Paper Bullets: Modeling Propaganda with the Help of Metaphor

Daniel Baleato Rodríguez
ILLC, University of Amsterdam
daniel@codealia.com

Verna Dankers
ILCC, University of Edinburgh
vernadankers@gmail.com

Preslav Nakov
MBZUAI
preslav.nakov@mbzuai.ac.ae

Ekaterina Shutova
ILLC, University of Amsterdam
e.shutova@uva.nl

Abstract

Propaganda aims to persuade an audience by appealing to emotions and using faulty reasoning, with the purpose of promoting a particular point of view. Similarly, metaphor modifies the semantic frame, thus eliciting a response that can be used to tune up or down the emotional volume of the message. Given the close relationship between them, we hypothesize that, when modeling them computationally, it can be beneficial to do so jointly. In particular, we perform multi-task learning with propaganda identification as the main task and metaphor detection as an auxiliary task. To the best of our knowledge, this is the first work that models metaphor and propaganda together. We experiment with two datasets for identifying propaganda techniques in news articles and in memes shared on social media. We find that leveraging metaphor improves model performance, particularly for the two most common propaganda techniques: loaded language and name-calling.

1 Introduction

Propaganda aims to influence an audience. It is a type of information that, whether true or false, tries to promote a particular agenda (Cantril, 1938) by appealing to emotions or by using faulty reasoning (Miller, 1939). Although this communication strategy comes in many forms, it is conveyed using specific persuasion techniques that exploit our psychology to sell us an idea or a point of view (Da San Martino et al., 2019b). In Figure 1, we can see an example of such techniques used in a meme shared on social media.

Another rhetorical device at the heart of many successful communication strategies is *metaphor*. Postulated as a primordial mechanism to conceptualize what we think and experience (Lakoff, 1980), metaphor works by mapping a concept in one domain (often a physical domain) to another domain (usually an abstract one) by means of a systematic

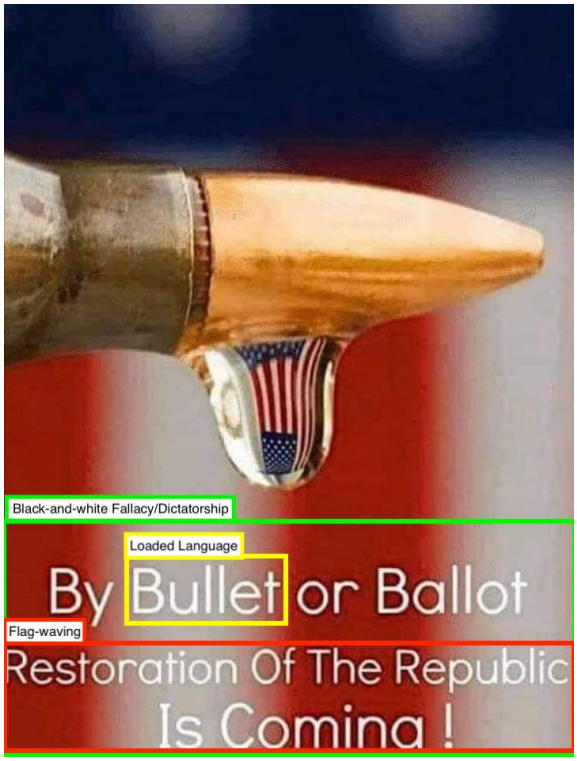

Figure 1: Meme containing propagandistic techniques (Dimitrov et al., 2021). These techniques are highlighted with bounding boxes for illustration purposes.

association. For instance, the term "*paper bullets*"[1] connects the domains of information and war, illustrating the weaponization of information.

In the same way that propaganda can exploit automatic shortcuts our brain uses to process information (*e.g.*, stereotypes) (Tversky and Kahneman, 1974), metaphors can affect how we reason about a particular situation or issue by evoking a different semantic frame (Fillmore et al., 2006). Research shows that characterizing crime as a *beast* delivered more punishment-oriented strategies to *fight* crime (Thibodeau and Boroditsky, 2011). Con-

[1]The metaphor "*paper bullets*" was used during World War II, where the Germans used tactical aircrafts to drop anti-Semitic leaflets over American troops (Margolin, 1946) as a way of psychological warfare.

Findings of the Association for Computational Linguistics: EACL 2023, pages 472–489
May 2-6, 2023

versely, referring to crime as a *virus* gathered a more significant number of preventive measures to *cure* it. As a persuasive device, framing has successfully been used in politics (Howe, 1988; Ana, 1999; Lakoff, 2009) to shift the public opinion about a particular topic. Moreover, the use of metaphors by politicians in their posts on social media increases engagement with their electorate (Prabhakaran et al., 2021).

Some propagandist techniques and metaphors can exhibit a similar intention by the author. For instance, the most common technique is the use of *loaded language* to increase the emotional response of the audience (e.g., "... *disastrous* [nuclear deal]"). Likewise, metaphor can also elicit an emotionally charged reaction (Mohammad et al., 2016). The following example combines both: "the *ruinous* reforms". Similarly, *name-calling* connects the object of the propaganda campaign with terms the target audience sees positively or negatively (Miller, 1939). This technique seeks a love or hate emotional response, and it could also alter the semantic frame (e.g., "*Crooked* Hillary" or "*Deep* State officials").

Other salient examples where different propagandist techniques employ metaphor can be found in the Propaganda Techniques Corpus (PTC) (Da San Martino et al., 2019b), including *exaggeration* ("a *tsunami* of lies and smear"), *appeal to fear* ("[bubonic plague in Madagascar] could even *spill over* into neighboring countries and beyond"), *doubt* ("Why is the U.S. *singling out* Iran ...") and *flag-waving* ("it is time to *take* our government *back* ..."), among others.

We explore how metaphor detection can aid propaganda technique classification under the multi-task learning paradigm. Computational modeling for propaganda detection was initially studied as a document-level classification task in news articles (Rashkin et al., 2017; Barrón-Cedeño et al., 2019; Martino et al., 2020). More recently, annotation efforts produced datasets that identify the text spans where particular forms of propaganda are used. Our work builds upon the most extensive corpus of fragment-level propaganda techniques to date (Da San Martino et al., 2019b) and on shared task 6 from SemEval-2021 (Dimitrov et al., 2021) to identify persuasive techniques in both news articles and internet memes, respectively. We analyze how a multi-task learning approach that leverages metaphor detection can improve results in propaganda identification.

To our knowledge, this is the first study of the role of metaphor in computational propaganda identification. We produce the first models that combine the two phenomena and analyze their predictive capability, both quantitatively and qualitatively.

Our findings show that metaphor detection can increase performance for certain types of propaganda. We see improvements across multiple tasks covering both datasets. The gains are more pronounced for *name-calling*, with significant results for the news domain. Furthermore, our models' predictions suggest that propagandist content uses figurative language more extensively than non-propagandist text.

2 Related work

2.1 Metaphor detection

NLP applications need to distinguish the particular intent that metaphor plays in context (Veale et al., 2016). Metaphor detection research has studied various approaches: hand-crafted features and word classes (Beigman Klebanov et al., 2016), concreteness and imageability word ratings (Broadwell et al., 2013; Turney et al., 2011), semantic classification making use of lexical databases (e.g., WordNet, VerbNet, ConceptNet) (Wilks et al., 2013; Neuman et al., 2013; Mohler et al., 2013; Tsvetkov et al., 2013), distributional semantic models (Gutierrez et al., 2016; Bulat et al., 2017; Hovy et al., 2013), and even visual (Shutova et al., 2016) or sensorial features (Tekiroglu et al., 2015). More recently, deep learning methods (Mao et al., 2019; Dankers et al., 2020; Gao et al., 2018; Rei et al., 2017; Wu et al., 2018) have been used to detect metaphors.

Current state-of-the-art textual metaphor detection is powered by large pre-trained neural network models (Su et al., 2020; Chen et al., 2020; Gong et al., 2020; Choi et al., 2021) that have been trained using datasets of billions of words. These models can leverage word representations that carry context-sensitive semantic information. As the latest shared task on metaphor detection highlights (ACL 2020) (Leong et al., 2020), more than half of the participants used BERT (Devlin et al., 2019) or its variants, widely successful pre-trained models that perform well on downstream tasks.

In addition, metaphor detection has successfully been used as an auxiliary task in multi-task learning (MTL) (Caruana, 1993) for emotion classification

(Dankers et al., 2019), political perspective, affiliation, and framing (Huguet Cabot et al., 2020); and aspect-based sentiment analysis (Mao and Li, 2021), among others. The MTL approach builds on the idea that the same model can encode valuable features for different tasks that would help each other's performance. As metaphor is extensively used in everyday language and dramatically influences the expressiveness of the message, it can help in a significant number of semantic tasks.

2.2 Propaganda detection

Propaganda is closely related to political bias and misinformation (colloquially referred to as *fake news*) (Guess and Lyons, 2020). This area of research has gained popularity in the last decade due to concerns regarding the weaponization of social media and how it can negatively affect political discourse (Wardle and Derakhshan, 2017). Work on political bias commonly uses lexicon-based approaches to detect sentiment on political topics, while models to expose fake stories usually rely on publishing patterns and knowledge graphs (Haq et al., 2020).

However, propaganda does not necessarily have to be politically driven or rely on untrue or incorrect information. While some instances of propaganda usually do (e.g., clickbait) (Martino et al., 2020), propagandist content varies in accuracy and the acknowledgment of its sources (Jowett et al., 2012). In essence, propaganda aims to influence an audience to exercise a particular agenda (Cantril, 1938) by appealing to emotions or faulty reasoning (Miller, 1939).

Computational approaches to propaganda detection are relatively recent and were initially directed to the document classification of varying sizes, from news articles to tweets (Barrón-Cedeño et al., 2019; Rashkin et al., 2017; Volkova et al., 2017). Proposed models used BERT, LSTM (Hochreiter and Schmidhuber, 1997), Convolutional Neural Networks (CNN) (LeCun et al., 1995), and Naive Bayes models powered by Glove (Pennington et al., 2014) embeddings. These works rely to different degrees on the labeling of information sources by crowd-sourced groups or non-profit organizations (e.g., MBFC[2], PropOrNot[3]). Unfortunately, this categorization approach can introduce noise into the system. Reliable news agencies might occasionally include a propagandist article to fulfill their interest. Conversely, highly propagandist media could publish a non-propagandist piece to boost their credibility.

The latest propaganda detection approaches take advantage of the rhetorical devices that propaganda uses to influence reasoning. Although the literature compiles different accounts of propagandist or persuasive techniques (Miller, 1939; Shah, 2005; Abd Kadir and Abu Hasan, 2014), they are mainly sub-types of the general principles first proposed in Cantril (1938), which share the aim of connecting an idea or propagandist object to an attitude or emotion.

The PTC corpus (Da San Martino et al., 2019b) was the first effort to classify propaganda at a more granular level. It identifies 18 persuasive techniques across 451 news articles, making it the largest of its kind. It annotates the start and end of each propagandist fragment. This corpus, and a later variant, were used in shared tasks on propaganda detection (Da San Martino et al., 2019a, 2020). The best systems used pre-trained Transformer-based models and ensembles (Yoosuf and Yang, 2019; Jurkiewicz et al., 2020; Morio et al., 2020; Chernyavskiy et al., 2020).

More recently, SemEval 2021 Task 6 (Dimitrov et al., 2021) has expanded fragment-level propaganda identification efforts outside the news corpora. It identifies propaganda techniques ingrained in the combination of textual and image data. The task's dataset consists of 950 internet memes posted on social media with topics related to politics, vaccines, COVID-19, and gender equality. Apart from identifying 20 textual propagandist techniques, it also identifies two that are only present when in combination with the image. The most common and best-performing models used for textual tasks were the transformer-based models BERT and RoBERTa (Kaczyński and Przybyła, 2021; Gupta et al., 2021).

3 Tasks and datasets

In this work, we examine six tasks for fragment-level propagandist technique identification. Half of them use labeled data from news articles, while the others use textual information from memes shared on social media. For each domain, we perform a multi-label classification task — to identify all propagandist techniques in the dataset — and two single-label classification tasks to detect the two

[2] https://mediabiasfactcheck.com

[3] http://www.propornot.com/p/the-list.html

most common persuasive techniques: *loaded language* and *name-calling*. The single-label tasks ignore the rest of the labels in the dataset while using the same textual input as the multi-label tasks.

In addition, MTL models include metaphor detection as an auxiliary task. This task aims to detect all content words used as metaphors in a given text.

3.1 VUA Metaphor Corpus

We use the data from the ACL 2020 shared task on metaphor detection (Leong et al., 2020). Specifically, the all-POS subtask that identifies which content words (i.e., nouns, verbs, adjectives, and adverbs) are used in their metaphorical sense. The data for the task comes from the VU Amsterdam Metaphor Corpus, (Steen et al., 2010) which contains annotations for all words in 117 texts from the British National Corpus (Clear, 1993) and across four different registers: academic text, conversation, fiction, and news. The dataset covers 190K lexical units over 16,189 sentences with a train/test split of 12,109 and 4,080 sentences. The prevalence of metaphorical use for content words is 6.8% for the training set and 7.7% for the test set. We randomly sample 10% of the training split for validation.

3.2 Propaganda Techniques Corpus

The PTC corpus (Da San Martino et al., 2019b) identifies 18 propaganda techniques across 451 articles (350K tokens) from 49 news outlets. The annotations were produced by separate teams of annotators and merged through a consolidation process where all disagreements were discussed before becoming part of the final version. Each annotation identifies the technique used and its start and end within the news article. The dataset contains 20,339 sentences split into training, validation, and test sets with 14,263, 2,034, and 4,042 sentences, respectively.

The number of instances per technique and its length varies widely. The most common classes are *loaded language* with 2,547 occurrences and *name-calling* with 1,294. Those techniques have been used an average of 6.7 and 4.7 times per article, whereas all others appear a maximum of twice per article. We evaluated these two techniques separately as they provide a larger number of positive examples and can relate to metaphor as described in Section 1. Details on the number of annotations per split and their average length are shown in Table 1.

Dataset	#Annotations				Length
Prop. technique	Total	Train	Val	Test	Avg. ± SD
News					
Loaded language	2,547	1,811	304	432	23.70 ± 25.30
Name-calling	1,294	931	154	209	26.10 ± 19.88
All combined	7,480	5,114	927	1,439	46.99 ± 61.45
Memes					
Loaded language	761	543	68	150	14.87 ± 18.17
Name-calling	408	301	37	70	17.00 ± 11.65
All combined	2,083	1,498	182	403	40.43 ± 48.91

Table 1: Statistics on propaganda technique annotations and their average length in characters.

3.3 Propaganda detection in memes

SemEval-2021 Task 6 (Dimitrov et al., 2021) aims to identify the propagandist technique used in memes shared on social media. The images were collected from 26 public Facebook groups, which provided memes on the following topics: politics, COVID-19, pro-vaccines, anti-vaccines, and gender equality. The annotation process involved a heterogeneous group of annotators and a consolidation step. The text of the images was retrieved automatically using Google Vision API[4] and manually corrected afterward. We focus on subtask two, which only uses the textual data of the meme to predict where in the text a particular technique is present.

The dataset contains 951 examples (16,840 tokens) divided into 688, 63, and 200 samples for train, validation, and test splits. The average number of sentences per meme is 1.68, with a maximum of 13 sentences in one image alone. Again, the most common techniques are *loaded language* with 761 annotations (36.5%) and *name-calling* 408 occurrences (19.6%) from 2,083 propagandist fragments.

We provide a summary of the textual persuasion techniques in the Appendix Section A.1 and examples in the Appendix Table 12.

4 Methods

4.1 Models

We employ the pre-trained RoBERTa-BASE model (Liu et al., 2019). RoBERTa shares its architecture with its counterpart BERT (Devlin et al., 2019), but it improves performance across many tasks due to its highly optimized training and the use of ten times more data.

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

RoBERTa tokenizes inputs using byte-pair encodings (Sennrich et al., 2016) and computes contextualised embeddings for these input tokens. We add task-specific classifiers on top of RoBERTa, consisting of a linear layer followed by the sigmoid activation function. During inference, tokens with predicted targets over 0.5 are assigned to the class corresponding to the classifier. We fine-tune all of the model parameters in our respective tasks. Since the datasets provide labels at the word level, we aggregate predictions for words consisting of multiple tokens. If the model predicts any token to belong to a particular class, we assign the label to the whole word.

4.2 Single-task learning

Our main task is to detect the text span of each propaganda technique from news articles and memes. When solely training the model on a propaganda task, we refer to this as *single-task* learning (STL). The standard propaganda task as introduced in Section 3 is *multi label*. Since propagandist fragments can overlap, we perform multi-label classification by predicting the presence of each technique independently at each token, using separate task classifiers per technique as described in Section 4.1.

In addition to the multi-label propaganda technique identification, we generate two *single-label* tasks targeting the most frequent persuasion techniques (i.e., *loaded language* and *name-calling*). Both techniques share common aspects with metaphor discussed in Section 1, making them particularly interesting for experimentation.

4.3 Multi-task learning

In the MTL setup, we train the model jointly on two tasks: one of the propaganda identification tasks and the metaphor detection task. Similar to the STL setup, we do this both for single-label and multi-label classification. As the model learns to identify metaphors, we hypothesize that the metaphor-related features benefit the propaganda technique identification.

We extend the STL models with an additional classifier to predict metaphor as the auxiliary task. All tasks share the pre-trained model (RoBERTa) in a hard parameter sharing fashion. For fine-tuning, we reuse the best configuration from the single-task models to facilitate comparison between the two strategies. We experiment with different MTL regimes attending to the following hyper-parameters:

- Task sampling ratios (r_a, r_m): these ratios are used to select a task at each update step during training. With a probability of $p_m = r_m/(r_a + r_m)$ the main task is selected, and with $p_a = r_a/(r_a + r_m)$ the auxiliary task is selected.
- Epoch sampling coefficients (c_a, c_m): these coefficients are used to update the sampling ratios at every epoch. At epoch n, $r_{m_n} = r_{m_{n-1}} \times c_m$, and , $r_{a_n} = r_{a_{n-1}} \times c_a$.
- Loss scaling factors (s_a, s_m): these hyper-parameters are used to scale the losses for the main task (s_m) and the auxiliary task (s_a).

Although the MTL models have access to more data, as they are trained on two datasets, we limit their computational budget to match the one available for STL models. Every epoch, the model is trained in iterations, where the number of iterations is the same as for the STL model. Each iteration randomly selects a task for training according to its sampling probability p. We shuffle all examples in the training set at the start and after exhaustion of the training split. We fill each batch with samples from the selected task at random without replacement.

5 Experiments and results

5.1 Experimental Setup

In the implementation, we use the PyTorch framework and the pre-trained RoBERTa-base[5] model from the *transformers* library (Wolf et al., 2020). We trained all models using a maximum sequence length of 512 tokens, a weight decay of 0.01, and the AdamW optimizer (Loshchilov and Hutter, 2017) with a 10% warm-up period and a cosine-based learning rate decay function. All hyper-parameter search trials and the selected configurations for each task are listed in Appendix Table 11. We use the binary cross-entropy loss with modified class weights to account for class imbalances. Hyper-parameter search trials are performed over five different random seeds that dictate the order of data presentation and the initialisation of the task-specific classifiers. For the final configuration, performance is computed over ten different random seeds.

To ensure that the MTL models do capture meaningful features for metaphor identification – in spite

[5] https://huggingface.co/roberta-base

of it being the auxiliary task – we discard hyperparameter combinations with a median F1-score below 0.6 for metaphor identification, asserting the models exceed the baseline level set out by baselines one and two of Leong et al. (2020).

We evaluate the performance of propaganda detection based on the micro-averaged F1 score using precision (P) and recall (R) metrics defined in Da San Martino et al. (2019b). These metrics give partial credit to imperfect matches to account for overlap between techniques and the significant variation in length between propagandistic fragments. We provide details of these calculations in the Appendix A.2. We use statistical bootstrapping (Efron, 1979) to test the significance of our results and detail the procedure in Appendix A.3. We detail the system and configurations used for this work in Appendix A.4 for reproducibility.

5.2 Results

5.2.1 Single-label propaganda technique detection

Table 2 shows the performance for single-label propaganda detection tasks. The MTL approach improves results for all single-label tasks. Adding metaphor increases performance in news articles by 1.02 points for *name-calling*, from 28.72 to 29.74. This growth is statistically significant under the paired bootstrap test between learning strategies. The improvement is milder for *loaded language*, with a gain of 0.22 points, although results were more stable, almost halving the standard deviation for the metric.

We observe similar results in the memes dataset. Detection of *name-calling* improves the F1 metric by 1.24 points to 57.77 when training with metaphor as an auxiliary task. This increase is also more stable, lowering the standard deviation from 2.44 to 1.26. *Loaded language* improvements are smaller, adding 0.34 points to a total of 65.5 with lower variability.

5.2.2 Multi-label propaganda technique detection

Table 3 shows the results for the multi-label propaganda identification task in the news dataset. We compare our models to previous work (Da San Martino et al., 2019b) and achieve better results using a similar pre-trained model with the same number of parameters. The multi-task models obtained the best overall performance with an F1 of 24.32 and a standard deviation of 0.48. The single-task models averaged 23.78 F1 score with a much higher variation ($\sigma = 2.03$).

	Loaded Language			Name Calling		
Model	P	R	F1	P	R	F1
News						
STL	32.67	**48.04**	38.72 ± 1.00	24.48	35.25	28.72 ± 1.14
MTL	**33.60**	46.88	**38.94** ± 0.51	**26.30**	**35.35**	**29.74** ± 1.64
Memes						
STL	**68.88**	62.21	65.16 ± 2.16	52.39	**61.70**	56.53 ± 2.44
MTL	66.29	**64.90**	**65.50** ± 1.62	**59.03**	57.13	**57.77** ± 1.26

Table 2: Propaganda detection performance for single-label models. Statistically significant differences between STL and MTL are underlined ($p < 0.05$).

Model	P	R	F1
Da San Martino et al. (2019b)	**24.42**	**21.05**	**22.58**
Multi-label STL	20.37	**30.42**	23.78 ± 2.03
Multi-label MTL	**21.98**	27.46	**24.32** ± 0.48

Table 3: Propaganda technique identification results in news articles. The highest performance per model type is shown in bold. Underlined values denote statistical significance ($p < 0.05$) via paired bootstrap test between single-task and multi-task models.

Results for multi-label propaganda detection in memes are shown in Table 4. State-of-the-art performance for the task reaches an F1 score of 47.6, (Gupta et al., 2021) but it uses a model with 340M parameters. This model is three times larger than the ones we used (110M parameters). Comparing performance across same-size models, we see that our STL model performs best with an average F1 score of 46.22 and a standard deviation of 1.82. In contrast, the multi-task model achieves 44.81±1.31. Both models outperform the value of 43.9±0.9 reported in Gupta et al. (2021).

In the Appendix, Tables 8 and 9 show the performance of multi-label models for all techniques in the news and memes datasets, respectively.

6 Analysis and discussion

Given the shared traits between the use of metaphor and specific propagandist techniques, we hypothesized that it can be beneficial to model them jointly. We split the analysis into two subsections discussing quantitative and qualitative aspects.

6.1 Quantitative analysis

The results show improvements across most propaganda detection tasks when trained in a multi-

Model	P	R	F1
Volta (RoBERTa-Large)	-	-	47.6 ± 1.5
Volta (RoBERTa-Base)	-	-	43.9 ± 0.9
Multi-label STL	**46.02**	46.51	**46.22** ± 1.82
Multi-label MTL	42.62	**47.82**	44.81 ± 1.31

Table 4: Propaganda technique identification results in memes. We include the winning team for the shared task: *Volta* Gupta et al. (2021). The highest performance per model type is shown in bold. Underlined values denote statistical significance ($p < 0.05$) via paired bootstrap test between single-task and multi-task models.

task setting with metaphor as the auxiliary task. We hypothesised metaphor detection would benefit the single-label tasks, due to the use of a different semantic frame in *name-calling* and emotionally charged vocabulary in *loaded language*. Improvements were more pronounced for *name-calling* in both datasets, which suggests that, as anticipated, metaphorical framing plays a role in this propaganda technique. The fact that the gain in F1-score is the largest for name calling in both datasets further strengths this conclusion.

To further consolidate the relationship between proganda and metaphor our models identify, we investigate the prevalence of metaphors' predictions in propagandistic text fragments. We use our MTL models to predict metaphors on the propaganda corpora, and observe a higher percentage of metaphors in propagandist fragments than for non-propagandist content, and even higher for *loaded language* and *name-calling*. This is shown in the Appendix, in Figures 2 and 3. These model predictions hint at the likelihood that propagandist content, and some techniques in particular, may resort to metaphor more often than non-propagandist text does. Manual annotation of metaphors in propaganda datasets will allow asserting this with certainty, yet, we leave this for future work.

Although a slight improvement in task performance was observed for multi-label propaganda identification in the news dataset, this was not the case for the memes task. This task was the only one for which the MTL strategy was not superior. The memes dataset is 20 times smaller than the news dataset and includes two more labels. These challenges of size and sparsity could play a role in the utility of the MTL architecture, particularly when imposing on it the best hyper-parameters from the single-task models. We did this to facilitate the comparison between models, but we risk ending up with a configuration especially harmful to the MTL approach. Further experimentation is needed to investigate this drop in performance.

6.2 Qualitative analysis

To validate the effect of metaphor for the tasks, we pooled the predictions for all ten models of the same type trained with different seeds. We use simple majority voting to harmonize predictions across the different runs. Next, we identify the difference in the predicted spans between single-task and multi-task models. We include gold labels and the predicted metaphors by multi-task models for analysis. Examples of models' predictions for news articles and memes are shown in Table 5.

MTL models can detect figurative language, which contributes to detecting propaganda techniques that use this device. Idioms such as "*throw out the window*" (ref. LL.N.1) and "*kick the can down the road*" (ref. LL.N.2) are correctly identified, albeit partially, as *loaded language* in the context of the news article. This is also the case for the metaphorical use of the word "*dinosaurs*", present in an example of *name-calling*, to convey the point of view that current social media platforms will *go extinct* (ref. NC.N.1).

Other instances of non-literal meaning deliver incorrect predictions. However, we believe that some of those instances could be considered correct. In the case of *name-calling*, the models detect "*poor sport*" (ref. NC.N.2), which is alluding to a defeated candidate in an electoral race. Similarly, the phrase "*you can throw us in jail, but you will never defeat us*" (ref. LL.N.3) signals defiance with a considerable degree of emotion which borderlines the *loaded language* category.

Conversely, the label *hardworking* used in *"hardworking Georgians"* (ref. NC.N.3) cannot be attributed to *name-calling* as it does not refer to the propagandist target of the article: Georgia gubernatorial candidate Stacy Abrams. This mislabeled example highlights the task's difficulty and need for a broader context. Our models received individual sentences for training and inference, which is insufficient in this instance to identify the object of the propaganda campaign.

Looking at predictions on the memes dataset, we observe that the gains in *name-calling* for multi-task models were driven primarily by minimizing incorrect predictions. The examples NC.M.1 and

PT	Reference	Text fragments
Name-calling	NC.News.1	... we will **rise** from the **ashes** of the social media **dinosaurs** to help **build** and **create** new **platforms** ...
	NC.News.2	Talk about a **poor sport**, but Democrats are often like that in these races.
	NC.News.3	"The election is over and **hardworking** Georgians are **ready** to **move forward**," he said.
	NC.Memes.1	**HOLD** UP!!! Sleepy Joe **broke** my record?!?!?!?
	NC.Memes.2	... the most **corrupt**, **lying** and **despised member** of Congress and the **WORST** Speaker of the house ...
	NC.Memes.3	So Don King and Beetlejuice had a baby...
	NC.Memes.4	WARNING **SIGNS** OF A CULT // ...
	NC.Memes.5	ATTENTION PATRIOTS // **MEET** YOUR CIVIL WAR OPPONENTS
Loaded Language	LL.News.1	Political correctness **needs** to be **thrown out** the **window** when **dealing** with those who...
	LL.News.2	In other words, let's just **kick** the **can** down the **road** and hope for a more reasonable Iranian regime ...
	LL.News.3	You can **throw** us in **jail**, but you will never **defeat** us.
	LL.Memes.1	WHEN TRUMP IS REELECTED THERE WILL BE **BLOOD** !
	LL.Memes.2	WE ARE AT **WAR** !
	LL.Memes.3	**FAKE** WINNER
	LL.Memes.4	... UNDERCOVER FEDS DOCUMENTING THE FRAUD AND THEY'VE STEPPED INTO A **TRAP**

Table 5: Example predictions of propaganda techniques. Gold labels in yellow, predictions in blue, and their intersection in green. The underline style identifies predictions only produced by one learning strategy. Predicted metaphors from MTL models are shown in bold.

NC.M.2 were the only ones containing prediction spans singular to the multi-task models. Both instances correctly label parts of the text that do not include predicted metaphors, although they contain metaphors in their vicinity. In contrast, the single-task models produced more mislabels on nouns or noun phrases, see examples NC.M.3, NC.M.4 and NC.M.5. With respect to *loaded language*, we observe metaphor predictions falling equally into correct and incorrect spans, see examples LL.M.1, LL.M.2, LL.M.3, LL.M.4.

7 Conclusion and Future Work

In this work, we explored the influence of metaphor detection on propaganda technique identification in a multi-task learning setup. Joint modelling of metaphor and propaganda was performed using two propaganda datasets from different domains: news articles and internet memes. We experimented with six different propaganda detection tasks, including multi-label propaganda technique identification and single-label tasks for the two most common propagandist techniques: *name-calling* and *loaded language*, for each dataset. Incorporating metaphor detection yielded performance improvements in five of the six tasks considered, with the highest improvements observed for the *name-calling* technique. Moreover, the different datasets showed similar patterns in performance changes. We supplemented the task performance results with an analysis of the prevalence of metaphor in the propaganda corpora and qualitatively examined a range of examples of metaphorical language use in propagandist fragments. We are the first to investigate the interaction of these two phenomena and our promising results encourage further research in this direction.

In future work, we plan to extend our analysis to other propaganda techniques. In view of the emergence of datasets for other languages, such as an the Arabic propaganda detection shared task at WANLP'2022 and the multilingual SemEval-2023 task 3 subtask 3 on propagada detection in English, French, German, Italian, Polish, and Russian, we plan future multi/cross-lingual experiments.

Limitations

Although we established a positive influence of metaphor detection on propaganda technique identification, our work also has some limitations. (1) Considering that this work focused on the two most common propagandist techniques, future work could extend this analysis to cover others, although we should note that these analyses are limited by a data scarcity issue (in particular in the memes dataset). (2) While we considered six tasks, these tasks used one MTL architecture. Previous work has experimented with more advanced MTL methods (e.g., soft parameter sharing) and in the future, these methods could also benefit joint learning of metaphor and propaganda. (3) Finally, it should be emphasised that both types of propaganda employed and the types of figurative language used are very specific to cultures and languages. As such, the techniques applied in this study might not deliver the same effect when using data from different geographical locations, or data from languages other than English. Moreover, the prevalence of metaphor varies across different propagandist techniques, meaning that not every propaganda-related task will benefit from joint learning with metaphor.

Ethics and Broader Impact

Intended Use and Misuse Potential Our models can be of interest to the general public, fact-checkers, and journalists. However, they could also be misused by malicious actors. We, therefore, ask researchers to exercise caution.

Environmental Impact We would like to warn that the use of large language models requires a lot of computations and the use of GPUs/TPUs for training, which contributes to global warming (Strubell et al., 2019). This is a bit less of an issue in our case, as we do not train such models from scratch, we just fine-tune them.

References

Shamsiah Abd Kadir and Ahmad Sauffiyan Abu Hasan. 2014. A content analysis of propaganda in harakah newspaper. *Journal of Media and Information Warfare (JMIW)*, 5:73–116.

Otto Santa Ana. 1999. 'Like an Animal I was Treated': Anti-Immigrant Metaphor in US Public Discourse. *Discourse & Society*, 10(2):191–224.

Alberto Barrón-Cedeño, 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.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106, Berlin, Germany. Association for Computational Linguistics.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using Imageability and Topic Chaining to Locate Metaphors in Linguistic Corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 102–110, Berlin, Heidelberg. Springer Berlin Heidelberg.

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Modelling metaphor with attribute-based semantics. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 523–528, Valencia, Spain. Association for Computational Linguistics.

Hadley Cantril. 1938. Propaganda analysis. *The English Journal*, 27(3):217–221.

Richard A. Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Machine Learning Proceedings 1993*, pages 41–48. Elsevier.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go Figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online. Association for Computational Linguistics.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Jeremy H. Clear. 1993. *The British National Corpus*, page 163–187. MIT Press, Cambridge, MA, USA.

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*, pages 162–170, Hong Kong, China. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5635–5645, Hong Kong, China. Association for Computational Linguistics.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. 2020. Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 227–234, Online. Association for Computational Linguistics.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

B. Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1).

Charles J Fillmore et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural Metaphor Detection in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. IlliniMet: Illinois System for Metaphor Detection with Contextual and Linguistic Information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153, Online. Association for Computational Linguistics.

Andrew M. Guess and Benjamin A. Lyons. 2020. *Misinformation, Disinformation, and Online Propaganda*, SSRC Anxieties of Democracy, page 10–33. Cambridge University Press.

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 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1075–1081, Online. Association for Computational Linguistics.

E.Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and Metaphorical Senses in Compositional Distributional Semantic Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193, Berlin, Germany. Association for Computational Linguistics.

Ehsan Ul Haq, Tristan Braud, Young D. Kwon, and Pan Hui. 2020. A Survey on Computational Politics. *IEEE Access*, 8:197379–197406. Conference Name: IEEE Access.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dirk Hovy, Shashank Shrivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying Metaphorical Word Use with Tree Kernels. In *Proceedings of the First Workshop on Metaphor in {NLP}*, pages 52–57, Atlanta, Georgia.

Nicholas Howe. 1988. Metaphor in Contemporary American Political Discourse. *Metaphor and Symbolic Activity*, 3(2):87–104.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Garth Jowett, Victoria O'Donnell, and Garth Jowett. 2012. *Propaganda & persuasion*, 5th edition. SAGE, Thousand Oaks, Calif. OCLC: ocn674939375.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Konrad Kaczyński and Piotr Przybyła. 2021. HOMADOS at SemEval-2021 task 6: Multi-task learning for propaganda detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1027–1031, Online. Association for Computational Linguistics.

George Lakoff. 1980. *Metaphors we live by*. University of Chicago Press, Chicago [etc.

George Lakoff. 2009. Metaphor and War: The Metaphor System Used to Justify War in the Gulf. *Cognitive Semiotics*, 4(2).

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Rui Mao and Xiao Li. 2021. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13534–13542.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Leo Jay Margolin. 1946. *Paper Bullets: A Brief Story of Psychological Warfare in World War II.* New York: Froben Press.

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 Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 4826–4832, Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization.

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*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic Signatures for Example-Based Linguistic Metaphor Detection. In *Proceedings of the First Workshop on Metaphor in {NLP}*, pages 27–35, Atlanta, Georgia. Association for Computational Linguistics.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online). International Committee for Computational Linguistics.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor Identification in Large Texts Corpora. *PLoS ONE*, 8(4):e62343.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vinodkumar Prabhakaran, Marek Rei, and Ekaterina Shutova. 2021. How Metaphors Impact Political Discourse: A Large-Scale Topic-Agnostic Study Using Neural Metaphor Detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:503–512.

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 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1537–1546, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Anup Shah. 2005. War, propaganda and the media. *Global Issues*, 31.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the 2016 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. VU amsterdam metaphor corpus. Oxford Text Archive.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A Reading Comprehension Paradigm for Token-level Metaphor Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring Sensorial Features for Metaphor Identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors We Think With: The Role of Metaphor in Reasoning. *PLoS ONE*, 6(2):e16782.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-Lingual Metaphor Detection Using Common Semantic Features. In *Proceedings of the First Workshop on Metaphor in {NLP}*, pages 45–51, Atlanta, Georgia. Association for Computational Linguistics.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor : A Computational Perspective.* Number Vol. 31 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*, 27:1–107.

Yorick Wilks, Adam Dalton, James Allen, and Lucian Galescu. 2013. Automatic Metaphor Detection using Large-Scale Lexical Resources and Conventional Metaphor Extraction. In *Proceedings of the First Workshop on Metaphor in {NLP}*, pages 36–44. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. THU NGN at NAACL-2018 Metaphor Shared Task: Neural Metaphor Detecting with CNN-LSTM Model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.

A Appendix

A.1 Persuasion techniques

The following list compiles the descriptions of propaganda techniques present in the PTC corpus (Da San Martino et al., 2019b) and the dataset used by SemEval-2021 task 6 (Dimitrov et al., 2021).

1. Appeal to authority: stating the validity of a claim because an expert or authority has issued it without providing any other evidence. The datasets include *Testimonials* as part of this technique, although they might not refer to an expert or authority.

2. Appeal to fear/prejudice: building support for an idea by provoking anxiety/panic to the alternative. In some instances, it leverages prejudices to obtain the desired response.

3. Bandwagon: invites the target audience to support an idea or action with the pretext that "*everyone is doing the same*".

4. Black-and-white Fallacy (Dictatorship): introduces two alternatives as the only possible options to weaken or strengthen one of them. In the extreme, it morphs into *dictatorship* when the choice is made for the audience, and all other options are considered impossible.

5. Causal Oversimplification: assuming a single cause for an issue when there might be many factors at play in reality. The data also includes *scapegoating* in this category - moving the blame to a person or group without considering the issue's complexities.

6. Doubt: questioning the credibility of something or someone.

7. Exaggeration/Minimisation: representing something as more extreme/dramatic than it is or, conversely, downplaying its significance.

8. Flag-waving: rally around a solid national sentiment to justify an action or idea.

9. Glittering generalities (Virtue) [6]: words or symbols that produce a positive image of the propagandist object by association with the preferences of the target audience.

10. Loaded Language: the use of emotionally charged words to influence an audience. It often exploits stereotypes and vagueness.

11. Name-calling: referring to the object of the propagandist campaign with a label that connects the target audience with an emotion, either positive (love, praise) or negative (fear, hate).

12. Obfuscation, Intentional Vagueness, Confusion: deliberately use unclear statements forcing the audience to produce their interpretation.

13. Red Herring: presenting irrelevant data to divert attention away from the discussed issue.

14. Reductio ad Hitlerum: seek disapproval of a position by suggesting that it is popular with a group the target audience hates.

15. Repetition: repeating the same message to subdue the audience into acceptance.

16. Slogans: brief and memorable motto or phrase to persuade the audience.

17. Smears [6]: effort to damage or question someone's reputation by propounding negative propaganda.

18. Straw Man: misrepresentation of someone's position to disprove it leaving the original argument unaddressed.

19. Thought-terminating cliché: using expressions to prevent critical thinking and meaningful discussions.

20. Whataboutism: replying with a counter-question or counter-accusation that suggests the rival is hypocritical concerning their position without refuting their argument.

[6] Only present for propaganda in memes, not for propaganda in the news dataset

A.2 Evaluation metrics for propaganda

To evaluate the model's performance in identifying propagandist instances, we follow the methods used by preceding works. The authors of the PTC corpus (Da San Martino et al., 2019b) propose precision and recall metrics based on the overlaps between the target and predicted spans. These metrics are then used to calculate the F1 score for each technique and all techniques combined.

Should document ***d*** be a sequence of characters, we can represent a propaganda technique span by $t = [t_i, ..., t_j] \subseteq \boldsymbol{d}$. This ground truth will be compared against the predicted model outputs $s = [s_i, ..., s_j]$. The labeling function $l(x)$ will return the propaganda technique associated with the fragment x. The function $\delta(l_a, l_b)$ will return 1 when l_a equals l_b and 0 otherwise. The groups T and S denote the group of propagandist fragments for gold labels and predictions respectively. Equation 1 calculates the overlapping number of character between two spans and divides it by a given length h.

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

In turn, Equation 2 reuses C to calculate the precision metric as the average proportion of correct prediction spans. Conversely, the Equation 3 defines recall as the average proportion of ground truth fragments covered by the predicted spans. Both metrics are similar, but while precision uses the number and length of the predictions, recall uses the gold label spans instead.

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

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

In contrast, precision and recall metrics for metaphor are calculated as a binary classification task at the word level. Only content words are considered for this task.

A.3 Significance testing

To check the statistical significance of our results, we use statistical bootstrapping (Efron, 1979). This powerful non-parametric method is recommended for evaluation metrics such as precision, recall, and F-score in NLP tasks (Dror et al., 2018). The main idea is to assess whether differences in performance between two models originate from variability in the data rather than from the superiority of one model over the other.

First, we create 100 different bootstrap samples ($B_{1..100}$) from the test data (T) by sampling with replacement (i.e., an example can appear multiple times within the same sample while others might not be present at all). Our examples are either individual sentences from news articles or the textual information of a meme, depending on the dataset used for the task. Each bootstrap sample has the same size as the test set ($|T| = |B_n| \forall n \in \{1, 100\}$). The premise is that being the test set a representative sample from all possible data for the task; we can get a sense of the variability of the task's data by comparing performance across multiple bootstrap samples.

After randomly generating the samples, we performed a paired bootstrap test as suggested in Dror et al. (2018). We calculate the *p-value* as the proportion of bootstrap samples where one type of model outperforms another. Since we use ten different seeds for each setup, comparing results between single-task and multi-task learning strategies requires calculating the mean across multiple models' performances. We start by calculating the performance of all models of a particular type by averaging their scores on each bootstrap sample. We do this for single-task and multi-task models. Then, we count the number of times one strategy achieves higher performance than the other. Finally, we calculate the *p-value* as the proportion of samples where that strategy was superior. We use the standard confidence level of 95% ($\alpha = 0.05$).

A.4 Reproducibility

We adapted our source code [7] to achieve reproducible results. First, we enabled the use of deterministic algorithms in the *PyTorch* framework. Next, we manually set the seed for all packages involved in random number generation. We use natural numbers for the seeds starting at one and up to the number of runs for each set of hyperparameters tested. Finally, we pined the versions for all dependencies.

The system we used had the following software: Python/3.8.2, GCCcore/9.3.0, CUDA/11.2, cuDNN/8.2.1.32. Additionally, we assigned the value ":4096:8" to the environment vari-

[7] https://github.com/baleato/paper-bullets

Task	Duration STL	MTL
News		
Multi-label	1:31:14	1:31:01
Name Calling	38:53	52:59
Loaded Language	37:06	53:47
Memes		
Multi-label	12:38	36:48
Name Calling	6:33	23:59
Loaded Language	15:06	22:36

Table 6: Average training runtime per task. This includes models discarded in hyper-parameter search trials.

able "CUBLAS_WORKSPACE_CONFIG" as suggested by Nvidia documentation [8] to avoid non-deterministic behavior. Our models used a single GeForce 1080Ti GPU for training. The average training runtime per task is shown in Appendix Table 6.

A.5 Preprocessing

The PTC dataset used NLTK sentence splitter[9] to break news articles into individual sentences. We detected duplicates driven primarily by boilerplate content regarding site functionality (e.g., invitation to participate in an online poll or request to subscribe to their newsletter). Duplicates were mainly short sentences that did not include any labels. We removed these instances from the training set.

We observed that the text in 454 examples (47% of the data) for the memes dataset was upper-cased. Since our model is case-sensitive, we true-cased all instances to minimize the number of out-of-vocabulary words by the tokenizer.

[8] https://docs.nvidia.com/cuda/cublas/index.html

[9] https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt

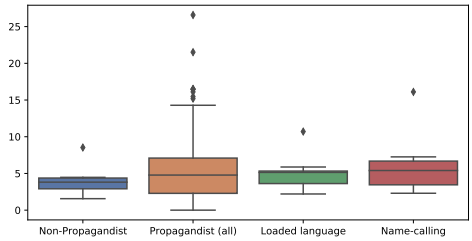

Figure 2: Percentages of metaphorical open-class words predicted by multi-label MTL models in news articles (test set).

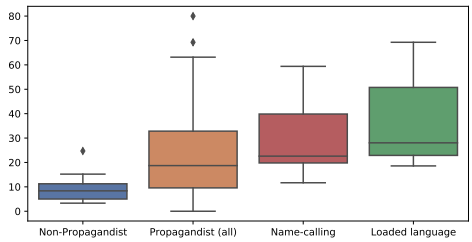

Figure 3: Percentages of metaphorical open-class words as predicted by multi-label MTL models in social memes (test set).

Model	F1 score
Multi-label (news)	66.77 ± 0.61
Name Calling (news)	62.95 ± 2.05
Loaded Language (news)	61.42 ± 1.64
Multi-label (memes)	64.09 ± 4.63
Name Calling (memes)	56.23 ± 4.68*
Loaded Language (memes)	63.05 ± 5.73

Table 7: Metaphor F1 score performance for multi-task models. *The first five runs had a median of 60.79; however, adding five extra seeds brought it down to 56.23.

Model / Propagandist Technique	P	R	F1
Single-task learning			
- Appeal to Authority	**7.58**	**1.14**	**1.67** ± 0.80
- Appeal to fear-prejudice	23.66	**26.87**	24.02 ± 2.67
- Bandwagon	0.00	0.00	0.00 ± 0.00
- Black-and-White Fallacy	8.16	**14.97**	**10.20** ± 2.48
- Causal Oversimplification	4.11	8.02	5.14 ± 1.91
- Doubt	7.02	20.89	10.01 ± 1.42
- Exaggeration, Minimisation	18.16	**24.25**	20.16 ± 2.12
- Flag-Waving	31.66	**49.44**	37.83 ± 3.44
- Loaded Language	28.26	**47.64**	34.79 ± 3.16
- Name Calling	23.74	**37.70**	28.58 ± 1.63
- Obfuscation, Vagueness, Confusion	0.00	0.00	0.00 ± 0.00
- Red Herring	**2.26**	1.33	**1.65** ± 3.32
- Reductio ad hitlerum	**16.81**	**17.73**	**15.88** ± 4.26
- Repetition	9.82	**6.75**	**7.23** ± 1.76
- Slogans	31.20	**32.68**	31.42 ± 1.95
- Straw Men	0.00	0.00	0.00 ± 0.00
- Thought-terminating Cliches	3.82	11.07	5.36 ± 2.69
- Whataboutism	**13.42**	**4.73**	5.56 ± 3.67
Multi-task learning			
- Appeal to Authority	5.57	0.95	1.50 ± 1.02
- Appeal to fear-prejudice	**24.58**	24.71	**24.44** ± 2.13
- Bandwagon	0.00	0.00	0.00 ± 0.00
- Black-and-White Fallacy	**9.05**	11.47	9.72 ± 3.44
- Causal Oversimplification	**5.39**	**8.42**	**6.35** ± 1.88
- Doubt	**7.88**	**15.46**	**10.31** ± 1.16
- Exaggeration, Minimisation	**19.82**	21.13	**20.29** ± 1.06
- Flag-Waving	**34.16**	46.40	**39.16** ± 1.76
- Loaded Language	**29.51**	43.40	**34.96** ± 0.95
- Name Calling	**25.73**	34.44	**29.36** ± 1.52
- Obfuscation, Vagueness, Confusion	0.00	0.00	0.00 ± 0.00
- Red Herring	1.61	**2.00**	1.59 ± 2.77
- Reductio ad hitlerum	11.78	16.85	13.59 ± 3.71
- Repetition	**10.03**	4.54	6.12 ± 1.98
- Slogans	**35.35**	31.94	**32.74** ± 5.45
- Straw Men	0.00	0.00	0.00 ± 0.00
- Thought-terminating Cliches	**6.18**	**13.57**	**8.32** ± 3.72
- Whataboutism	12.12	3.99	**5.89** ± 3.62

Table 8: Performance on propaganda technique identification in news articles by multi-label models on every technique. The highest performance for each metric is in bold.

Model / Propagandist Technique	P	R	F1
Single-task learning			
- Appeal to authority	**61.76**	45.29	**49.67** ± 9.13
- Appeal to fear/prejudice	**16.17**	**6.82**	**9.05** ± 7.45
- Bandwagon	0.00	0.00	0.00 ± 0.00
- Black-and-white Fallacy	**67.70**	**30.70**	**41.55** ± 4.01
- Causal Oversimplification	12.20	8.70	8.61 ± 7.63
- Doubt	**45.90**	13.95	20.98 ± 6.92
- Exaggeration/Minimisation	**44.71**	35.90	**39.44** ± 3.43
- Flag-waving	**52.13**	35.18	40.88 ± 8.91
- Glittering generalities (Virtue)	**33.83**	6.74	9.92 ± 7.09
- Loaded Language	**60.48**	**68.93**	**64.39** ± 1.95
- Name calling	**52.90**	55.88	54.21 ± 3.23
- Obfuscation, Vagueness, Confusion	0.00	0.00	0.00 ± 0.00
- Red Herring	0.00	0.00	0.00 ± 0.00
- Reductio ad hitlerum	0.00	0.00	0.00 ± 0.00
- Repetition	0.00	0.00	0.00 ± 0.00
- Slogans	**32.74**	25.58	27.75 ± 6.00
- Smears	30.57	**37.19**	**33.13** ± 2.43
- Straw Man	0.00	0.00	0.00 ± 0.00
- Thought-terminating cliché	**23.33**	**10.16**	**13.43** ± 10.76
- Whataboutism	21.69	**26.25**	**22.98** ± 6.57
Multi-task learning			
- Appeal to authority	51.53	**50.93**	49.52 ± 6.90
- Appeal to fear/prejudice	11.81	6.80	7.96 ± 5.79
- Bandwagon	0.00	0.00	0.00 ± 0.00
- Black-and-white Fallacy	42.31	28.17	33.23 ± 5.65
- Causal Oversimplification	**13.55**	**19.10**	**12.74** ± 7.71
- Doubt	43.46	**22.48**	**28.60** ± 6.31
- Exaggeration/Minimisation	37.45	**39.07**	37.40 ± 5.86
- Flag-waving	45.99	**52.85**	**48.17** ± 6.56
- Glittering generalities (Virtue)	32.43	**11.59**	**16.07** ± 6.80
- Loaded Language	56.68	68.82	61.51 ± 2.90
- Name calling	52.49	**57.49**	**54.50** ± 1.87
- Obfuscation, Vagueness, Confusion	0.00	0.00	0.00 ± 0.00
- Red Herring	0.00	0.00	0.00 ± 0.00
- Reductio ad hitlerum	0.00	0.00	0.00 ± 0.00
- Repetition	**10.42**	**12.50**	**11.25** ± 19.65
- Slogans	32.49	**27.27**	**28.87** ± 5.63
- Smears	**31.49**	34.70	32.28 ± 3.52
- Straw Man	0.00	0.00	0.00 ± 0.00
- Thought-terminating cliché	12.87	5.42	6.72 ± 6.86
- Whataboutism	**23.49**	25.89	22.49 ± 6.66

Table 9: Performance on propaganda technique identification in memes by multi-label models on every technique. The highest performance for each metric is in bold.

Dataset / Model	P	R	F1
News			
- STL Multi-label	24.92	29.26	26.18 ± 1.87
- MTL Multi-label	26.41	26.59	26.39 ± 0.78
- STL Loaded Language	39.03	50.23	43.80 ± 0.77
- MTL Loaded Language	40.20	48.40	43.70 ± 0.99
- STL Name-Calling	30.32	38.29	33.67 ± 0.88
- MTL Name-Calling	30.96	38.68	34.08 ± 0.62
Memes			
- STL Multi-label	57.84	54.83	56.23 ± 0.79
- MTL Multi-label	54.75	56.68	55.59 ± 1.20
- STL Loaded Language	77.13	73.07	74.84 ± 1.67
- MTL Loaded Language	71.89	75.62	73.64 ± 1.77
- STL Name-Calling	71.71	69.33	70.32 ± 1.91
- MTL Name-Calling	74.56	71.66	72.88 ± 1.65

Table 10: Performance on the validation set for propaganda technique identification.

Task	Parameter	Values
All	dropout	0.0
	LR scheduler	cosine
	warmup	10%
	weight decay	0.01
News - Multi label	batch size	8, 16, **32**
	learning rate	1e-5, 3e-5, 4e-5, **5e-5**
	max epochs	35
	patience	7
	task sampling ratio ∗	(1/6, 5/6), **(1/5, 4/5)**, (1/4, 3/4), (1/3, 2/3), (1/2, 1/2), (2/3, 1/3)
	epoch factor ∗	(0.95, 1.0), (0.96, 1.0), (0.97, 1.0), (0.98, 1.0), (0.99, 1.0), **(1.0, 1.0)**
News - Name calling	batch size	16, **32**
	learning rate	5e-6, **1e-5**, 3e-5, 5e-5
	task sampling ratio ∗	(1/4, 3/4), (1/3, 2/3), **(1/2, 1/2)**
	epoch factor ∗	(0.99, 1.0), **(1.0, 1.0)**
News - Loaded language	batch size	16, **32**
	learning rate	5e-6, **1e-5**, 3e-5, 5e-5
	task sampling ratio ∗	(1/4, 3/4), **(1/3, 2/3)**, (1/2, 1/2)
	epoch factor ∗	(0.99, 1.0), **(1.0, 1.0)**
Memes - Multi label	batch size	**8**, 16, 32
	learning rate	1e-5, 3e-5, 4e-5, **5e-5**
	max epochs	150
	patience	50
	epoch factor ∗	(0.98, 1.0), **(0.99, 1.0)**, (0.995, 1.0) (1.0, 1.0)
	task sampling ratio ∗	(1/4, 3/4), (1/3, 2/3), **(1/2, 1/2)**, (6/10, 4/10), (7/10, 3/10)
	loss scaling ∗	(3/4, 1), **(1, 1)**
Memes - Name calling	batch size	8, 16, **32**
	learning rate	1e-5, 3e-5, **4e-5**, 5e-5
	task sampling ratio ∗	(1/5, 4/5), **(1/4, 3/4)**, (1/3, 2/3), (1/2, 1/2), (6/10, 4/10)
	epoch factor ∗	(0.98, 1), (0.99, 1.0), (0.995, 1) **(1.0, 1.0)**
	loss scaling ∗	(3/4, 1), **(1, 1)**, (5/4, 1) (1, 5/4), (1, 3/2)
Memes - Loaded language	batch size	8, **16**, 32
	learning rate	1e-5, 2e-5, 3e-5, **4e-5**, 5e-5
	task sampling ratio ∗	(1/5, 4/5), (1/4, 3/4), (1/3, 2/3) (1/2, 1/2), (6/10, 4/10), **(7/10, 3/10)** (4/5, 1/5)
	epoch factor ∗	(0.98, 1.0), **(0.99, 1.0)**, (0.995, 1.0) (1.0, 1.0)
	loss scaling ∗	(3/4, 1), **(1, 1)**

Table 11: Best performance parameters after five runs are in bold. Multi-task parameters are identified with an asterisk, and their values belong to the auxiliary and main tasks.

Technique	Example
Appeal to authority	"... information released by investigative reporter **Laura Loomer proves that authorities have directly lied to the American people** about the case at least once ...
Appeal to fear	"... students told her daughter that she was **going to hell**.
Bandwagon	"... the likelihood that this disease will move to other more densely populated regions of the planet has become a **huge concern for many.**
Black-and-white Fallacy	**Either you stand with BDS, Hamas, blood libels and those who want to destroy Israel or with Jews.**
Causal Oversimplification	On the other hand, it knows that **by seeking continued secrecy, it's essentially an implicit acknowledgment of guilt.**
Doubt	What happened during the 6 minutes between Campos being shot and Paddock opening fire, and **why weren't the police rushing to the scene immediately?**
Exaggeration/Minimisation	Whatever definition that one might put on that nebulous term, **no reasonable person can honestly believe** that the release of 50-year-old records are going to **result in the United States falling into the ocean** or even that the communists are going to take over the federal government.
Flag-waving	"I want to get **our soldiers** out. I want to bring **our soldiers** back home," Trump said.
Glittering generalities	"... to show the **enormous, enthusiastic crowd in front of him**.
Loaded Language	On both of their blogs the pair called their bans from entering the UK "**a striking blow against freedom**" and said the "**the nation that gave the world the Magna Carta is dead**".
Straw Man	His opinion is: "Take it seriously, but with a large grain of salt." **Which is just Allen's more nuanced way of saying: "Don't believe it."**
Name-calling	"It's embarrassing for this **so-called land of democracy and freedom of speech,**" he said.
Obfuscation	Accordingly, he rushed to the defense of Bergoglio and his corrupt regime against "**a radicalization of religious conservatism in the neo-traditionalism sense**...
Red Herring	"The jury of six men and six women, **including three immigrants**, found the Mexican national not guilty ...
Reductio ad Hitlerum	Exactly what this "special need" is that can constitute a **Gestapo like police state** surveilling its own citizens is a moving target that has already been proven to be abused over and over again.
Repetition	**Take notice**, Dutch Prime Minister Rutte. **Take notice**, Mrs. Merkel or President Macron. **Take notice**: the future is ours and not yours
Slogans	**Christianity is Europe's last hope**.
Smears	**No honor, no integrity, no principles, no morals**, ...
Thought-terminating cliché	This whole idea of a two-state solution, **it doesn't work**.
Whataboutism	"They interpreted the law in my case to say it was criminal," Saucier told Fox News, referring to prosecuting authorities in his case, "**but they didn't prosecute Hillary Clinton**.

Table 12: Examples of persuasion techniques are in bold.