Fine-Grained Analysis of Propaganda in News Articles

Giovanni Da San Martino\textsuperscript{1}  Seunghak Yu\textsuperscript{2}  Alberto Barrón-Cedeño\textsuperscript{3}  Rostislav Petrov\textsuperscript{4}  Preslav Nakov\textsuperscript{1}

\textsuperscript{1} Qatar Computing Research Institute, HBKU, Qatar  
\textsuperscript{2} MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA  
\textsuperscript{3} Università di Bologna, Forlì, Italy,  
\textsuperscript{4} A Data Pro, Sofia, Bulgaria

\{gmartino, pnakov\}@hbku.edu.qa
seunghak@csail.mit.edu,a.barron@unibo.it
rostislav.petrov@adata.pro

Abstract

Propaganda aims at influencing people’s mindset with the purpose of advancing a specific agenda. Previous work has addressed propaganda detection at the document level, typically labelling all articles from a propagandistic news outlet as propaganda. Such noisy gold labels inevitably affect the quality of any learning system trained on them. A further issue with most existing systems is the lack of explainability. To overcome these limitations, we propose a novel task: performing fine-grained analysis of texts by detecting all fragments that contain propaganda techniques as well as their type. In particular, we create a corpus of news articles manually annotated at the fragment level with eighteen propaganda techniques and we propose a suitable evaluation measure. We further design a novel multi-granularity neural network, and we show that it outperforms several strong BERT-based baselines.

1 Introduction

Research on detecting propaganda has focused primarily on articles (Barrón-Cedeno et al., 2019; Rashkin et al., 2017). In many cases, there are no labeled data for individual articles, but there are such labels for entire news outlets. Thus, often all articles from the same news outlet get labeled the way that this outlet is labeled. Yet, it has been observed that propagandistic sources could post objective non-propagandistic articles periodically to increase their credibility (Horne et al., 2018). Similarly, media generally recognized as objective might occasionally post articles that promote a particular editorial agenda and are thus propagandistic. Thus, it is clear that transferring the label of the news outlet to each of its articles, could introduce noise. Such labels can still be useful for training robust systems, but they cannot be used to get a fair assessment of a system at testing time.

One option to deal with the lack of labels for articles is to crowdsource the annotation. However, in preliminary experiments we observed that the average annotator cannot detach her personal mindset from the judgment of propaganda and bias, i.e., if a clearly propagandistic text expresses ideas aligned with the annotator’s beliefs, it is unlikely that she would judge it as such.

We argue that in order to study propaganda in a sound and reliable way, we need to rely on high-quality trusted professional annotations and it is best to do so at the fragment level, targeting specific techniques rather than using a label for an entire document or an entire news outlet.

Ours is the first work that goes at a fine-grained level: identifying specific instances of propaganda techniques used within an article. In particular, we create a corresponding corpus. For this purpose, we asked six experts to annotate articles from news outlets recognized as propagandistic and non-propagandistic, marking specific text spans with eighteen propaganda techniques. We also designed appropriate evaluation measures. Taken together, the annotated corpus and the evaluation measures represent the first manually-curated evaluation framework for the analysis of fine-grained propaganda. We release the corpus (350K tokens) as well as our code in order to enable future research.\textsuperscript{1}

Our contributions are as follows:

- We formulate a new problem: detect the use of specific propaganda techniques in text.
- We build a new large corpus for this problem.
- We propose a suitable evaluation measure.
- We design a novel multi-granularity neural network, and we show that it outperforms several strong BERT-based baselines.

\textsuperscript{1}The corpus, the evaluation measures, and the models are available at \url{http://propaganda.qcri.org/}
Our corpus could enable research in propagandistic and non-objective news, including the development of explainable AI systems. A system that can detect instances of use of specific propagandistic techniques would be able to make it explicit to the users why a given article was predicted to be propagandistic. It could also help train the users to spot the use of such techniques in the news.

The remainder of this paper is organized as follows: Section 2 presents the propagandistic techniques we focus on. Section 3 describes our corpus. Section 4 discusses an evaluation measures for comparing labeled fragments. Section 5 presents the formulation of the task and our proposed models. Section 6 describes our experiments and the evaluation results. Section 7 presents some relevant related work. Finally, Section 8 concludes and discusses future work.

2 Propaganda and its Techniques

Propaganda comes in many forms, but it can be recognized by its persuasive function, sizable target audience, the representation of a specific group’s agenda, and the use of faulty reasoning and/or emotional appeals (Miller, 1939). Since propaganda is conveyed through the use of a number of techniques, their detection allows for a deeper analysis at the paragraph and the sentence level that goes beyond a single document-level judgment on whether a text is propagandistic.

Whereas the definition of propaganda is widely accepted in the literature, the set of propaganda techniques differs between scholars (Torok, 2015). For instance, Miller (1939) considers seven techniques, whereas Weston (2018) lists at least 24, and Wikipedia discusses 69.\footnote{http://en.wikipedia.org/wiki/Propaganda_techniques; last visit May 2019.} The differences are mainly due to some authors ignoring some techniques, or using definitions that subsume the definition used by other authors. Below, we describe the propaganda techniques we consider: a curated list of eighteen items derived from the aforementioned studies. The list only includes techniques that can be found in journalistic articles and can be judged intrinsically, without the need to retrieve supporting information from external resources. For example, we do not include techniques such as card stacking (Jowett and O’Donnell, 2012, page 237), since it would require comparing against external sources of information.

The eighteen techniques we consider are as follows (cf. Table 1 for examples):

1. **Loaded language.** Using words/phrases with strong emotional implications (positive or negative) to influence an audience (Weston, 2018, p. 6). Ex.: “[... ] a lone lawmaker’s childish shouting.”

2. **Name calling or labeling.** Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises (Miller, 1939). Ex.: “Republican congressweasels”, “Bush the Lesser.”

3. **Repetition.** Repeating the same message over and over again, so that the audience will eventually accept it (Torok, 2015; Miller, 1939).

4. **Exaggeration or minimization.** Either representing something in an excessive manner: making things larger, better, worse (e.g., “the best of the best”, “quality guaranteed”) or making something seem less important or smaller than it actually is (Jowett and O’Donnell, 2012, p. 303), e.g., saying that an insult was just a joke. Ex.: “Democrats bolted as soon as Trump’s speech ended in an apparent effort to signal they can’t even stomach being in the same room as the president”; “I was not fighting with her; we were just playing.”

5. **Doubt.** Questioning the credibility of someone or something. Ex.: A candidate says about his opponent: “Is he ready to be the Mayor?”

6. **Appeal to fear/prejudice.** Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments. Ex.: “Stop those refugees; they are terrorists.”

7. **Flag-waving.** Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea (Hobbs and Mcgee, 2008). Ex.: “Entering this war will make us have a better future in our country.”

8. **Causal oversimplification.** Assuming one cause when there are multiple causes behind an issue. We include scapegoating as well: the transfer of the blame to one person or group of people without investigating the complexities of an issue. Ex.: “If France had not declared war on Germany, World War II would have never happened.”
9. **Slogans.** A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals (Dan, 2015). *Ex.:* “Make America great again!”

10. **Appeal to authority.** Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence (Goodwin, 2011). We include the special case where the reference is not an authority/expert, although it is referred to as testimonial in the literature (Jowett and O’Donnell, 2012, p. 237).

11. **Black-and-white fallacy, dictatorship.** Presenting two alternative options as the only possibilities, when in fact more possibilities exist (Torok, 2015). As an extreme case, telling the audience exactly what actions to take, eliminating any other possible choice (dictatorship). *Ex.:* “You must be a Republican or Democrat; you are not a Democrat. Therefore, you must be a Republican”; “There is no alternative to war.”

12. **Thought-terminating cliché.** Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought (Hunter, 2015, p. 78). *Ex.:* “it is what it is”; “you cannot judge it without experiencing it”; “it’s common sense”, “nothing is permanent except change”, “better late than never”; “mind your own business”; “nobody’s perfect”; “it doesn’t matter”; “you can’t change human nature.”

13. **Whataboutism.** Discredit an opponent’s position by charging them with hypocrisy without directly disproving their argument (Richter, 2017). For example, mentioning an event that discredits the opponent: “What about . . . ?” (Richter, 2017). *Ex.:* Russia Today had a proclivity for whataboutism in its coverage of the 2015 Baltimore and Ferguson protests in the US, which re-
revealed a consistent refrain: “the oppression of blacks in the US has become so unbearable that the eruption of violence was inevitable”, and that the US therefore lacks “the moral high ground to discuss human rights issues in countries like Russia and China.”

14. Reductio ad Hitlerum. Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation (Teninbaum, 2009). Ex.: “Only one kind of person can think this way: a communist.”

15. Red herring. Introducing irrelevant material to the issue being discussed, so that everyone’s attention is diverted away from the points made (Weston, 2018, p. 78). Those subjected to a red herring argument are led away from the issue that had been the focus of the discussion and urged to follow an observation or claim that may be associated with the original claim, but is not highly relevant to the issue in dispute (Teninbaum, 2009). Ex.: “You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son’s murderer to be fed and housed?”

16. Bandwagon. Attempting to persuade the target audience to join in and take the course of action because “everyone else is taking the same action” (Hobbs and Mcgee, 2008). Ex.: “Would you vote for Clinton as president? 57% say yes.”

17. Obfuscation, intentional vagueness, confusion. Using deliberately unclear words, so that the audience may have its own interpretation (Suprabandari, 2007; Weston, 2018, p. 8). For instance, when an unclear phrase with multiple possible meanings is used within the argument, and, therefore, it does not really support the conclusion. Ex.: “It is a good idea to listen to victims of theft. Therefore, if the victims say to have the thief shot, then you should do it.”

18. Straw man. When an opponent’s proposition is substituted with a similar one which is then refuted in place of the original (Walton, 1996). Weston (2018, p. 78) specifies the characteristics of the substituted proposition: “caricaturing an opposing view so that it is easy to refute.”
3.2 Manual Annotation

We aim at obtaining text fragments annotated with any of the 18 techniques described in Section 2 (see Figure 1 for an example). Since the time required to understand and memorize all the propaganda techniques is significant, this annotation task is not well-suited for crowdsourcing. We partnered instead with a company that performs professional annotations, A Data Pro.\(^4\) Appendix A shows details about the instructions and the tools provided to the annotators.

We computed the \(\gamma\) inter-annotator agreement (Mathet et al., 2015). We chose \(\gamma\) because (i) it is designed for tasks where both the span and its label are to be found and (ii) it can deal with overlaps in the annotations by the same annotator\(^5\) (e.g., instances of doubt often use name calling or loaded language to reinforce their message). We computed \(\gamma_s\), where we only consider the identified spans, regardless of the technique, and \(\gamma_{sl}\), where we consider both the spans and their labels.

Let \(a\) be an annotator. In a preliminary exercise, four annotators \(a_1 \ldots a_4\) annotated six articles independently, and the agreement was \(\gamma_s = 0.34\) and \(\gamma_{sl} = 0.31\). Even taking into account that \(\gamma\) is a pessimistic measure (Mathet et al., 2015), these values are low. Thus, we designed an annotation schema composed of two stages and involving two annotator teams, each of which covered about 220 documents. In stage 1, both \(a_1\) and \(a_2\) annotated the same documents independently. In stage 2, they gathered with a consolidator \(c_1\) to discuss all instances and to come up with a final annotation. Annotators \(a_3\) and \(a_4\) and consolidator \(c_2\) followed the same procedure. Annotating the full corpus took 395 man hours.

Table 4 shows the \(\gamma\) agreements on the full corpus. As in the preliminary annotation, the agreements for both teams are relatively low: 0.30 and 0.34 for span selection, and slightly lower when labeling is considered as well. After the annotators discussed with the consolidator on the disagreed cases, the \(\gamma\) values got much higher: up to 0.74 and 0.76 for each team. We further analyzed the annotations to determine the main cause for the disagreement by computing the percentage of instances spotted by one annotator only in the first stage that are retained as gold annotations.

Overall the percentage is 53\% (5,921 out of 11,122), and for each annotator is \(a_1 = 70\%\), \(a_2 = 48\%\), \(a_3 = 57\%\), \(a_4 = 31\%\). Observing such percentages together with the relatively low differences in Table 4 between \(\gamma_s\) and \(\gamma_{sl}\) for the same pairs \((a_1, a_2)\) and \((a_3, c_2)\), we can conclude that disagreements are in general not due to the two annotators assigning different labels to the same or mostly overlapping spans, but rather because one has missed an instance in the first stage.

<table>
<thead>
<tr>
<th>Annotations</th>
<th>spans ((\gamma_s))</th>
<th>+labels ((\gamma_{sl}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_1) (a_2)</td>
<td>0.30</td>
<td>0.24</td>
</tr>
<tr>
<td>(a_3) (a_4)</td>
<td>0.34</td>
<td>0.28</td>
</tr>
<tr>
<td>(a_1) (c_1)</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td>(a_2) (c_1)</td>
<td>0.74</td>
<td>0.72</td>
</tr>
<tr>
<td>(a_3) (c_2)</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>(a_4) (c_2)</td>
<td>0.42</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 4: \(\gamma\) inter-annotator agreement between annotators spotting spans alone \((\text{spans})\) and spotting spans+labeling \((\text{+labels})\). The top-2 rows refer to the first stage: agreement between annotators. The bottom 4 rows refer to the consolidation stage: agreement between each annotator and the final gold annotation.

Figure 1: Example given to the annotators.

3.3 Statistics about the Dataset

The total number of technique instances found in the articles, after the consolidation phase, is 7,485, with respect to a total number of 21,293 sentences (35.2\%). Table 5 reports some statistics about the annotations. The average propagandistic fragment has a length of 47 characters and the average length of a sentence is 112.5 characters.

On average, the propagandistic techniques are half a sentence long. The most common ones are loaded language and name calling, labeling with 2,547 and 1,294 occurrences, respectively. They appear 6.7 and 4.7 times per article, while no other technique appears more than twice. Note that repetition is inflated as we asked the annotators to mark both the original and the repeated instances.

\(^4\)http://www.aiidatapro.com
\(^5\)See (Meyer et al., 2014; Mathet et al., 2015) for other alternatives, which lack some properties; (ii) in particular.
4 Evaluation Measures

Our task is a sequence labeling one, with the following key characteristics: (i) a large number of techniques whose spans might overlap in the text, and (ii) large lengths of these spans. This requires an evaluation measure that gives credit for partial overlaps.\(^6\) We derive an ad hoc measure following related work on named entity recognition (NER) (Nadeau and Sekine, 2007) and (intrinsic) plagiarism detection (PD) (Potthast et al., 2010).

While in NER, the relevant fragments tend to be short multi-word strings, in PD—and in our propaganda technique identification task—the length varies widely (cf. Table 5), and instances span from single tokens to full sentences or even longer pieces of text. Thus, in our precision and recall versions, we give partial credit to imperfect matches at the character level, as in PD.

Let document \(d\) be represented as a sequence of characters. A propagandistic text fragment is then represented as \(t = [t_1, \ldots, t_j] \subseteq d\). A document includes a set of (possibly overlapping) fragments \(T\). Similarly, a learning algorithm produces a set \(S\) with fragments \(s = [s_m, \ldots, s_n]\), predicted on \(d\). A labeling function \(l(x) \in \{1, \ldots, 18\}\) associates \(s \in S\) to one of the eighteen techniques. Figure 2 gives examples of gold and predicted fragments.

---

\(^6\)The evaluation measures for the CoNLL’02 and CoNLL’03 NER tasks, where an instance is considered properly identified if and only if both the boundaries and the label are correct (Tsai et al., 2006), are not suitable in our context.

Table 5: Corpus statistics including instances per technique and their avg. length in terms of characters.

<table>
<thead>
<tr>
<th>Propaganda Technique</th>
<th>inst</th>
<th>avg. length</th>
</tr>
</thead>
<tbody>
<tr>
<td>loaded language</td>
<td>2,547</td>
<td>23.70 ± 25.30</td>
</tr>
<tr>
<td>name calling, labeling</td>
<td>1,294</td>
<td>26.10 ± 19.88</td>
</tr>
<tr>
<td>exaggeration, minimization</td>
<td>571</td>
<td>45.36 ± 35.55</td>
</tr>
<tr>
<td>doubt</td>
<td>562</td>
<td>123.21 ± 97.65</td>
</tr>
<tr>
<td>appeal to fear/prejudice</td>
<td>367</td>
<td>93.56 ± 74.59</td>
</tr>
<tr>
<td>flag-waving</td>
<td>330</td>
<td>61.88 ± 68.61</td>
</tr>
<tr>
<td>causal oversimplification</td>
<td>233</td>
<td>121.03 ± 71.66</td>
</tr>
<tr>
<td>slogans</td>
<td>172</td>
<td>25.30 ± 13.49</td>
</tr>
<tr>
<td>appeal to authority</td>
<td>169</td>
<td>131.23 ± 123.2</td>
</tr>
<tr>
<td>black-and-white fallacy</td>
<td>134</td>
<td>98.42 ± 73.66</td>
</tr>
<tr>
<td>thought-terminating cliches</td>
<td>95</td>
<td>34.85 ± 29.28</td>
</tr>
<tr>
<td>whataboutism</td>
<td>76</td>
<td>120.93 ± 69.62</td>
</tr>
<tr>
<td>reductio ad hitlerum</td>
<td>66</td>
<td>94.58 ± 64.16</td>
</tr>
<tr>
<td>red herring</td>
<td>48</td>
<td>63.79 ± 61.63</td>
</tr>
<tr>
<td>bandwagon</td>
<td>17</td>
<td>100.29 ± 97.05</td>
</tr>
<tr>
<td>obfusc., int. vagueness, confusion</td>
<td>17</td>
<td>107.88 ± 86.74</td>
</tr>
<tr>
<td>straw man</td>
<td>15</td>
<td>79.13 ± 50.72</td>
</tr>
<tr>
<td><strong>all</strong></td>
<td>7,485</td>
<td>46.99 ± 61.45</td>
</tr>
</tbody>
</table>

---

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

\[
C(s, t, h) = \frac{|(s \cap t)|}{h} \delta(I(s), I(t)),
\]

where \(h\) is a normalizing factor and \(\delta(a, b) = 1\) if \(a = b\), and 0 otherwise. In the future, \(\delta\) could be refined to account for custom distance functions between classes, e.g., we might consider mistaking **loaded language** for **name calling or labeling** less problematic than confusing it with **Reduction ad Hitlerum**. Given Eq. (1), we now define variants of precision and recall able to account for the imbalance in the corpus:

\[
P(S, T) = \frac{1}{|S|} \sum_{s \in S, t \in T} C(s, t, |s|),
\]

\[
R(S, T) = \frac{1}{|T|} \sum_{s \in S, t \in T} C(s, t, |t|),
\]

We define Eq. (2) to be zero if \(|S| = 0\) and Eq. (3) to be zero if \(|T| = 0\). Following Potthast et al. (2010), in Eqs. (2) and (3) we penalize systems predicting too many or too few instances by dividing by \(|S|\) and \(|T|\), respectively, e.g., in Figure 2 \(P(\{s_2, s_3\}, T) < P(\{s_3\}, T)\). Finally, we combine Eqs. (2) and (3) into an \(F_1\)-measure, the harmonic mean of precision and recall.

Having a separate function \(C\) to be responsible for comparing two annotations gives us some additional flexibility that is missing in standard NER measures that operate at the character/token level. For example, in Eq. (1) we could easily change the factor that gives credit for partial overlaps by being more forgiving when only few characters are wrong.
5 Tasks and Proposed Models

We define two tasks based on the corpus described in Section 3: (i) SLC (Sentence-level Classification), which asks to predict whether a sentence contains at least one propaganda technique, and (ii) FLC (Fragment-level classification), which asks to identify both the spans and the type of propaganda technique. Note that these two tasks are of different granularities, $g_1$ and $g_2$, i.e., tokens for FLC and sentences for SLC. We split the corpus into training, development and test, each containing 293, 57, 101 articles and 14,857, 2,108, 4,265 sentences.

5.1 Baselines

We depart from BERT (Devlin et al., 2019), as it has achieved state-of-the-art performance on multiple NLP benchmarks, and we design three baselines based on it.

BERT. We add a linear layer on top of BERT and we fine-tune it, as suggested in (Devlin et al., 2019). For the FLC task, we feed the final hidden representation for each token to a layer $L_{g_2}$ that makes a 19-way classification: does this token belong to one of the eighteen propaganda techniques or to none of them (cf. Figure 3-a). For the SLC task, we feed the final hidden representation for the special [CLS] token, which BERT uses to represent the full sentence, to a two-dimensional layer $L_{g_1}$ to make a binary classification.

BERT-Joint. We use the layers for both tasks in the BERT baseline, $L_{g_1}$ and $L_{g_2}$, and we train for both FLC and SLC jointly (cf. Figure 3-b).

BERT-Granularity. We modify BERT-Joint to transfer information from SLC directly to FLC. Instead of using only the $L_{g_2}$ layer for FLC, we concatenate $L_{g_1}$ and $L_{g_2}$, and we add an extra 19-dimensional classification layer $L_{g_{1,2}}$ on top of that concatenation to perform the prediction for FLC (cf. Figure 3-c).

5.2 Multi-Granularity Network

We propose a model that can drive the higher Granularity task (FLC) on the basis of the lower-granularity information (SLC), rather than simply using low-granularity information directly. Figure 3-d shows the architecture of this model. More generally, suppose there are $k$ tasks of increasing granularity, e.g., document-level, paragraph-level, sentence-level, word-level, subword-level, character-level.

Each task has a separated classification layer $L_{g_k}$ that receives the feature representation of the specific level of granularity $g_k$ and outputs $o_{g_k}$. The dimension of the representation depends on the embedding layer, while the dimension of the output depends on the number of classes in the task. The output $o_{g_k}$ generates a weight for the next granularity task $g_{k+1}$ through a trainable gate $f$:

$$w_{g_k} = f(o_{g_k})$$  (4)

The gate $f$ consists of a projection layer to one dimension and an activation function. The resulting weight is multiplied by each element of the output of layer $L_{g_{k+1}}$ to produce the output for task $g_{k+1}$:

$$o_{g_{k+1}} = w_{g_k} * o_{g_{k+1}}$$  (5)

If $w_{g_k} = 0$ for a given example, the output of the next granularity task $o_{g_{k+1}}$ would be 0 as well. In our setting this means that, if the sentence-level classifier is confident the sentence does not contain propaganda, i.e., $w_{g_k} = 0$, then $o_{g_{k+1}} = 0$ and there would be no propagandistic technique predicted for any span within that sentence. Similarly, when back-propagating the error, if $w_{g_k} = 0$ for a given example, the final entropy loss would become zero; i.e. the model would not get any information from that example. As a result, only examples strongly classified as negative in a lower granularity task would be ignored in the higher granularity task. Having the lower granularity as the main task means that higher-granularity information can be selectively used as additional information to improve the performance, but only if the example is not considered as highly negative. We show this in Section 6.3.
For the loss function, we use a cross-entropy loss with sigmoid activation for every layer, except for the highest-granularity layer \( L_{gK} \), which uses a cross-entropy loss with softmax activation. Unlike softmax, which normalizes over all dimensions, the sigmoid allows each output component of layer \( L_{gk} \) to be independent from the rest. Thus, the output of the sigmoid for the positive class increases the degree of freedom by not affecting the negative class, and vice versa. As we have two tasks, we use sigmoid activation for \( L_{g1} \) and softmax activation for \( L_{g2} \). Moreover, we use a weighted sum of losses with a hyper-parameter \( \alpha \):

\[
L_J = L_{g1} \ast \alpha + L_{g2} \ast (1 - \alpha)
\]

Again, we use BERT (Devlin et al., 2019) for the contextualized embedding layer and we place the multi-granularity network on top of it.

### 6 Experiments and Evaluation

#### 6.1 Experimental Setup

We used the PyTorch framework and the pre-trained BERT model, which we fine-tuned for our tasks. We trained all models using the following hyper-parameters: batch size of 16, sequence length of 210, weight decay of 0.01, and early stopping on validation \( F_1 \) with patience of 7. For optimization, we used Adam with a learning rate of 3e-5 and a warmup proportion of 0.1. To deal with class imbalance, we give weight to the binary cross-entropy according to the proportion of positive samples. For the \( \alpha \) in the joint loss function, we use 0.9 for sentence classification, and 0.1 for word-level classification. In order to reduce the effect of random fluctuations for BERT, all the reported numbers are the average of three experimental runs with different random seeds. As it is standard, we tune our models on the dev partition and we report results on the test partition.

#### 6.2 Fragment-Level Propaganda Detection

Table 6 shows the performance for the three baselines and for our multi-granularity network on the FLC task. For the latter, we vary the degree to which the gate function is applied: using ReLU is more aggressive compared to using the Sigmoid, as the ReLU outputs zero for a negative input. Note that, even though we train the model to predict both the spans and the labels, we also evaluated it with respect to the spans only.

<table>
<thead>
<tr>
<th>Model</th>
<th>Spans P</th>
<th>Spans R</th>
<th>Spans F1</th>
<th>Full Task P</th>
<th>Full Task R</th>
<th>Full Task F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>39.57</td>
<td>36.42</td>
<td>37.90</td>
<td>21.48</td>
<td>21.39</td>
<td></td>
</tr>
<tr>
<td>Joint</td>
<td>39.26</td>
<td>35.48</td>
<td>37.25</td>
<td>20.11</td>
<td>19.74</td>
<td>19.92</td>
</tr>
<tr>
<td>Granu</td>
<td>43.08</td>
<td>33.98</td>
<td>37.93</td>
<td>23.85</td>
<td>20.14</td>
<td>21.80</td>
</tr>
<tr>
<td>Multi-Granularity ReLU</td>
<td>43.29</td>
<td>34.74</td>
<td>38.28</td>
<td>23.98</td>
<td>20.33</td>
<td>21.82</td>
</tr>
<tr>
<td>Sigmoid</td>
<td>44.12</td>
<td>35.01</td>
<td>38.98</td>
<td>24.42</td>
<td>21.05</td>
<td>22.58</td>
</tr>
</tbody>
</table>

Table 6: Fragment-level experiments (FLC task). Shown are two evaluations: (i) **Spans** checks only whether the model has identified the fragment spans correctly, while (ii) **Full task** is evaluation wrt the actual task of identifying the spans and also assigning the correct propaganda technique for each span.

Table 6 shows that joint learning (BERT-Joint) hurts the performance compared to single-task BERT. However, using additional information from the sentence-level for the token-level classification (BERT-Granularity) yields small improvements. The multi-granularity models outperform all baselines thanks to their higher precision. This shows the effect of the model excluding sentences that it determined to be non-propagandistic from being considered for token-level classification. Nevertheless, the performance of sentence-level classification is far from perfect, achieving an \( F_1 \) of up to 60.98 (cf. Table 7). The information it contributes to the final classification is noisy and the more conservative removal of instances performed by the Sigmoid function yields better performance than the more aggressive ReLU.

#### 6.3 Sentence-Level Propaganda Detection

Table 7 shows the results for the SLC task. We apply our multi-granularity network model to the sentence-level classification task to see its effect on low granularity when we train the model with a high granularity task. Interestingly, it yields huge performance improvements on the sentence-level classification result. Compared to the BERT baseline, it increases the recall by 8.42%, resulting in a 3.24% increase of the \( F_1 \) score. In this case, the result of token-level classification is used as additional information for the sentence-level task, and it helps to find more positive samples. This shows the opposite effect of our model compared to the FLC task. Note also that using ReLU is more effective than using the Sigmoid, unlike in token-level classification.
Table 7: Sentence-level (SLC) results. *All-propaganda* is a baseline which always output the propaganda class.

Thus, since the performance range of the token-level classification is low, we think it is more effective to get additional information after aggressively removing negative samples by using ReLU as a gate in the model.

7 Related Work

Propaganda identification has been tackled mostly at the article level. Rashkin et al. (2017) created a corpus of news articles labelled as belonging to four categories: propaganda, trusted, hoax, or satire. They included articles from eight sources, two of which are propagandistic. Barrón-Cedeño et al. (2019) experimented with a binarized version of the corpus from (Rashkin et al., 2017): propaganda vs. the other three categories. The corpus labels were obtained with distant supervision, assuming that all articles from a given news outlet share the label of that outlet, which inevitably introduces noise (Horne et al., 2018).

A related field is that of computational argumentation which, among others, deals with some logical fallacies related to propaganda. Habernal et al. (2018b) presented a corpus of Web forum discussions with cases of *ad hominem* fallacy identified. Habernal et al. (2017, 2018a) introduced *Argotario*, a game to educate people to recognize and create fallacies. A byproduct of Argotario is a corpus with 1.3k arguments annotated with five fallacies, including *ad hominem*, *red herring* and *irrelevant authority*, which directly relate to propaganda techniques (cf. Section 2). Differently from (Habernal et al., 2017, 2018a,b), our corpus has 18 techniques annotated on the same set of news articles. Moreover, our annotations aim at identifying the minimal fragments related to a technique instead of flagging entire arguments.

8 Conclusion and Future Work

We have argued for a new way to study propaganda in news media: by focusing on identifying the instances of use of specific propaganda techniques. Going at this fine-grained level can yield more reliable systems and it also makes it possible to explain to the user why an article was judged as propagandistic by an automatic system.

In particular, we designed an annotation schema of 18 propaganda techniques, and we annotated a sizable dataset of documents with instances of these techniques in use. We further designed an evaluation measure specifically tailored for this task. We made the schema and the dataset publicly available, thus facilitating further research. We hope that the corpus would raise interest outside of the community of researchers studying propaganda: the techniques related to fallacies and the ones relying on emotions might provide a novel setting for the researchers interested in Argumentation and Sentiment Analysis.

We experimented with a number of BERT-based models and devised a novel architecture which outperforms standard BERT-based baselines. Our fine-grained task can complement document-level judgments, both to come out with an aggregated decision and to explain why a document—or an entire news outlet—has been flagged as potentially propagandistic by an automatic system.

We are collaborating with A Data Pro to expand the corpus. In the mid-term, we plan to build an online platform where professors in relevant fields (e.g., journalism, mass communication) can train their students to recognize and annotate propaganda techniques. The hope is to be able to accumulate annotations as a by-product of using the platform for training purposes.

Acknowledgments

This research is part of the Tanbih project,\(^7\) which aims to limit the effect of “fake news”, propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Qatar Computing Research Institute (QCRI), HBKU.

\(^7\)http://tanbih.qcri.org/
References


Benjamin D Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In International AAAI Conference on Web and Social Media, ICWSM ’18, Stanford, CA, USA.


