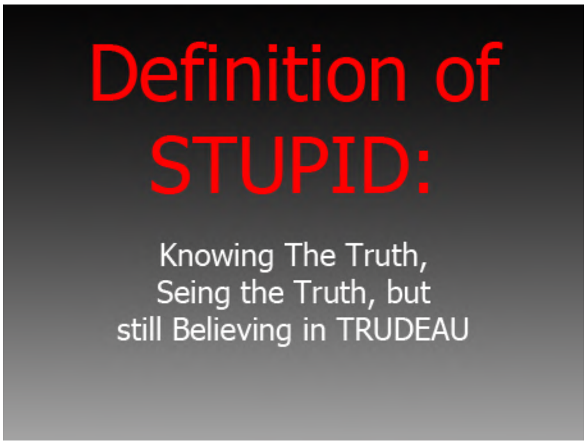
Phase 3: Text Consolidation Phase 3 is the consolidation step for the annotations from phase 2. The three annotators met with the rest of the team, who acted as consolidators, and discussed each annotation, so that a consensus could be reached.



(a) Example of a meme with a **graph** ***Source(s):*** Image ; License

CONAIR
TRUMP
NO MUSLIMS ALLOWED
BORDER CROSSING
TAJ MA WALL

(b) Example of a **cartoon** meme; ***Source(s):*** Image ; License .



Definition of STUPID:

Knowing The Truth,
Seing the Truth, but
still Believing in TRUDEAU

(c) Example of a meme with **only text modality**; License .

Figure 5: Examples of memes we filtered out.

We made sure to consider different interpretations and to anotate techniques corresponding to the most likely one. While this phase was devoted to checking the annotations from phase 2, when a novel instance of a technique was found, it could be added; conversely, an instance of a technique with perfect agreement from phase 2 could also be dropped. Phase 3 was essential for ensuring quality, and it served as an additional training opportunity for the entire team, which was very useful.

Phase 4: Multimodal Annotation In this phase, the goal is to identify which of the 22 techniques, discussed in Section A.4, appear in the meme: in the text and in the visual content. Note that some of the techniques occurring in the text might be identified only in this phase because the image provides the necessary context. Thus, we presented the meme with the consolidated propaganda labels from phase 3. We intentionally provided the consolidated text labels to the annotators in order to ensure that they focus their attention on identifying propaganda techniques that require both modalities rather than repeating what was already labeled in the earlier phases. In this phase, there were three annotators per example.

Phase 5: Multimodal Consolidation. In phase 5, we consolidated the annotations from phase 4 in a discussion of the entire team of six annotators (just as we did for phase 3).

A.3 Annotation Agreement

We assessed the quality for the individual annotators from phases 2 and 4 (i.e., when combining the annotations for the meme's text and for the entire meme) to the final consolidated labels at phase 5. Since our annotation is multi-label, we computed Krippendorff's α (Artstein and Poesio, 2008). The results are shown in Table 8, and the numbers indicate moderate to substantial agreement (Landis and Koch, 1977).

Agreement Pair	**Krippendorff's α**
Annotator 1 vs. Consolidated	0.83
Annotator 2 vs. Consolidated	0.91
Annotator 3 vs. Consolidated	0.56
Average	**0.77**

Table 8: Inter-annotator agreement in terms of Krippendorff's α between each of the annotators and the consolidated annotation.

A.4 Propaganda Techniques: Definitions

Below, we present the definitions of our 22 propaganda techniques, together with examples: both textual, and memes. Note that, for copyright reasons, we show our own recreated versions of actual memes from our dataset, where, for each meme, we indicate the image(s) we used and the corresponding license terms (as hyperlinks in the image caption).

1. Loaded Language: Using specific words and phrases with strong emotional implications (i.e., either positive or negative) to influence an audience.

An example meme is shown in Figure 6, which contains four instances of this persuasion technique in its text: *killed thousands of innocents*, *retaliate*, *kill*, and *warmonger*.

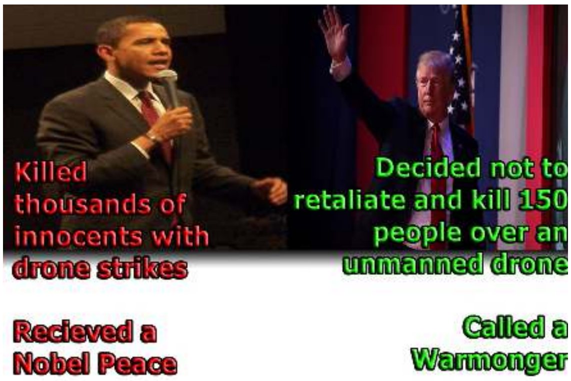

Killed thousands of innocents with drone strikes
Decided not to retaliate and kill 150 people over an unmanned drone
Recieved a Nobel Peace
Called a Warmonger

Figure 6: Example for **Loaded Language**; ***Source(s):*** Image 1, Image 2; License 1, License 2

2. Name Calling or Labeling: Labeling the object of the propaganda as either something the target audience fears, hates, finds undesirable, or loves, praises.

Figure 7 shows three instances of this technique: *the two biggest threats to America*, *the worst senate leader ever*, and *the most corrupt President ever*. Figure 6 also contains an instance: *warmonger*.

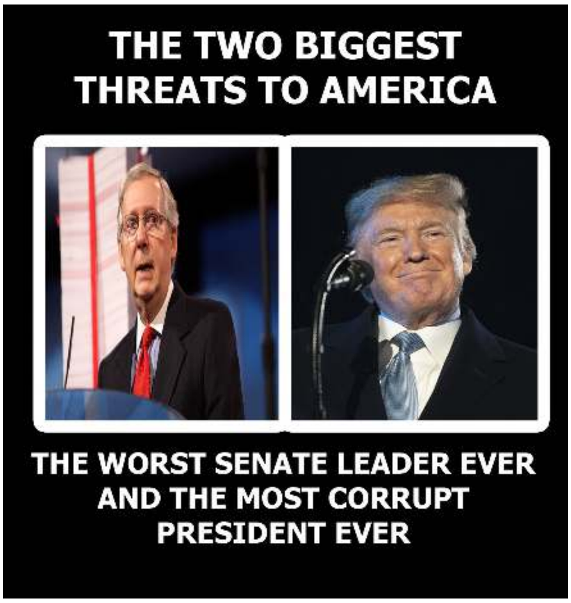

THE TWO BIGGEST THREATS TO AMERICA
THE WORST SENATE LEADER EVER AND THE MOST CORRUPT PRESIDENT EVER

Figure 7: Example for **Name Calling**; ***Source(s):*** Image 1, Image 2; License 1, License 2

3. Doubt: Questioning the credibility of someone or something.

An example is shown in Figure 8, where the entire text in the meme represents a span for this technique, while the image is just for illustration.

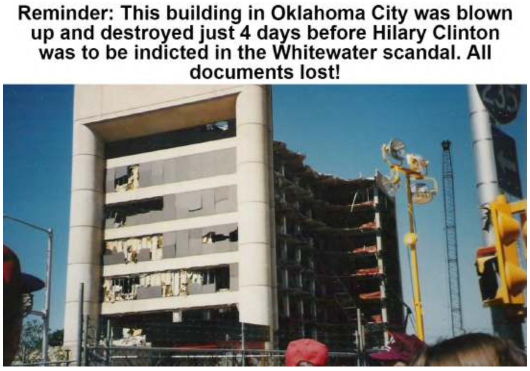

Reminder: This building in Oklahoma City was blown up and destroyed just 4 days before Hilary Clinton was to be indicted in the Whitewater scandal. All documents lost!

Figure 8: Example for **Doubt**; ***Source(s):*** Image ; License

4. Exaggeration or Minimisation: Representing something in an excessive manner, making it larger, better, worse (e.g., *the best of the best*); or making it seem less important or smaller than it really is (e.g., saying that an insult was just a joke).

An example is shown in Figure 9, where the entire meme conveys an exaggeration. Moreover, all three *Name Calling* instances in Figure 7 are also examples of *Exaggeration*.

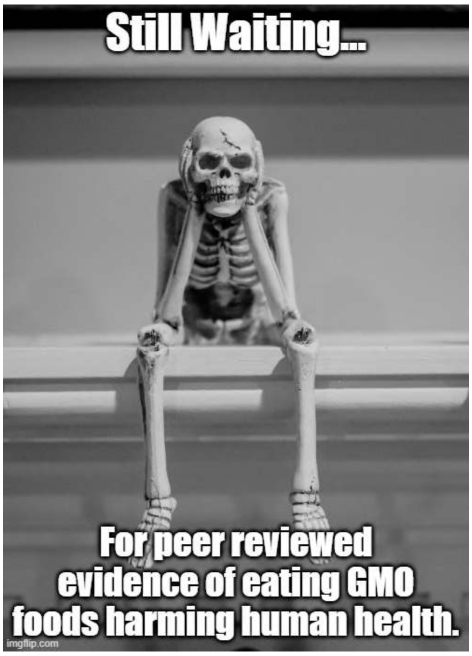

Still Waiting...
For peer reviewed evidence of eating GMO foods harming human health.
imgflip.com

Figure 9: Example for **Exaggeration**; ***Source(s):*** Image ; License

5. Appeal to Fear/Prejudice: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases, the support is built based on preconceived judgments.

An example is shown in Figure 10, where both the text and the image instill fear.

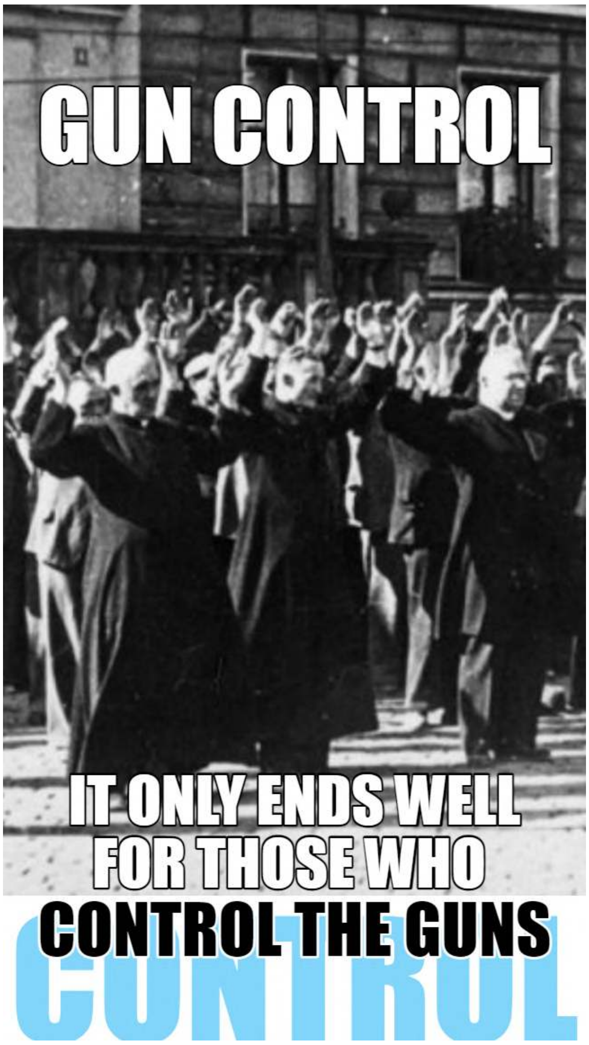


Figure 10: Example for **Appeal to Fear**; ***Source(s):*** Image ; License

6. Slogans: A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

An example is shown in Figure 11, which contains a slogan in its textual content: "*Vaccines. It isn't always about you.*"

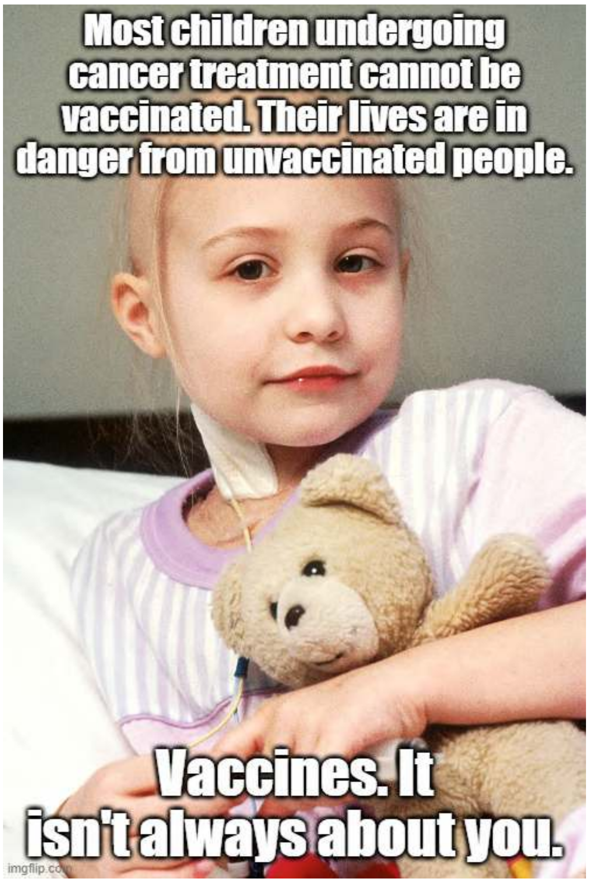

Most children undergoing cancer treatment cannot be vaccinated. Their lives are in danger from unvaccinated people.
Vaccines. It isn't always about you.
imgflip.com

Figure 11: Example for **Slogan**; ***Source(s):*** Image ; License

7. Whataboutism: A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

An example meme is shown in Figure 12, where the entire text represents a span for this technique, while the image is just for illustration.

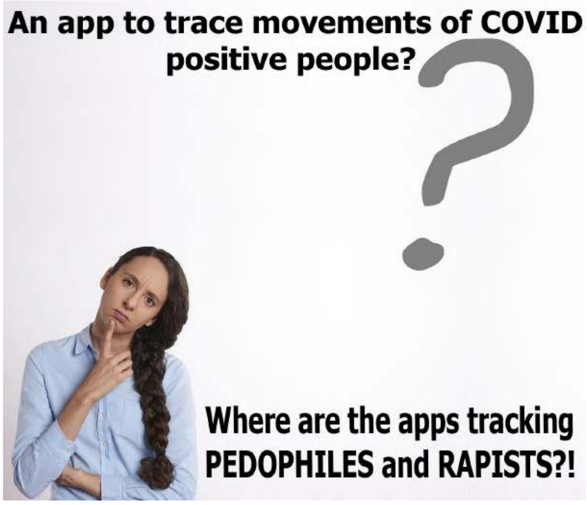

An app to trace movements of COVID positive people?
Where are the apps tracking PEDOPHILES and RAPISTS?!

Figure 12: Example for **Whataboutism**; ***Source(s):*** Image ; License

8. Flag-Waving: Playing on strong national feeling (or to any group such as race, gender, political preference) to justify or promote an action or idea.

An example is shown in Figure 13, with the technique expressed in the text and the image.

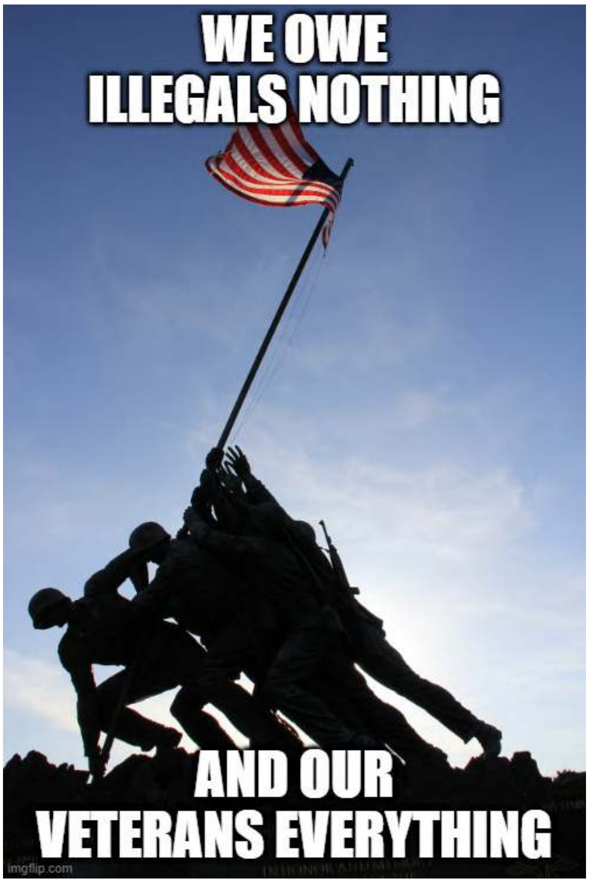


Figure 13: Example for **Flag-Waving**; ***Source(s):*** Image ; License

9. Misrepresentation of Someone's Position (Straw Man): An opponent's proposition is substituted with a similar one, which is then refuted in place of the original proposition.

An example meme is shown in Figure 14, which contains an instance of this technique in its text: here, the entire text in the meme represents a span for this technique, while the image is irrelevant for that technique (however, it is relevant for other techniques such as *Smears*).

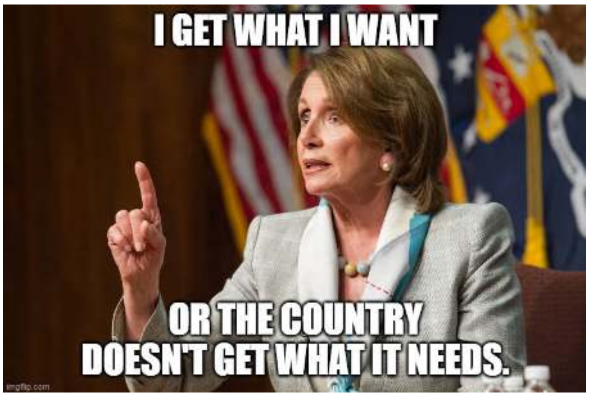


Figure 14: Example for **Misrepresentation of Someone's Position (Straw Man)**; ***Source(s):*** Image ; License

10. Causal Oversimplification: Assuming a single cause or reason when there are actually multiple causes for an issue. It includes transferring blame to one person or group of people without investigating the complexities of the issue.

An example meme is shown in Figure 15, which contains an instance of this technique in its text: "*You can't get rich in politics unless you are a crook.*" This statement says that if somebody got rich in politics, the only reason for this happening should be that this person is a crook, while in reality there are typically multiple causes. The image is irrelevant for that technique (however, it is relevant for other techniques such as *Smears*).

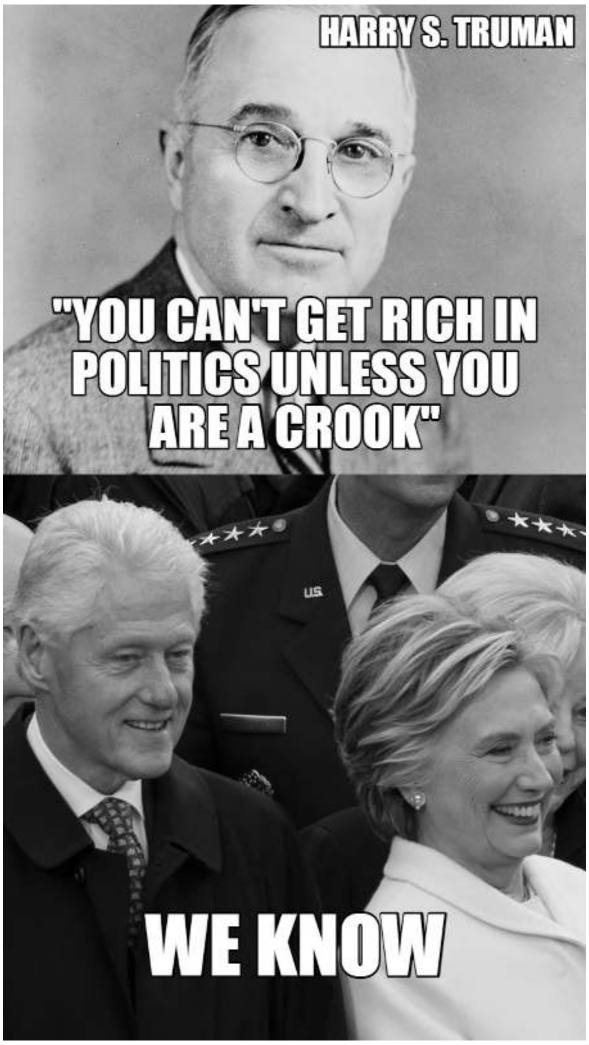


Figure 15: Example for **Causal Oversimplification**; ***Source(s):*** Image 1, Image 2; License 1, License 2

11. Appeal to Authority: Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered. We consider the special case in which the reference is not an authority or an expert in this technique, although it is referred to as *Testimonial* in literature.

An example meme is shown in Figure 16, which contains a quote by the 3rd President of the United States.

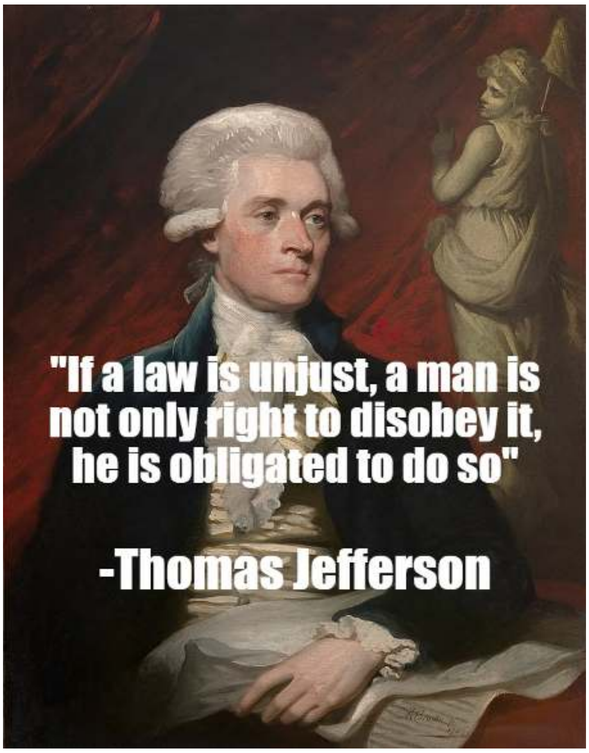

"If a law is unjust, a man is not only right to disobey it, he is obligated to do so"
-Thomas Jefferson

Figure 16: Example for **Appeal to Authority**; ***Source(s):*** Image ; License

12. Thought-Terminating Cliché: Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought.

Figure 17 shows a meme with an instance of this technique in its text: "*PERIOD.*"

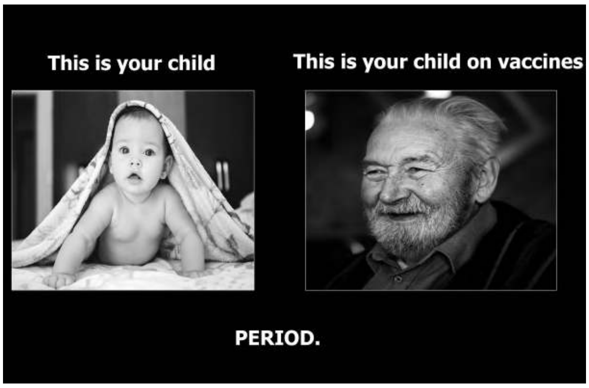


Figure 17: Example for **Thought-Terminating Cliché**; ***Source(s):*** Image 1, Image 2; License 1, License 2

13. Black-and-White Fallacy: Presenting two alternative options as the only possibilities, when in fact more possibilities exist. We also include dictatorship, where one tells the audience exactly what actions to take, eliminating any other choices.

An example of this technique is shown in Figure 18, which offers only two choices.

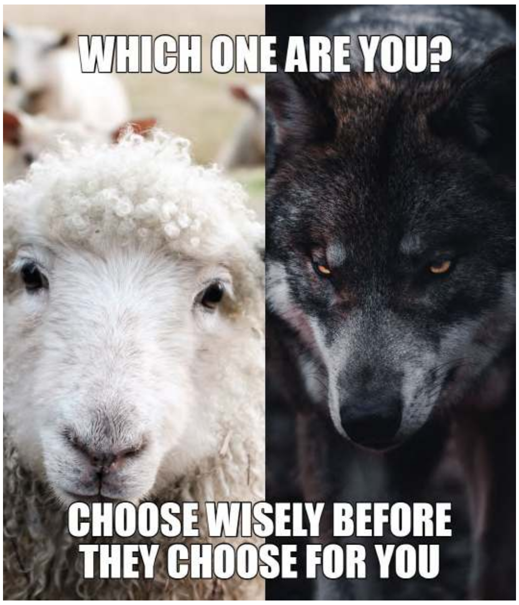


Figure 18: Example for **Black-and-White Fallacy**; ***Source(s):*** Image 1, Image 2; License 1, License 2

14. Reductio ad Hitlerum: Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated or in contempt by the target audience. It can refer to any person or concept with a negative connotation.

Figure 19 shows a meme trying to discredit the idea of being anti-union by saying that so is Donald Trump, who in turn is shown in bad light.

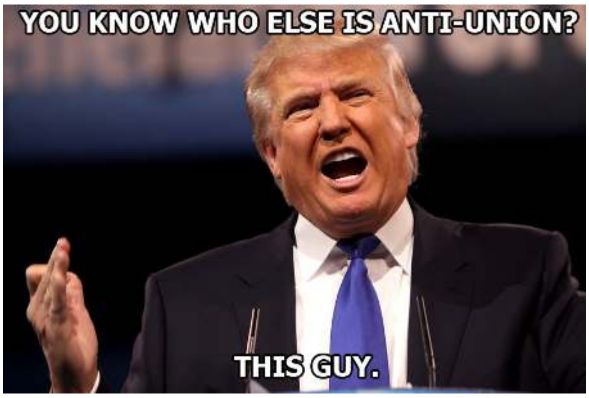


Figure 19: Example for **Reduction ad Hitlerum**; ***Source(s):*** Image , License

15. Repetition: Repeating the same message, so that the audience eventually accepts it.

An example is shown in Figure 20, where the repetition has a clear rhetorical function.

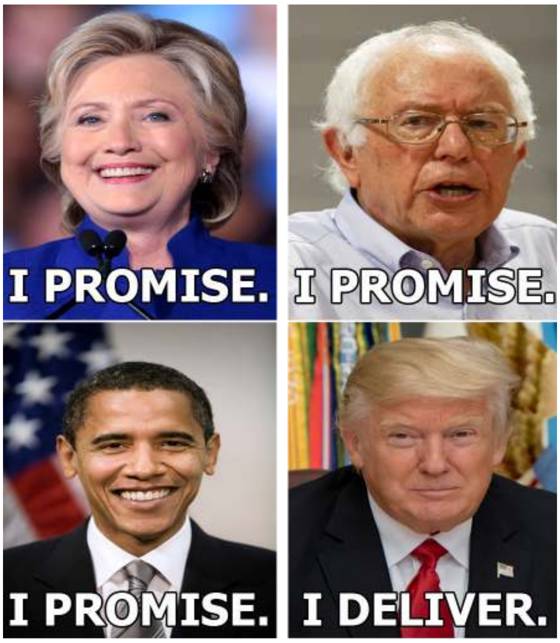


Figure 20: Example for **Repetition**; ***Source(s):*** Image 1, Image 2, Image 3, Image 4; License 1, License 2, License 3, License 4

16. Obfuscation, Intentional Vagueness, Confusion: Using words that are deliberately unclear, so that the audience may have their own interpretations.

Figure 21, shows an example, where the entire quote by Joe Biden is a span of this technique, as it is unclear what exactly is meant here.

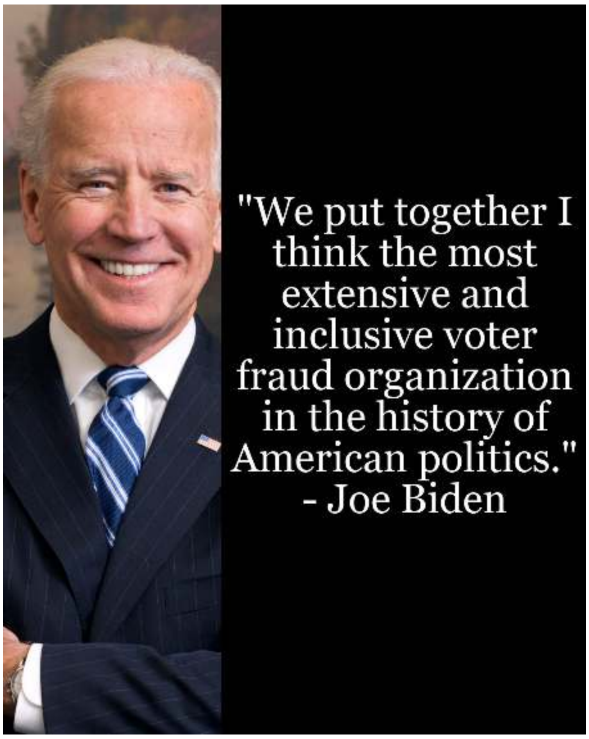

"We put together I think the most extensive and inclusive voter fraud organization in the history of American politics." - Joe Biden

Figure 21: Example for **Obfuscation, Intentional vagueness, Confusion**; ***Source(s):*** Image ; License

17. Presenting Irrelevant Data (Red Herring): Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.

An example meme is shown in Figure 22, which contains an instance of this technique in its text. We can see that there is no real connection between the two sentences. Here, the entire text represents a span for this technique, while the image is for reinforcement.

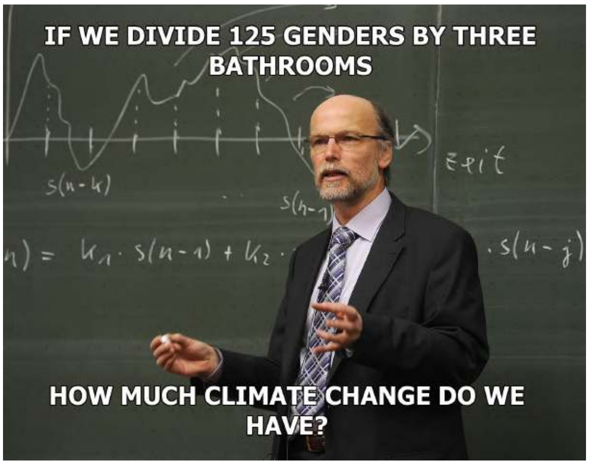

IF WE DIVIDE 125 GENDERS BY THREE BATHROOMS
HOW MUCH CLIMATE CHANGE DO WE HAVE?

Figure 22: Example for **Presenting Irrelevant Data (Red Herring)**; ***Source(s):*** Image ; License

18. Bandwagon: Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action."

Figure 23 shows an example that covers the entire text; the image less relevant.

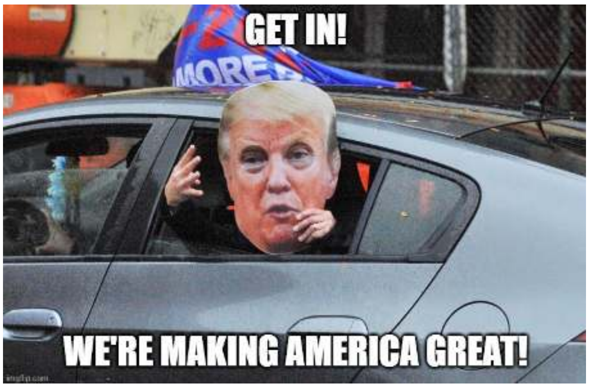


Figure 23: Example for **Bandwagon**; ***Source(s):*** Image ; License

19. Smears: A smear is an effort to damage or to call into question someone's reputation, by propounding negative propaganda. It can be applied to individuals or groups.

An example meme is shown in Figure 24, where the combination of the image and the text conveys the idea that Biden is unpopular.

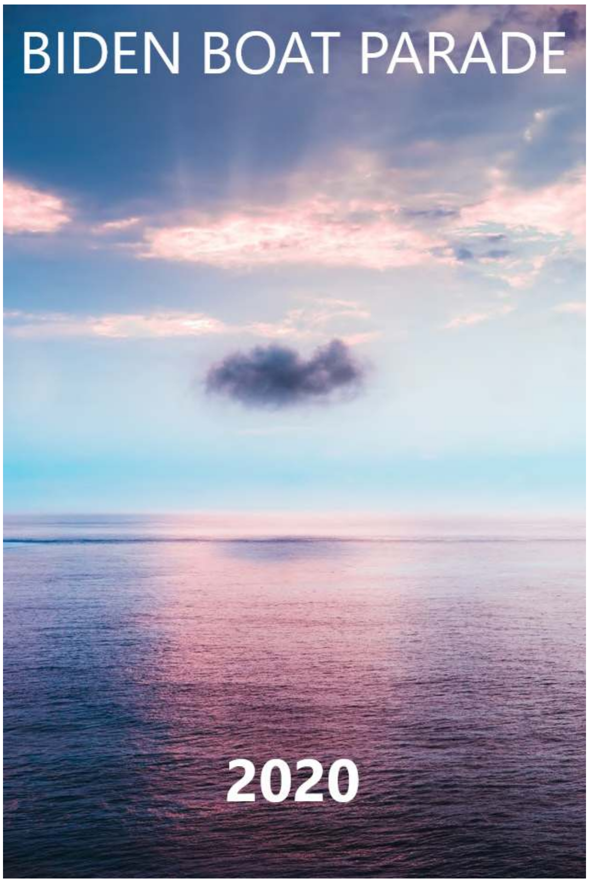


Figure 24: Example for **Smears**; ***Source(s):*** Image ; License

20. Glittering Generalities: These are words or symbols in the value system of the target audience that produce a positive image when attached to a person or issue. Peace, hope, happiness, security, wise leadership, freedom, "The Truth", etc. are virtue words. Virtue can be also expressed in images, where a person or an object is depicted positively.

Figure 25 shows an example of the use of this technique, in the right half of the meme. The technique covers the entire text span starting from "*2 & 1/2 years* . . ." until "*GDP up 3.2%* . . ." It is also expressed in the image, which depicts Donald Trump in a positive way. The text–image combination further strengthens the technique.

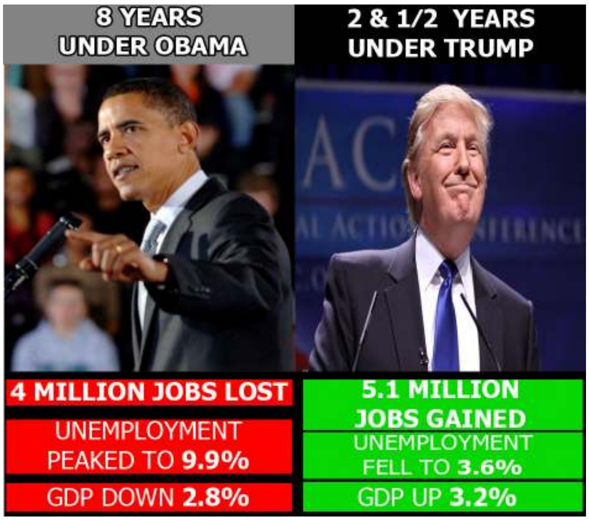


Figure 25: Example for **Glittering Generalities**; ***Source(s):*** Image 1, Image 2; License 1, License 2

21. Appeal to (Strong) Emotions: Using images with strong positive/negative emotional implications to influence an audience. We reserve this technique to the images content only.

An example is shown in Figure 26, which invokes strong emotions in the audience.

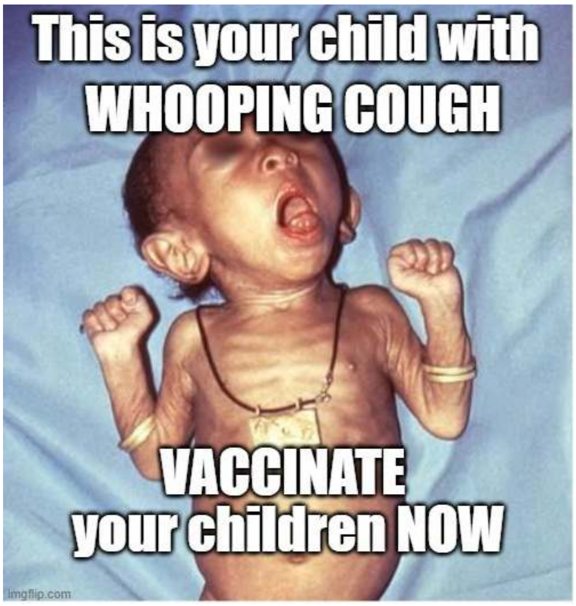

This is your child with WHOOPING COUGH
VACCINATE your children NOW
imgflip.com

Figure 26: Example for **Appeal to (Strong) Emotions**; ***Source(s):*** Image ; License

22. Transfer: Also known as *Association*, this is a technique of projecting positive or negative qualities (praise or blame) of a person, entity, object, or value onto another one to make the second one more acceptable or to discredit it. It evokes an emotional response, which stimulates the target to identify with recognized authorities. Often highly visual, this technique often utilizes symbols (for example, the swastikas used in Nazi Germany, originally a symbol for health and prosperity) superimposed over other visual images.

Figure 27 shows an example, where the *Transfer* technique makes use of a communist symbol (namely, hammer and sickle) on top of the pictures of two targeted politicians, with the aim of depicting them in a negative way. The technique is further reinforced by the use of the red color (which is also a symbol of Communism), and by the two instances of *Name Calling* ("*Moscow Mitch*" and "*Moscow's bitch*"), which make a connection to Moscow (which in turn was the capital of the former Communist block).

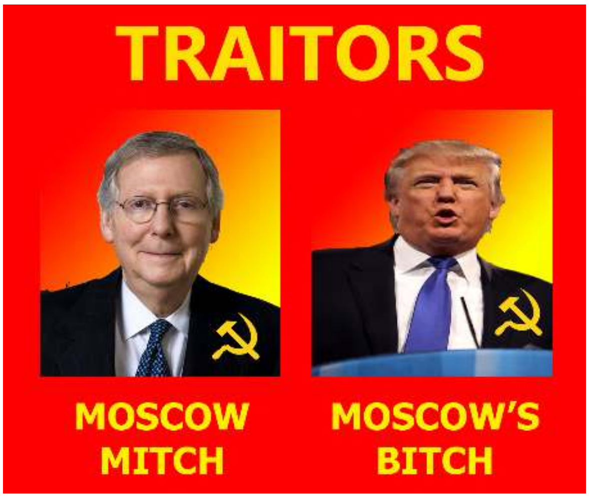


Figure 27: Example for **Transfer**; ***Source(s):*** Image 1, Image 2; License 1, License 2

B Subtasks: Definition, Data Format, and Data Examples

Below, we describe the three subtasks and the general data format for each of them. We further show an example of an annotated example for each subtask.

B.1 Subtask 1

This is a multi-label classification problem, defined as follows:

Subtask 1 (ST1) Given only the "textual content" of a meme, identify which of the 20 techniques are used in it.

The data for ST1 comes as a JSON object in the following format:

```
{
  id -> example identifier,
  labels -> list of persuasion
            techniques,
  text -> text of the meme
}
```

Here is an example:

```
{
 "id": "125",
 "labels": [
    "Loaded Language",
    "Name calling/Labeling"
 ],
 "text": "I HATE TRUMP\n\n
          MOST TERRORIST DO"
}
```

B.2 Subtask 2

ST2 is a more complex version of ST1, as it asks not only for the techniques but also for the exact spans of use each technique. This subtask is a combination of the two subtasks in *SemEval-2020 task 11*. It is a multi-label sequence tagging problem, defined as follows:

Subtask 2 (ST2) Given only the "textual content" of a meme, identify which of the 20 techniques are used in it together with the span(s) of text covered by each technique.

The data for ST2 comes as a JSON object with the following format:

```
{
  id -> example identifier,
  text -> text of the meme
  labels : [ -> list of objects
    {
      start -> start index,
      end -> end index,
      technique -> technique,
      text_fragment -> text
    }
  ]
}
```

Here is an example:

```
{
 "id": "125",
 "text": "I HATE TRUMP\n\n
         MOST TERRORIST DO"
 "labels": [
   {
    "start": 2,
    "end": 6,
    "technique": "Loaded Language",
    "text_fragment": "HATE"
   },
   {
    "start": 19,
    "end": 28,
    "technique": "Name calling/
    Labeling",
    "text_fragment": "TERRORIST"
   }
 ]
}
```

Note that the labels to be predicted for ST2 are the same ones as for ST1, but this time the spans are to be predicted as well.

B.3 Subtask 3

ST3 is a multi-modal version of ST1, where the image is also provided. It is a multi-label classification problem, defined as follows:

Subtask 3 (ST3) Given a meme, identify which of the 22 techniques are used both in the textual and in the visual content of the meme.

The data for ST3 comes as a JSON object with the following format:

```
{
  id -> example identifier,
  labels -> list of persuasion
            techniques,
  image -> name of the image file,
  text -> text of the meme
}
```

Here is an example:

```
{
 "id": "125",
 "labels": [
    "Loaded Language",
    "Name calling/Labeling",
    "Reductio ad hitlerum",
    "Smears",
 ],
 "image": "125_image.png"
}
```

Here, the image, which is shown in Figure 28), gives rise to two additional persuasion techniques compared to ST1: *Reductio ad Hitlerum* and *Smears*. These techniques are not clearly present in the text alone. Indeed, the image is needed for us to see that there is *Smears*, as this can be only seen when we understand that this is a dialog with a negative propaganda targeting one of the participants (Ilhan Omar). Similarly, we need the image for *Reductio ad Hitlerum*: the image shows us that Ilhan Omar is depicted as a bad person (she is targeted by the *Name Calling* "*terrorist*", and she is also the target of the *Smears*), and thus the message being conveyed is that any choice that such a bad person does has to be a bad choice, i.e., hating Trump is a bad thing to do as this is something terrorists do.

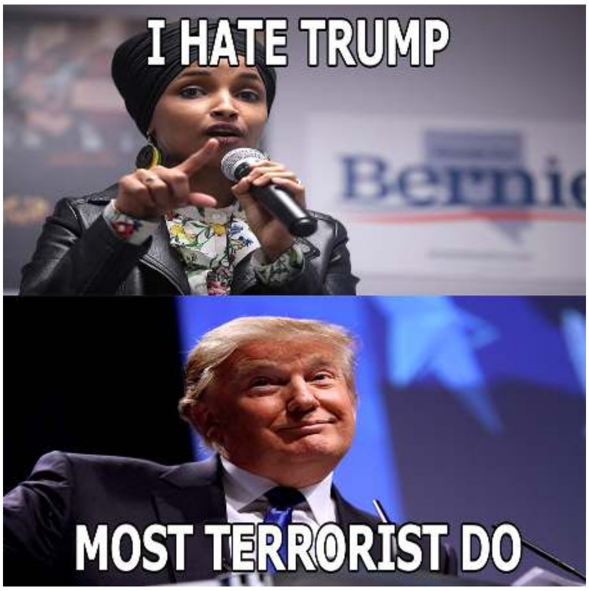


Figure 28: The meme with id=125; ***Source(s):*** Image 1, Image 2; License 1, License 2

C Participating Systems

Below, we give a brief description of the participating systems, listed in alphabetical order, with reference to the corresponding task description paper. The numbers in square brackets refer to the official ranking of the target system on the individual subtasks.

1213Li (Peiguang et al., 2021)[ST3: 3rd] used RoBERTa and ResNet-50 as feature extractors for texts and images. They used a label embedding layer with a multi-modal attention mechanism to measure the similarity between labels with the multi-modal information and fused features for label prediction.

AIMH (Messina et al., 2021) [ST1: 5th, ST3: 4th] used transformer-based models and proposed visual–textual transformers to mainly address subtask 3 (ST3). For the visual part, they used ResNet50, and for the textual part, they used BERT. The same network used the multi-label classification on text (ST1) by using only the textual part of the network.

Alpha (Feng et al., 2021) [ST1:2nd, ST3:1st] team pre-trained a transformer using text with visual features. They extract grid features, using ResNet50, and salient region features, using BUTD. They used grid features to capture the high-level semantic information found in the images. Additionally, they used salient region features to describe objects and to caption the event present in the memes. For ST1, they combined the text and the text representation of the visual features, and trained DeBERTa. For ST3, they built an ensemble of fine-tuned DeBERTA+ResNet, DeBERTA+BUTD, and ERNIE-VIL.

HOMADOS (Kaczyński and Przybyła, 2021) [ST2: 2nd] used a multi-task learning (MTL) approach with additional datasets such as the PTC corpus from SemEval-2020 (Da San Martino et al., 2020a), and a fake news corpus (Przybyla, 2020). The model was trained using BERT followed by several output layers, which solve auxiliary tasks of propaganda detection and credibility assessment in two distinct scenarios: sequential and parallel MTL, effectively accelerating the training process. The final submission used a parallel MTL approach on the propaganda detection of SemEval-2020, which ranked second.

TeamFPAI (Xiaolong et al., 2021) (ST2: 3rd) formulated the task as a question answering one in a machine reading comprehension (MRC) framework, which achieved better results compared to an ensemble-based approach (Liu et al., 2018). Moreover, data augmentation and loss design techniques were also explored to alleviate the problem of data sparseness and imbalance. Their system was ranked 3rd in the final evaluation phase.

CSECUDSG (Hossain et al., 2021) (ST1: 13th, ST2: 6th, ST3: 6th) participated in all three subtasks. For ST1, they used a majority vote late fusion on top of logistic regression, decision tree, and fine-tuned DistilBERT models. For ST2, they reformulated the task as one of multi-label classification, where a pre-trained BERT model was used to design binary classifiers for each technique in a multi-label classification setting. For ST3, they used a majority voting late fusion on top of fine-tuned DistilBERT, ResNet50, and a predicted label from an early fusion model. The early fusion model consisted of features from (*i*) multi-kernel CNN on top of the LSTM model with word embeddings including (*ii*) word2vec (Mikolov et al., 2013), (*iii*) word embeddings fine-tuned FastBERT (Liu et al., 2020), (*iv*) RoBERTa, (*v*) sentence embeddings from FastBERT, (*vi*) image features from YouTube-8M (Abu-El-Haija et al., 2016), and (*vii*) multimodal features from VisualBERT (Li et al., 2019).

LeCun (Dia et al., 2021) [ST1: 6th] trained five models and combined them in an ensemble. Initially, they pre-processed text using stemming. Later, they trained DebERTA and RoBERTa models with augmented data using synonym replacement, random insertion, random swap, random deletion and back-translation. They first trained the five models separately, and then they fine-tuned the ensemble on the official non-augmented data.

LIIR (Ghadery et al., 2021)[ST3: 8th] used data augmentation through back-translation and CLIP to obtain image and text representations, which were then fed to a chained classifier that uses the correlations between the output techniques.

LT3-UGent (Singh and Lefever, 2021) [ST3: 14th] participated in subtask 3 only. They used Multimodal Compact Bilinear Pooling to combine representations from ResNet-51 and BERT. They further fine-tuned on the PTC corpus (Da San Martino et al., 2020a).

MinD (Tian et al., 2021) [ST1: 1st, ST3: 2nd] used five pre-trained models for ST1: BERT, RoBERTa, XLNet, DeBERTa, and ALBERT. They first fine-tuned them on the PTC corpus (Da San Martino et al., 2020a), and then on the training data. For the final prediction, they averaged the probabilities of the models. They also used a post-processing rule: a bigram that appeared more than three times was flagged as a *Repetition*. The system for ST1 was also used for ST3, combined with (*i*) ResNet-34, a face recognition system, (*ii*) OCR-based positional embeddings for text boxes in the image, and (*iii*) Faster R-CNN to extract region-based image features. They combined the textual and the visual representations by averaging their probabilities. Other multimodal fusion strategies included concatenation of the representation and mapping them to the space using a multilayer perceptron.

NLP-IITR (Gupta and Sharma, 2021) [ST1: 15th] used an ensemble that included included fine-tuned RoBERTa, BERT, and three additional models. They further used pre-processing. To tackle data scarceness for some rare labels, they used data augmentation using back-translation.

NLyticsFKIE (Pritzkau, 2021) [ST1: 9th, ST3: 13th] used RoBERTa as a text encoder in ST1 and ST3. For ST1, they used RoBERTa's output to build a classifier to predict each label separately. For ST3, they still used RoBERTa to encode the text and a VGG-16 layer to encode the image. They used multiple copies of a cross-modality encoder that outputs an encoding of the image features with respect to the text features, and vice versa. The concatenation of the two cross-encoders' outputs was then passed through a residual layer followed by layer normalization.

Volta (Gupta et al., 2021) [ST1: 3rd, ST2: 1st, ST3: 5th] used a combination of transformers for all subtasks. For ST1, they used RoBERTa's [CLS] token, which they fed to a feed-forward neural network, and example weighting to take care of class imbalance. For ST2, they predicted token classes by considering the output of each token embedding as obtained by RoBERTa. To account for subwords' class, they merged each subword belonging to the same token and assigned the union of the subwords' labels. For ST3, they separately encoded the textual features (extracted using RoBERTa) and the multi-modal features (extracted using UNITER, VisualBERT, and LXMERT). This layer's input was a sequence of textual subwords and visual tokens extracted by keeping the top 36 regions of interest as returned by Faster R-CNN. A concatenation of the two different [CLS] tokens was then fed into an MLP, and weighted labels were used with a cross-entropy loss.

WVOQ (Roele, 2021) [ST2: 5th] used a novel approach to ST2 consisting of adopting an encoder–decoder strategy. The encoder encodes the passage, while the decoder generates a marked version of the input, where the markup outlines the various spans along with the classes they belong to. In this way, the system performed simultaneous span detection and classification. The encoder–decoder used a specialization of BART.

YNU-HPCC (Zhu et al., 2021) [ST1: 12th, ST2: 7th, ST3: 11th] For ST1, they used a CNN on top of ALBERT and fine-tuned the model for multi-label classification. For ST2, each propaganda technique was considered as an independent task, and features were extracted from the pre-trained BERT model. Subsequently, the problem was addressed as a multi-task sequence labeling one, and the results for each task were combined. For ST3, a multi-modal network was used, where embeddings from textual and visual networks were concatenated, which was followed by a fully connected layer. For the text, the same approach was used for ST1, and for the image, ResNet and VGGNet were used for image feature extraction.