TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations

Prashanth Vijayaraghavan
IBM Research
Almaden Lab
San Jose, CA
prashanthv@ibm.com

Soroush Vosoughi
Dartmouth College
Department of Computer Science
Hanover, NH
soroush@dartmouth.edu

Abstract

Recently, several studies on propaganda detection have involved document and fragment-level analyses of news articles. However, there are significant data and modeling challenges dealing with fine-grained detection of propaganda on social media. In this work, we present TWEETSPIN, a dataset containing tweets that are weakly annotated with different fine-grained propaganda techniques, and propose a neural approach to detect and categorize propaganda tweets across those fine-grained categories. These categories include specific rhetorical and psychological techniques, ranging from leveraging emotions to using logical fallacies. Our model relies on multi-view representations of the input tweet data to (a) extract different aspects of the input text including the context, entities, their relationships, and external knowledge; (b) model their mutual interplay; and (c) effectively speed up the learning process by requiring fewer training examples. Our method allows for representation enrichment leading to better detection and categorization of propaganda on social media. We verify the effectiveness of our proposed method on TWEETSPIN and further probe how the implicit relations between the views impact the performance. Our experiments show that our model is able to outperform several benchmark methods and transfer the knowledge to relatively low-resource news domains.

1 Introduction

Propaganda refers to any idea or information, that is often false or exaggerated, and is used to promote or publicize a particular cause or point of view. In recent years, there has been a surge in research and development of methods to detect propaganda from text. For example, some of the earlier works like (Rashkin et al., 2017a) and (Barrón-Cedeno et al., 2019) released a corpus of news articles containing coarse-grained document-level annotation of propaganda. Da San Martino et al. (2019b) described a corpus of news articles containing annotations of 18 fine-grained propaganda techniques. Following this work, two subtasks were presented as a part of NLP4IF workshop (Da San Martino et al., 2019a) that focused on the identification of propagandist text units at fragment and sentence level.

We believe that the challenge for this task lies in the varied nature of propaganda techniques, including cognitive and information distortion and logical fallacies. The challenges are further exacerbated in the social media setting, which has become a key battleground in the spread of propaganda. Research on propaganda detection in social media platforms like Twitter has been limited by the: (a) lack of sufficiently annotated social media propaganda data, (b) idiosyncratic nature of the content on social media, (c) difficulty in modeling the social context in which propaganda is disseminated, and (d) varying propaganda techniques that require factual knowledge, structural relationships, and reasoning abilities.

In this work, we address a subset of both the data and modeling challenges. First, we introduce TWEETSPIN, a corpus of tweets containing weak labels of fine-grained propaganda techniques. We accomplish this through a data collection pipeline that incorporates keyword-based search and users calling out propaganda techniques publicly on Twitter (see Figure 1 for examples). Next, we present a transformer-based multi-view propaganda detection model, MV-PROP, that identifies varied aspects of the textual data using multi-view contextual embeddings and captures their interaction via pairwise cross-view transformers. The main contributions of this work are described as follows:
(1) Creation of the TWEETSPIN corpus containing weak annotations of fine-grained propaganda techniques for tweets.
(2) An end-to-end Transformer-based MV-PROP model augmented with multiple views that infuse context, relational information and external knowl-

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3433 - 3448

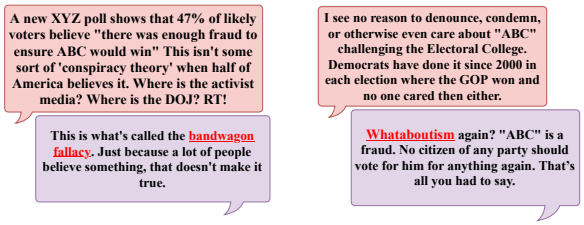

Figure 1: Data from TWEETSPIN where users are being called out for using specific propaganda techniques.

edge into the representation and capture their pairwise interactions through cross-view transformers. To the best of our knowledge, this is the first study that incorporates such multi-view representations for the propaganda detection task.

(3) We conduct experiments using TWEETSPIN to demonstrate the capability of our model in detecting fine-grained propaganda techniques for tweets. We also demonstrate the transferability of our model to data from the news domain.

2 Related Work

This work is closely related to a broad spectrum of topics, including offensive language detection, computational argumentation, and fake news detection. Some of the propaganda techniques might overlap with the categories of argumentation techniques (e.g., strawman argument) or offensive language (e.g., name-calling). However, there are considerable variations in propaganda techniques across different social contexts. Thus, methods used for the detection of offensive language or argumentation techniques may not be directly transferable to fine-grained propaganda detection. In this section, we review the prior researches in two main areas that are relevant to our work – (a) propaganda analyses and detection and (b) multi-view representation learning.

Propaganda Analyses and Detection Research in propaganda detection has primarily focused on document-level analysis. A work by Rashkin et al. (Rashkin et al., 2017b) constructed a corpus of news stories from eight different sources and labeled them with four broad categories – propaganda, hoax, trusted, and satire. Barron et al. (Barrón-Cedeno et al., 2019) addressed some of the limitations of the previous work by obtaining more data using distant supervision. They explored a range of features such as keywords, rich representations, writing styles, and readability level to discern propaganda text from other forms of news stories by casting it as a binary classification problem. More recently, there has been a dedicated line of research aimed at identifying fine-grained propaganda methods at the fragment level (Da San Martino et al., 2019b,a). This involved a manually annotated corpus flagging specific text spans in news articles as containing one of 18 propaganda techniques beyond the binarized setting used earlier. Various text classification models (Da San Martino et al., 2019a; Alhindi et al., 2019) incorporating TF-IDF features (Li et al., 2019) and contextual representations (e.g. BERT (Zellers et al., 2019), RoBERTa and ELMo (Cruz et al., 2019)) have been proposed to handle this fragment-level task. To detect propaganda on social media, many studies (Williamson III and Scrofani, 2019; Caldarelli et al., 2020) utilize datasets that contain propaganda content spread on social media by Russian-based IRA (Farkas and Bastos, 2018; Miller, 2019) or extremists (Johnston and Weiss, 2017; Nizzoli et al., 2019). Subsequent works investigated the influence of propaganda on public opinion (Caldarelli et al., 2020) and the techniques applied to disseminate targeted political agenda (Gorrell et al., 2019). Another recent work (Wang et al., 2020b) leveraged cross-domain learning approach to label propagandistic content. Wang et al. (2020b) implement different classifiers using different informative features and constraints based on labeled documents and sentences from news and tweets to detect propaganda within and across domains. Most of these works either apply bot or troll detection techniques or apply feature engineering methods to conduct binary classification of propagandistic content.

Multi-View Representation Learning Multi-view representation learning has numerous applications involving images, texts, graphs or videos. A line of work in computer vision has extensively studied the benefits of multi-view representation in embedding social images (Gong et al., 2014), object detection (Chen et al., 2017), viewpoint classification (Su et al., 2009), shape/face recognition (Chen et al., 2017; Su et al., 2015; Li et al., 2016), to list a few. However, there has been a limited exploration of multi-view representation explicitly for texts. For instance, the widely used Seq2Seq with attention module (Bahdanau et al., 2014), used in several state-of-the-art NLP tasks, can be seen from the the perspective of multi-view fusion where information from different time-steps are fused and encoded together into a semantic representation.

More recently, a work by (Bian et al., 2020) focused on learning to match a resume with a relevant job using multi-view representation learning approach.

Unlike previous studies that either focused on news articles or individual social media posts, our work performs a fine-grained analysis of propaganda for tweets considering the importance of the discussion context. Furthermore, we collect a weakly-annotated corpus of tweets associated with different propaganda techniques and develop a multi-view learning approach that has not yet to the best of our knowledge been explored for propaganda detection.

3 Problem Definition

Given an input tweet text along with a discussion context, our goal is to predict if the input text exhibits any propaganda techniques. Determining the propaganda technique of the input tweet can be formulated as a multi-class classification problem. In this work, we denote the input tweet text as $T_i = [w_1, w_2, .., w_{N_i}]$, where N_i is the sequence length of the input text. Each tweet text may be accompanied by a context $C_i = [T_1, T_2, ..., T_{i-1}]$, referring to M prior tweets in the discussion thread related to the input tweet T_i. Every input text might not necessarily contain a context (in such a case, $M = 0$). In this task, the target is to develop a model that can learn a mapping function $f : T \mapsto p_k$, where $k \in 1, 2, ..., L$, $p_k \in P$ indicates one of the L propaganda labels and p_1 denotes the special case of non-propaganda category or absence of any propaganda technique.

4 TWEETSPIN

We construct a dataset of English-language tweets, referred to as the TWEETSPIN corpus, containing weak annotations of 18 propaganda techniques same as in (Da San Martino et al., 2019b). Table B1 shows the full statistics of our dataset.

Our data collection pipeline consists of three components: (a) propaganda keyword expansion, (b) keyword-based tweet retrieval, (c) tweet filtration and (d) data augmentation.

Propaganda Keyword Expansion: For each propaganda technique, we select the name of the technique as the initial keyword. This works well for techniques that cannot be easily characterized by specific lexical patterns. Typical examples of such techniques include red herring, obfuscation, causal oversimplification, and strawmen. Additionally, we expand the list of keywords related to some of the propaganda techniques by combining information from publicly available resources including technique-specific phrases, idioms, and examples listed in Table B2.

Keyword-based Tweet Retrieval: We use Twitter's standard search API to ingest tweets based on the keywords identified from the previous step. We enclose the keywords or phrases within quotation marks. We observed that many Twitter users explicitly call out any usage of specific propaganda techniques in their discussions on the platform. Therefore, we select quoted tweets and replies containing keywords related to the propaganda techniques. Additionally, we search for tweets containing both the word "you" along with the identified keywords to capture instances where the users call out the usage of propaganda techniques.

Tweet Filtration: Given the tweets retrieved based on the keywords, we remove tweets that are extremely short (less than 5 words) and those that are replies to tweets from deleted or protected accounts. For the remaining tweets, we collect the discussion thread for each tweet which provide the context for the tweet. Though the tweet threads may involve complex tree structures with different reply branches, we are only interested in the specific branch of the tree that contains the tweet being called out for using specific propaganda techniques. In tweet threads that are long (>50 tweets in a discussion thread), not all tweets in the discussion context might be important for the classification task. Thus, we apply temporal filtering on such discussion threads, where we only retain the source tweet and discussion context that falls in a 7-day window before a particular tweet was being called out as propagandistic.

Data Augmentation: We adopt data augmentation strategies to handle potential data scarcity and also address the problem of overfitting. We employ linguistically informed transformations of text to prevent meaning distortion leading to new misclassification errors (Li et al., 2020). Hence, we randomly select 10% of examples from each class from our training set and perform linguistically informed augmentations using the code from Li et al. (2020).

Manual Validation of TWEETSPIN We randomly sampled 1,000 samples from the TWEETSPIN corpus containing the tweet to be classified, the prior

context, and anonymized user information indicating if the text in the discussion context is from the same user or a different user. Three MTurk workers annotated each of these samples with one of the 19 propaganda techniques (18 techniques + 1 for non-propaganda). The definitions of the propaganda techniques were made available to the MTurk workers for reference. The inter-rater reliability as calculated using Fleiss' κ was 0.85, indicating a substantial agreement between the annotators. The agreement between the labels in our corpus and the labels provided by the annotators (through majority agreement) was 89.3%, indicating the relative high fidelity of our corpus, especially given its reliance on a weak-annotation scheme.

5 MV-PROP: Multi-View Propaganda Detection

Here, we describe our proposed model for fine-grained propaganda detection on social media. Unlike the fragment-level classification task used for propaganda detection in news articles, we formulate this problem as a text classification task where we aim to map the input tweet text to one of the several propaganda labels conditioning on the discussion thread context if it exists (as explained in Section 3). Figure 2 illustrates the overview of our model architecture. Inspired from the literature in multi-modal learning (Tsai et al., 2019), we propose a transformer-based multi-view propaganda detection model, MV-PROP, that integrates multi-view contextual embeddings via pairwise cross-view transformers. The main motivation behind such a modeling choice comes from the fact that different propaganda techniques require focus on varying aspects of the data. For example, propaganda techniques like loaded language can be identified from the usage of specific words or phrases, while repetition or red herring necessitates a contextual understanding of the tweet. Similarly, most propaganda techniques can benefit from word sense disambiguation and entity information that can be accumulated from external knowledge sources. Thus, the fine-grained differences between these propaganda techniques call for multi-view representations that can unravel such variations in the data. Our MV-PROP model comprises the following components:

Multi-View Encoding Layer, which computes multi-view representation from input tweet T_i and context C_i.

Cross-View Transformer, that reinforces representations obtained from a specific view with those computed from another view. We compute this for all pairs of such cross-view transformers.

Classification Layer, which fuses the embeddings from the previous step and computes the likelihood of the input tweet with the given context belonging to a particular propaganda label.

5.1 Multi-View Encoding Layer

In this work, we compute three different views: (a) context-aware semantic view, which derives semantic representation from text and context by leveraging a pre-trained language model, (b) relationship structure view, that calculates a relational representation by applying relation-based graph neural network on dependency graph and speaker-dependent context graph, and (c) knowledge-enriched view, which enriches the input tweet embedding with different kinds of knowledge. We discuss them in detail in subsequent sections.

5.1.1 Context-Aware Semantic View

The accompanying discussion context of a tweet can significantly shift how the tweet is perceived and hence plays a critical role in determining the propaganda technique used in the given tweet. Therefore, we employ a hierarchical incremental transformer encoder to obtain a context-aware semantic representation of the tweet text. Inspired from some of the existing hierarchical approaches (Zhang et al., 2019; Liu and Lapata, 2019), we implement a two-level transformer encoding process: (a) a tweet encoder f_T that operates at the word-level to transform each tweet into an embedding including those in the discussion thread and (b) a context encoder f_C that enriches the input tweet representation by capturing the influence of the previous tweets in the discussion thread relevant to our classification task. Thus, we learn a context-aware semantic view of the given input tweet.

We utilize a BERT-based pretrained language model (Devlin et al., 2018) as our tweet encoder, f_T. Each tweet $T_j = [w_1, w_2, ...w_{N_j}]$ is fed to the wordpiece tokenization algorithm. Here N_j is the number of words in the tweet text. We add special tokens $[CLS]$ and $[SEP]$ at the start and end of the tokenized tweet token list. We feed each tweet text into the BERT model and produce contextual word embedding as:

$$\hat{H}_j = [\hat{h}_j^1, \hat{h}_j^2, ..., \hat{h}_j^{N_j}] = f_T(T_j) = BERT(T_j) \quad (1)$$

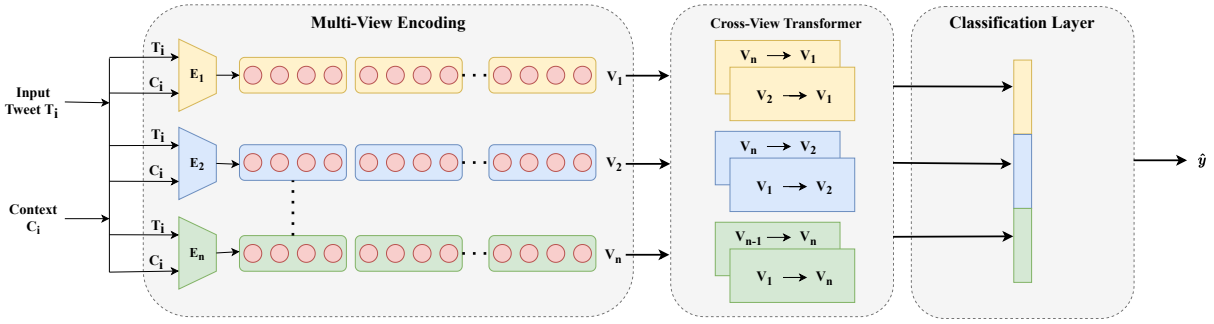

Figure 2: Illustration of our MV-PROP model.

We obtain an overall tweet embedding by performing a maxpool operation on the contextual word embeddings intended primarily to retain the important information in each dimension:

$$h_j^s = maxpool(\hat{H}_j) \tag{2}$$

The choice of maxpool was made after considering other options like average pooling and $[CLS]$ token representation. We found the maxpool worked best for our classification task.

Given an input tweet T_i and its prior context $\{T_k\}_{k=1}^{i-1}$, we account for the sequence information in the discussion context by performing an element-wise summation of the tweet embedding h_j^s with the positional embedding p_k. We compute the position incorporated tweet context embeddings as:

$$C^{(<i)} = [h_1^s, h_2^s, ..., h_{i-1}^s] \odot [p_1, p_2..., p_{i-1}] \tag{3}$$

Since our goal is to detect the propaganda technique used in the input tweet by conditioning on the prior context, we enrich the input tweet embedding using the information from the computed context embeddings $C^{(<i)}$. This is done by feeding the input tweet and the context representations to a context encoder, f_C, comprising N_C transformer encoding layers. However, we introduce an additional context-attention sub-layer in the transformer layer that integrates discussion context into the encoder. This is implemented as:

$$\begin{aligned} U^{(l)} &= \textsc{Mhatt}(\hat{H}_i^{(l-1)}, \hat{H}_i^{(l-1)}, \hat{H}_i^{(l-1)}) \\ V^{(l)} &= \textsc{Mhatt}(U^{(l)}, C^{(<i)}, C^{(<i)}) \\ H_i^{(l)} &= FFN(V^{(l)}) \\ H_{cas} &= H_i^{(N_C)} \end{aligned} \tag{4}$$

where l refers to the l^{th} context encoding layer, $l \in 1, 2, ..., N_C$, MHATT and FFN refer to the multi-head attention step and feed-forward network in each transformer encoding layer and $H_i^1 = \hat{H}_i$, $C^{(<t)}$ is the prior discussion context embedding as computed in Equation 3 and $H_i^{(l)}$ is the embedding of the input tweet at the l^{th} layer. The output from the N_C-th layer is the final context-aware semantic view H_{cas} of the input tweet T_i.

5.1.2 Relationship Structure View

The goal of this view is to compute a hierarchical relational embedding that captures two main aspects: (a) dependency graph-based structural information from individual tweets and (c) speaker-dependent structural relationships from the discussion context. First, we intuit that a representation that encapsulates the syntactic structures explicitly and learns relationships between specific words and phrases can better guide the propaganda detection model. We explain the reason as follows. Certain words can express an attitude or sentiment towards specific key terms or entities in the sentence. Despite the advantages of using flat attention-based models, the limitations of assigning higher attention scores to irrelevant words or wrong associations can lead to performance degradation. Additionally, we differentiate between the input tweet user's previous tweets and the other users' tweets in the discussion context. We believe that self-dependency (relationship between input tweet user's previous tweets in the context) and inter-speaker dependency (relationship between input tweet with other users' tweets in the context) can be critical to understanding the speaker motivation (or intention) or attitudinal/sentiment shifts in the conversation. Therefore, both these aspects require the extraction of some form of structural relationship. Recently, graph neural networks (Scarselli et al., 2008; Schlichtkrull et al., 2018; Kipf and Welling, 2016; Veličković et al., 2017) have been applied to tackle challenges in effective representation of nodes from graph-structured data and have proven effective in a number of NLP applications

such as aspect-level sentiment analysis (Huang and Carley, 2019; Wang et al., 2020a), reading comprehension (Tu et al., 2019; Zhang, 2020; Song et al., 2018) and relation extraction (Fu et al., 2019; Zhang et al., 2018). However, most of the models like graph-convolution networks (GCN) (Kipf and Welling, 2016) or graph attention networks (GAT) (Veličković et al., 2017) operate on homogeneous links or edges. In our work, we introduce a hierarchical relation-based graph neural network that can aggregate incoming information from neighbors depending on the edge type.

First, we perform coreference resolution[1] given the discussion context tweets and replace the mentions with proper entity information. Next, We apply a dependency parser from (Kong et al., 2014; Liu et al., 2018) to transform a tweet text into a dependency parse graph $\mathcal{G}$. This graph is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ where a node refers to a word in the tweet text ($v_k \in V$) and a labeled edge indicating a dependency relation between two words $(v_k, r, v_m) \in \mathcal{E}$, where $r \in \mathcal{R}$ is a relation type. Each tweet T_j is converted into a graph $\mathcal{G}_j$. Since the traditional GAT model fails to consider different relation types into consideration, there is a significant loss of crucial dependency information. Following some of the prior work on relation-based graphical propagation of information (Busbridge et al., 2019; Veličković et al., 2017; Ishiwatari et al., 2020), we aggregate embeddings of the k^{th} node, $\bar{h}_k^{(l)}$, using varied relation-specific influences of the relation-specific neighborhood nodes $\mathcal{N}_k^r$ computed using an attention mechanism. Stacking N_S layers allows information from nodes N_S-hops away to propagate to a particular node. Therefore, we implement a relational graph attention network (R-GAT) that intuitively aggregates the incoming information from different relations with varying influences. These steps are defined by:

$$\begin{aligned} \bar{h}_k^{(l)} &= \textstyle\sum_{r=1}^{R} \bar{h}_{kr}^{(l-1)} \\ \bar{h}_{kr}^{(l-1)} &= \textstyle\sum_{m \in \mathcal{N}_k^r} \beta_{kj}^{r(l-1)} W_r^{(l-1)} \bar{h}_m^{(l-1)} \\ \beta_{km}^{r(l-1)} &= attention(\bar{h}_k^{(l-1)}, \bar{h}_m^{(l-1)}) \\ \bar{H}_j &= [\bar{h}_1^{N_S}, \bar{h}_2^{N_S}, ..., \bar{h}_{N_j}^{N_S}] \end{aligned} \tag{5}$$

where $\beta_{kj}^{r(l-1)}$ denotes the normalized attention coefficient calculated using dot-product mechanism for the node k based on its neighborhood node m under relation type r. Additionally, we compute multi-head attention and concatenate its outputs.

[1] https://spacy.io/universe/project/neuralcoref

We also incorporate relational position embeddings as in (Ishiwatari et al., 2020). At $l = 1$, $\bar{h}_j^1$ is assigned to the contextual word embeddings obtained in Equation 1. Similar to Equation 2, we obtain a structural information enriched tweet embedding h_j^{rel} for a tweet T_j using maxpool operation on $\bar{H}_j$.

For the input tweet T_i, we differentiate discussion context tweets into two types: (i) T^A: tweets that are produced by the same user as the input tweet and (ii) T^B: tweets that are produced by all users other than the input tweet user. Using the computed tweet-level structural embedding h_j^{rel}, we construct a graph to account for speaker-dependent structural relationships (An example is shown in Figure B1). This includes two labeled edges indicating how the input tweet T_i is influenced by T^A (self-dependency) and T^B (inter-speaker dependency). Once the graph is constructed, we run R-GAT as in equations 5. Finally, we obtain relationship structure view of the input tweet T_i as:

$$H_{rel} = \bar{H}_i^{(N_R)} \tag{6}$$

where N_R refers to the number of layers in the speaker-dependent relationship extraction layer.

5.1.3 Knowledge-Enriched View

Some of the propaganda techniques involve distortion of facts and data to promote their cause or point of view. Thus, we intuit that models which infuse external knowledge could potentially improve the overall performance in our task. Therefore, the primary aim of the knowledge-enriched view is to compute a text representation by enriching them with different kinds of knowledge. We leverage K-ADAPTER$(F + L)$ (Wang et al., 2020c) that combines both factual knowledge and linguistic knowledge to derive the knowledge-enriched view. K-ADAPTER acquires factual knowledge from the relationships among entities in text by training on a large scale alignment dataset between Wikipedia abstracts and Wikipedia triples. We modify the input by concatenating the context C_i and input tweet T_i: "<SEP> context</SEP>input tweet</SEP>" and use the embedding of the first token to get the knowledge-enriched view. This is given as:

$$\widetilde{H}_{know} = \text{K-ADAPTER}(T_i, C_i) \tag{7}$$

5.2 Cross-View Transformer

First, we fuse the cross-view information between any two views A and B using an additional multi-attention sub-layer as in Equation 4. We denote the three views as S, R, K. Next, we introduce a

transformer encoder layer on top of the cross-view transformer layer. Finally, we perform a maxpool operation on the output from the previous transformer layer. These steps are defined by:

$$\begin{aligned} H_{R\mapsto S} &= \text{MH-ATT}(Query_S, Key_R, Value_R) \\ H_{K\mapsto S} &= \text{MH-ATT}(Query_S, Key_K, Value_K) \\ z_S &= maxpool(Transformer([H_{R\mapsto S}; H_{K\mapsto S}])) \end{aligned} \tag{8}$$

Similarly, we compute z_R, z_K and feed the concatenated outputs to the final classification layer.

5.3 Classification Layer

We calculate the probability that the input tweet uses a particular propaganda technique using a softmax layer, where f_p is a fully-connected layer, and z is the embedding produced by concatenating the outputs from the previous layer as:

$$z = z_S||z_R||z_K q = softmax(f_p(z)) \tag{9}$$

6 Training & Implementation Details

We optimize the categorical cross-entropy loss between the predicted and true propaganda labels as in Equation 10, where L is the total number of propaganda labels, q_j is the predicted distribution that the input tweet falls under propaganda technique j, and $p_j \in \{0, 1\}$ denotes the ground-truth of whether the input tweet can be categorized under the j^{th} propaganda label.

$$L_{CE} = \sum_{j=1}^{L} -p_j log(q_j) \tag{10}$$

We use the publicly released default pre-trained model parameters for the BERT variants used. We perform a grid-search and optimize the hyperparameters using the validation set: $N_C = 3, N_S = N_R = 2$. We used Adam with a learning rate of α=2e-5 and a warmup proportion of 0.1 for optimization. To account for randomness, we report the numbers which are the mean of five experimental runs with different random seeds. To alleviate the problem of unbalanced datasets, we utilize class weights in categorical cross-entropy loss based on the training and validation sets. See Appendix A for details on the hardware.

7 Experiments

Our experiments are designed to investigate the following research questions:
RQ1: How well does our MV-PROP model perform compared to the other baselines in the propaganda detection task on social media data?
RQ2: What are the influences of different views and their interactions on the overall performance?
RQ3: Can our model be applied to detect propaganda on in-domain and cross-domain datasets?

7.1 Dataset

We run experiments using TWEETSPIN dataset containing 210,392 tweets labeled with 19 propaganda types (referring to 18 propaganda techniques and 1 non-propaganda label). Due to the imbalance of the TWEETSPIN dataset, we divide our TWEETSPIN dataset into training (70%), validation (10%), and test (20%) sets using a stratified shuffle split[2].

7.2 Baselines

We use the following baselines in our experiments:
BERT FT (Devlin et al., 2018) is a fine-tuned version of BERT_{base} model on the input tweets with and without considering the discussion context.
ROBERTA FT (Liu et al., 2019) is a fine-tuned version of ROBERTA_{base} model on the input tweets conditioning on the discussion context tweets.
LATEXPRO (Wang et al., 2020d) leverages the declarative knowledge expressed in both first-order logic and text. We reimplement a variant of this model without the token-level loss to suit the sentence-level classification task. We further investigate the importance of the discussion context.

7.3 Model Variants

We investigate the importance of different modeling components by introducing variants to our proposed model and evaluating their performance on the TWEETSPIN validation set. These variants assess the influence of critical aspects: (a) discussion context, (b) different views, and (c) different fusion techniques. Depending on the fusion technique, we replace the cross-view transformer with simple late fusion techniques involving concatenation, mean, and sum of embeddings obtained from multiple views. View-specific variants include:
MV-PROP, which refers to our full model comprising all the three views as shown in Section 5.
MV-PROP-K, which integrates semantic and relational views while removing the knowledge-enriched view from our model.
MV-PROP-R, which combines semantic and knowledge-enriched views while removing the relationship structure view from our model.

[2]In this paper, we performed the stratified shuffle split using Python's Scikit-learn module

Models	P	R	F1	Std.
BERT FT	28.18	27.52	27.85	5.23
BERT FT w/ ctx	34.86	31.19	32.92	4.79
ROBERTA FT w/ ctx	35.16	32.20	33.72	3.27
LATEXPRO	32.27	29.40	30.77	3.96
LATEXPRO w/ ctx	43.76	34.56	38.62	3.24
MV-PROP	**68.48**	**59.62**	**63.74**	1.06

Table 1: Evaluation results on the TWEETSPIN test set. Std. refers to the standard deviation of the F1-scores across five runs.

Model	P	R	F1
Views			
MV-PROP - K	59.18	50.71	54.62
MV-PROP - R	62.95	49.09	55.16
MV-PROP - S	61.54	52.38	56.60
Fusion			
Concat	63.86	54.80	59.56
Mean	64.02	53.34	58.19
Sum	53.91	53.09	57.99
MV-PROP	**70.17**	**59.63**	**64.47**

Table 2: Ablation study on the TWEETSPIN validation set. We observe that the performance degrades when a specific view is removed or the cross-view transformer is replaced with other fusion techniques.

MV-PROP-S, which computes knowledge and relational views while removing the context-aware semantic view from our model.

7.4 Results

We report the precision, recall, micro-averaged F1 scores and standard deviation of the computed F1 scores across five runs. in Table 1. Notably, the context plays a critical role in determining the propaganda label. This is evident from an average $\sim 17.8\%$ drop in F1 without the context information in the baseline methods. We also find that the context information reduces the sensitivity of the models as indicated by a diminished standard deviation value whenever context comes into play. The results in Table 1 also demonstrate the ability of our multi-view representations to surpass the other baseline models by a large margin.

7.4.1 Effect of Multi-view Representations

In addition to experiments that emphasize the importance of the context in Table 1, we study the necessity of each view by discarding one view at a time and reporting the relative impact on the performance. It is clear from Table 2 that removal of each view leads to a significant drop in performance with

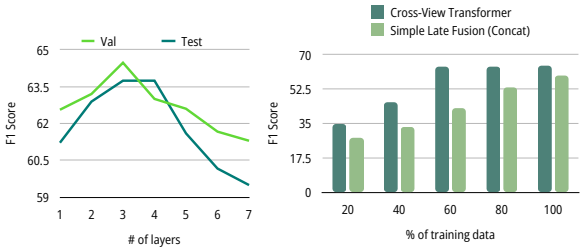

Figure 3: Analysis of cross-view transformer on TWEETSPIN. Left: Effect of number of cross-view transformer layers on performance. Right: Impact of cross-view transformers over simple late fusion method (concatenation) by varying the proportion of training samples.

knowledge-enriched view diminishing the performance by $\sim 14\%$. Since K-Adapter model injects additional knowledge into a pre-trained language model, it retains the benefits of the semantic representation obtained from the language model and also additionally incorporates both factual knowledge and entity information into the embedding. Therefore, discarding this component leads to a significant loss of information. This is the same reason why the drop in performance is relatively smaller when the context-aware semantic view is removed. The embedding from the knowledge-enriched view partially compensates for the information loss when the semantic view is discarded. We also highlight that the relationship structure view has a noticeable effect on the recall. We show sample tweets that were misclassified by our variants compared to our full model in Table B3 to further illustrate the importance of our views.

7.4.2 Effect of Cross-View Transformer

To study the effect of the cross-view transformer, we analyze the following: (a) effect of number of cross-view transformer layers, (b) impact of cross-view transformers over simple late fusion methods such as concat, mean, and sum of embeddings obtained from different views. Figure 3 (left) shows model performance with varying number of layers. The performance improves initially with the increase in number of layers and then drops beyond a point. The optimal number of layers on the validation and test datasets is 3. Moreover, Table 2 reports the performance for model variants involving simple late fusion techniques instead of the cross-view transformers. Visibly, the best performing late fusion technique (concat) lags behind the full MV-PROP model containing cross-view transformer layers. With a performance drop of $\sim 8\%$, it is evident that the cross-view transformer efficiently

Models	F1
LR*	81.7
SVM*	79.5
LSTM*	80.7
LSTMR*	78.9
MV-PROP	84.36

Table 3: Evaluations results on the TWE dataset. * indicates the scores reported in the original paper (Wang et al., 2020b).

computes the interaction between multi-view representations. Further, we vary the proportion of the training data and evaluate the potential impact of the cross-view transformer. We observe that our model with the cross-view transformer allows for quicker learning than the simple concatenation-based late fusion strategy. Figure 3 (right) shows that our full MV-PROP model plateaus closer to the best F1 score with $\sim 60\%$ of the training data.

7.5 Performance on Unseen In-Domain dataset

Wang et al. (2020b) constructed a TWE dataset by combining two pre-existing datasets, the Twitter IRA corpus (Edgett, 2017) and the "twitter7" data from SNAP (Yang and Leskovec, 2011) as propagandistic and non-propagandistic data respectively. However, this dataset doesn't contain the discussion context that is critical to exploit the full potential of our MV-PROP model. Table 3 shows evaluations on the TWE dataset. Though certain views like the relationship structure view may not be utilized to their capabilities due to the lack of the discussion context, our MV-PROP model is able to significantly outperform the baselines used in the original work. Notably, our trained MV-PROP model performs well on the unseen in-domain dataset.

7.6 Performance on Cross-Domain Dataset

We hypothesize that our full model trained on the TWEETSPIN dataset is transferable to a cross-domain dataset like news articles. To verify this, we conduct an experiment on the Propaganda Techniques Corpus (PTC) (Da San Martino et al., 2019b), which is a manually annotated dataset for propaganda detection. Given that our model detects propaganda at the tweet level, we perform the sentence level propaganda detection (SLC) task from Da San Martino et al. (2019b) that determines whether a given sentence from a news article is propagandistic. We train a zero-shot (MV-PROP (ZS)) variant of our model that directly takes the input from the PTC dataset and outputs the likelihood of it being propagandistic. With a threshold of 0.6, our model performs comparably to the fine-tuned BERT model, showing that our model demonstrates transfer capability to a similar task in the news domain. Further, we fine-tune our MV-PROP model (MV-PROP (FT)) using the PTC training set and observe that we outperform other benchmarks for the SLC task. Results are shown in Table 4.

Models	P	R	F1
Random	30.48	51.04	38.16
All-Propaganda	30.54	100.00	46.80
Fine-tuned BERT*[1]	63.20	53.16	57.74
BERT-Joint*[1]	62.84	55.46	58.91
MGN*[1]	60.41	61.58	60.98
Proper Gander*[2]	56.50	70.10	62.56
LatexPRO*[3] (L)	56.53	73.17	63.79
LatexPRO*[3] (L+T)	59.04	71.66	64.74
MV-PROP (ZS)	54.08	62.75	58.09
MV-PROP (FT)	**64.35**	**84.58**	**73.09**

Table 4: Evaluation results on the test set of PTC dataset for the sentence level propaganda classification (SLC) task.* refers to the scores reported from their original work. 1 refers to (Da San Martino et al., 2019b), 2 refers to (Madabushi et al., 2020), 3 refers to (Wang et al., 2020d). ZS refers to zero-shot and FS to fine-tuned variants of our model. All-Propaganda model always classifies the input text as propagandstic.

8 Conclusion

We introduced TWEETSPIN, a corpus of tweets containing weak labels of fine-grained propaganda techniques. Next, we presented a transformer-based multi-view propaganda detection model, MV-PROP, that integrates multi-view contextual embeddings via pairwise cross-view transformers. We demonstrate how the semantic, relational, and knowledge view enrichment of the input tweet text leads to significant performance improvement over other baseline methods. Our experiments also demonstrated the transferability of our trained model to propaganda detection for news articles. The main limitation of our work is the reliance on weak annotations of Twitter data, which is unavoidable given the scale of our dataset. Future work could investigate leveraging the multi-view representations for span-level detection of fine-grained propaganda techniques.

9 Ethics Statement

Social media has become a battleground for propaganda and influence campaigns. This paper is an attempt to provide a dataset and models for detecting various propaganda techniques on Twitter to aid with the fight against this scourge on society. We release TWEETSPIN, a Twitter corpus containing weak-annotations of fine-grained propaganda. Consistent with Twitter TOS, TWEETSPIN contains only tweet IDs (with code provided to hydrate them) and no identifying information. Given the nature of the task the dataset contains potentially offensive and hateful language which should be taken into consideration. Additionally it is possible that our models, analyses, and dataset can potentially be used to create more advanced and harder to detect propaganda techniques. Though we should be aware of this possibility it is imperative that we in the research community stay ahead of miscreants by actively pushing this field forward. Finally, our models can lead to false positives where a user is falsely accused of spreading propaganda. Thus, it is important that the techniques presented here be used as a part of a larger effort to combat propaganda with humans in the loop for checks and balances.

TWEETSPIN was validated using MTurk. The annotators were paid 0.08 USD per task which took on average 30 seconds, for an hourly rate of 9.6 USD, above the federal and our state's minimum wage.

References

Tariq Alhindi, Jonas Pfeiffer, and Smaranda Muresan. 2019. Fine-tuned neural models for propaganda detection at the sentence and fragment levels. *arXiv preprint arXiv:1910.09702*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

Shuqing Bian, Xu Chen, Wayne Xin Zhao, Kun Zhou, Yupeng Hou, Yang Song, Tao Zhang, and Ji-Rong Wen. 2020. Learning to match jobs with resumes from sparse interaction data using multi-view co-teaching network. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 65–74.

Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. 2019. Relational graph attention networks. *arXiv preprint arXiv:1904.05811*.

Guido Caldarelli, Rocco De Nicola, Fabio Del Vigna, Marinella Petrocchi, and Fabio Saracco. 2020. The role of bot squads in the political propaganda on twitter. *Communications Physics*, 3(1):1–15.

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915.

André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2019. On sentence representations for propaganda detection: From handcrafted features to word embeddings. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 107–112.

Giovanni Da San Martino, Alberto Barron-Cedeno, 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.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, 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 5636–5646.

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

Sean Edgett. 2017. Testimony of sean j. edgett. *United States Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism*.

Johan Farkas and Marco Bastos. 2018. Ira propaganda on twitter: Stoking antagonism and tweeting local news. In *Proceedings of the 9th International Conference on social media and society*, pages 281–285.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233.

Genevieve Gorrell, Mehmet E Bakir, Ian Roberts, Mark A Greenwood, Benedetta Iavarone, and Kalina Bontcheva. 2019. Partisanship, propaganda and post-truth politics: Quantifying impact in online debate. *arXiv preprint arXiv:1902.01752*.

Binxuan Huang and Kathleen M Carley. 2019. Syntax-aware aspect level sentiment classification with graph attention networks. *arXiv preprint arXiv:1909.02606*.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Andrew H Johnston and Gary M Weiss. 2017. Identifying sunni extremist propaganda with deep learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–6. IEEE.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012.

Chuanrong Li, Lin Shengshuo, Leo Z Liu, Xinyi Wu, Xuhui Zhou, and Shane Steinert-Threlkeld. 2020. Linguistically-informed transformations (lit): A method for automatically generating contrast sets. *arXiv preprint arXiv:2010.08580*.

Jianshu Li, Jian Zhao, Fang Zhao, Hao Liu, Jing Li, Shengmei Shen, Jiashi Feng, and Terence Sim. 2016. Robust face recognition with deep multi-view representation learning. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1068–1072.

Jinfen Li, Zhihao Ye, and Lu Xiao. 2019. Detection of propaganda using logistic regression. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 119–124.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A Smith. 2018. Parsing tweets into universal dependencies. *arXiv preprint arXiv:1804.08228*.

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 preprint arXiv:1907.11692*.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2020. Cost-sensitive bert for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*.

Daniel Taninecz Miller. 2019. Topics and emotions in russian twitter propaganda. *First Monday*.

Leonardo Nizzoli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2019. Extremist propaganda tweet classification with deep learning in realistic scenarios. In *Proceedings of the 10th ACM Conference on Web Science*, pages 203–204.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017a. 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.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017b. 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.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953.

Hao Su, Min Sun, Li Fei-Fei, and Silvio Savarese. 2009. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *2009 IEEE 12th International Conference on Computer Vision*, pages 213–220. IEEE.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020a. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362*.

Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. 2020b. Cross-domain learning for classifying propaganda in online contents. *arXiv preprint arXiv:2011.06844*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020c. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Ruize Wang, Duyu Tang, Nan Duan, Wanjun Zhong, Zhongyu Wei, Xuanjing Huang, Daxin Jiang, and Ming Zhou. 2020d. Leveraging declarative knowledge in text and first-order logic for fine-grained propaganda detection. *arXiv preprint arXiv:2004.14201*.

William Williamson III and James Scrofani. 2019. Trends in detection and characterization of propaganda bots. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Jaewon Yang and Jure Leskovec. 2011. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 177–186.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. *arXiv preprint arXiv:1905.06566*.

Xuanyu Zhang. 2020. Cfgnn: Cross flow graph neural networks for question answering on complex tables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9596–9603.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*.

A Experiment Platform

All the experiments were conducted on a Ubuntu 20.04 system, with 2.10GHz Intel(R) Xeon(R) CPU and 8-core NVIDIA GeForce GTX 1080Ti/11GB. Our models were implemented using Pytorch 1.4 with CUDA 10.1.

B Additional Information and Examples

Table B1 shows the statistics of the TWEETSPIN dataset. Table B2 lists the common phrases related to different propaganda techniques used in our data collection process. Table B3 shows sample tweets that were misclassified by our model variants in comparison to our full MV-PROP model. Finally, Figure B1 illustrates the speaker dependency graph related to a sample tweet thread.

Dataset Statistics	
#Total Propaganda Tweets	157,327
#Total Non-Propaganda Tweets	53,165
% of tweets with discussion context	59.06
Avg. discussion context length	3.15
Avg. #users in discussion context	2.26
Propaganda technique	# Tweets
Loaded Language	18,365
Name Calling/Labeling	17,096
Reductio Ad Hitelerium	15,677
Doubt	14,993
Appeal To Fear/Prejudice	14,654
Whataboutism	13,887
Repetition	13,285
Slogans	10,190
Appeal To Authority	8,539
Flag-Waving	7,675
Exaggeration, Minimization	5,416
Black-And-White Fallacy	4,872
Thought-terminating cliches	3,781
Bandwagon	2,547
Red Herring	2,315
Causal oversimplification	1,790
Straw man	1,265
O, I, C	1,048

Table B1: TWEETSPIN Dataset statistics . O, I, C refers to "Obfuscation, Intentional Vagueness, Confusion".

Propaganda Techniques	Common Phrases
Loaded Language	List of words/phrases: https://examples.yourdictionary.com/loaded-language-examples.html
Name Calling	"Commie", "Fascist", "Pig", "Yuppie", "Libtard", "Extremist", "Terrorist", "Snowflake", "Cuck"
Appeal to Authority	"Experts have warned", "Experts say ... ", "As an expert in ... ", "As a[n] [occupation*], I can say ....", "[PERSON] advises/ urges/ suggests" *https://learnersdictionary.com/3000-words/topic/jobs-professions
Doubt	"Lied to us", "Lying to us", "covering up", "cover up", "not being told the truth", "not adding up", "official story", "fake story"
Bandwagon	"Almost all/ Most/ Majority of [Nation or ethnic groups*]" *https://en.wikipedia.org/wiki/Lists_of_people_by_nationality
Flag Waving	"[Nation/State] first", "Nation" + [Positive Word*], "Anti-[Nation/State]", "True patriots/nationalist" *https://ptrckprry.com/course/ssd/data/positive-words.txt
Reductio Ad Hitlerum	"Hitler, Stalin", "[PERSON] is a communist/marxist/nazi/fascist", "[ORG] are communists/marxists/nazis/fascists"
Black & White Fallacy	"No other way...", "No alternative to ...", "No other option ...", "no better way"
Whataboutism	"The media ignores", "Nobody talks/mentions/speaks about ...", "But What about", "Don't focus on ..., but ..."

Table B2: List of common phrases related to different propaganda techniques used in the data collection. Note that some of the language used here is potentially offensive.

Tweet	True Label	Model	Predicted Label
Context: A discussion about AR15 Guns for hunting And no one needs a BMW o a$1000 suit or a Rolex instead of a Timex. It's what one wants, not needs. Many people want an AR15...You remind me of an unelected bureaucrat in a position of power,...	Straw man	MV-PROP-S	[Loaded language, Whataboutism]
		MV-PROP	[Straw man, Whataboutism]
Context: A discussion about student loans You're right. Being able to sign away your life & go to war & die is definitely less that debt.	Red Herring	MV-PROP - R	[Straw man, Exaggeration]
		MV-PROP	[Red Herring, Straw man]
Context: A discussion about Presidential Elections 2020 This is 'Democracy', Venezuela-style. Or Cuba. Or China. Or the Soviet Union. Or a certain Central European country in the 1930's.	Reductio Ad Hitlerum	MV-PROP - K	[Non Propaganda, Red Herring]
		MV-PROP	[Reductio Ad Hitlerum, Straw man]

Table B3: Sample tweets which were misclassified by our model variants in comparison to our full MV-PROP model. We report the top 2 ranked predictions.

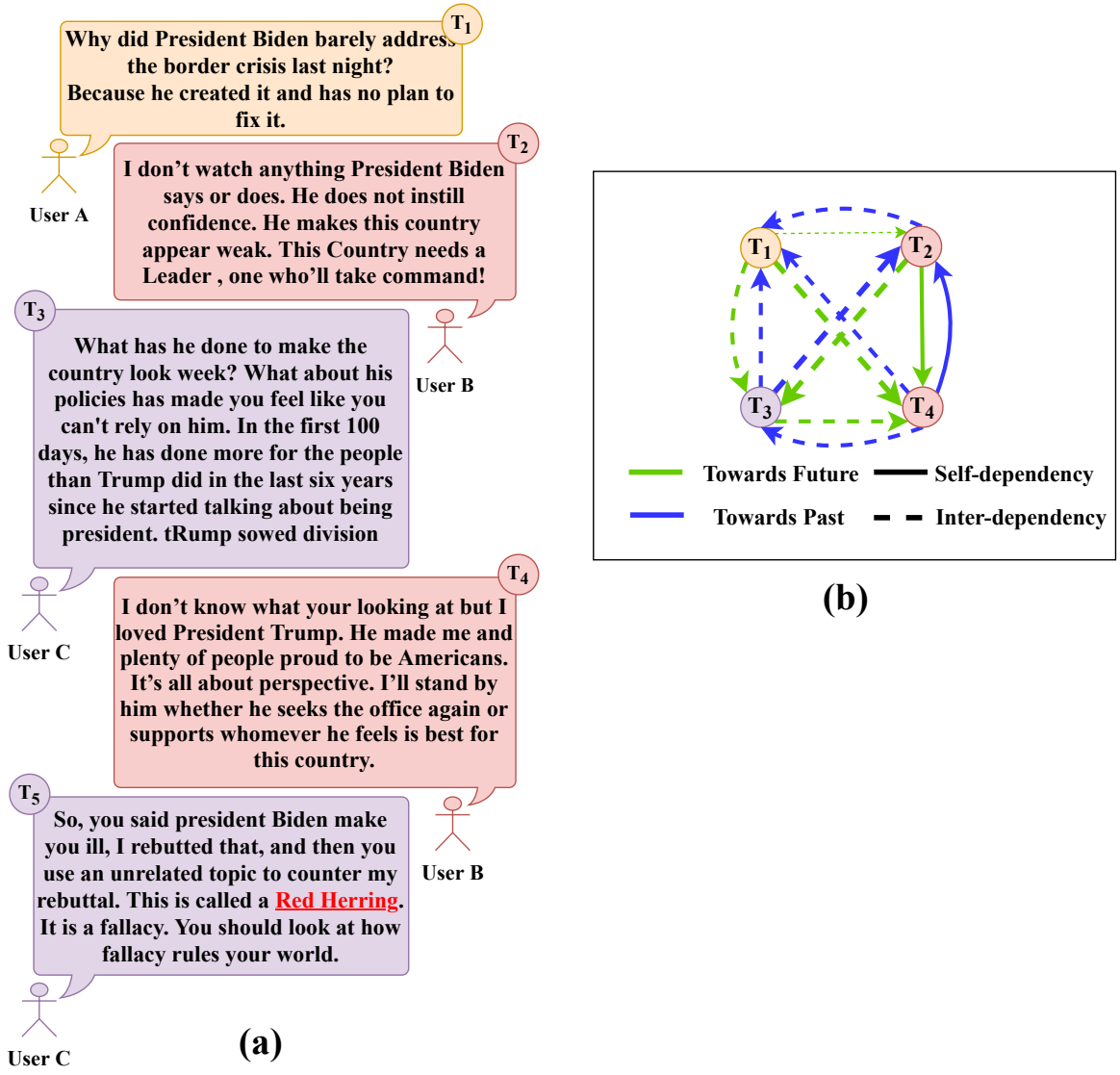

Figure B1: (a) Sample tweet thread where speaker dependencies can determine the propaganda technique; (b) Speaker-dependent structural relationship for the discussion thread in (a).