

Giovanni Da San Martino<sup>1</sup> Seunghak Yu<sup>1,2</sup> Alberto Barrón-Cedeño<sup>3</sup>  
Rostislav Petrov<sup>4</sup> Preslav Nakov<sup>1</sup>

<sup>1</sup>Qatar Computing Research Institute, HBKU, Qatar

<sup>2</sup>MIT CSAIL, USA <sup>3</sup>Università di Bologna, Italy <sup>4</sup>A Data Pro, Bulgaria

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

## Computational Propaganda

- Propaganda has been widely used since the advent of mass media



- However, Internet and social media "have allowed cross-border computational propaganda by foreign states or even private organizations" (Bolsover and Howard, Big Data 2017)

- Can we automatically detect the use of propaganda?  
Can we make America (and the world) aware again?
- Current approaches provide document-level predictions
  - rely on gold labels based on distant supervision → noisy
  - lack model explainability

## Propaganda Techniques

- Propaganda is conveyed through a series of rhetorical and psychological techniques



Figure 1: *Bandwagon*: join in because everyone else is taking the same action.

reductio ad Hitlerum  
flag-waving  
minimisation  
exaggeration  
intentional vagueness  
black-and-white fallacy  
appeal to prejudice  
whataboutism  
bandwagon  
causal oversimplification  
appeal to authority  
obfuscation  
name calling  
cognitive dissonance  
loaded language  
thought-terminating cliches  
labeling



Figure 2: *Name calling*.



Greta Thunberg: "We are in the middle of the sixth mass extinction, with more than 200 species getting extinct every day"

Figure 3: *Appeal to fear* (Can you spot another example in this poster?)

## Propaganda Techniques Corpus

450 news articles from 48 sources (21,230 sentences, 350K tokens) annotated at the fragment level with 18 propaganda techniques.

1	Manchin says Democrats acted like babies at the SOTU	Stereotyping, name calling or labeling
2	Democrat West Virginia Sen. Joe Manchin says his colleagues' refusal to stand or applaud during President Donald Trump's State of the Union speech was disrespectful and a signal that the party is more concerned with obstruction than it is with progress.	Black-and-white Fallacy
4	In a glaring sign of just how stupid and petty things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech	Loaded language
	not looking as though Trump killed his grandma.	Exaggeration, Loaded language
6	As Manchin noted, many 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	Exaggeration

## Annotation Process

- Phase 1: two annotators,  $a_i$  and  $a_j$ , independently annotated the same article
- Phase 2:  $a_i$  and  $a_j$  discussed with a consolidator  $c_1$  all instances to come up with a final annotation.

The table shows  $\gamma$  inter-annotator agreement for spans only and spans + labels between two annotators and one annotator and one consolidator.

Annotations		spans ( $\gamma_s$ )	+labels ( $\gamma_{sl}$ )
$a_1$	$a_2$	0.30	0.24
$a_3$	$a_4$	0.34	0.28
$a_1$	$c_1$	0.58	0.54
$a_2$	$c_1$	0.74	0.72
$a_3$	$c_2$	0.76	0.74
$a_4$	$c_2$	0.42	0.39

## Tasks and Evaluation Measure

- FLC**: detect the text fragments in which a propaganda technique is used and identify the technique. **Spans** is a lighter version of the task in which only the span has to be identified.
- SLC** detect the sentences that contain one or more propaganda techniques (binary task).

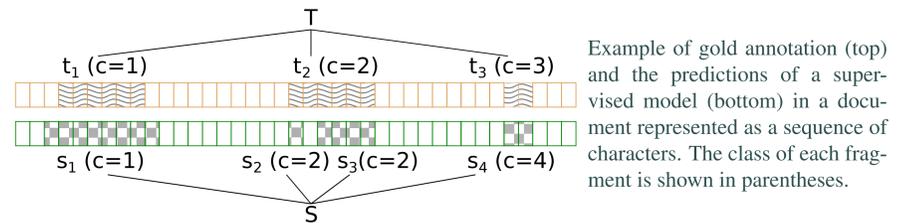
An evaluation measure for Task FLC needs to be defined. We use a variant of the standard  $F_1$  (and Precision, P, and Recall, R) taking into account partially overlapping spans:

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

where

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

here  $\delta(a, b) = 1$  if  $a = b$ , and 0 otherwise.



Example of gold annotation (top) and the predictions of a supervised model (bottom) in a document represented as a sequence of characters. The class of each fragment is shown in parentheses.

## Models

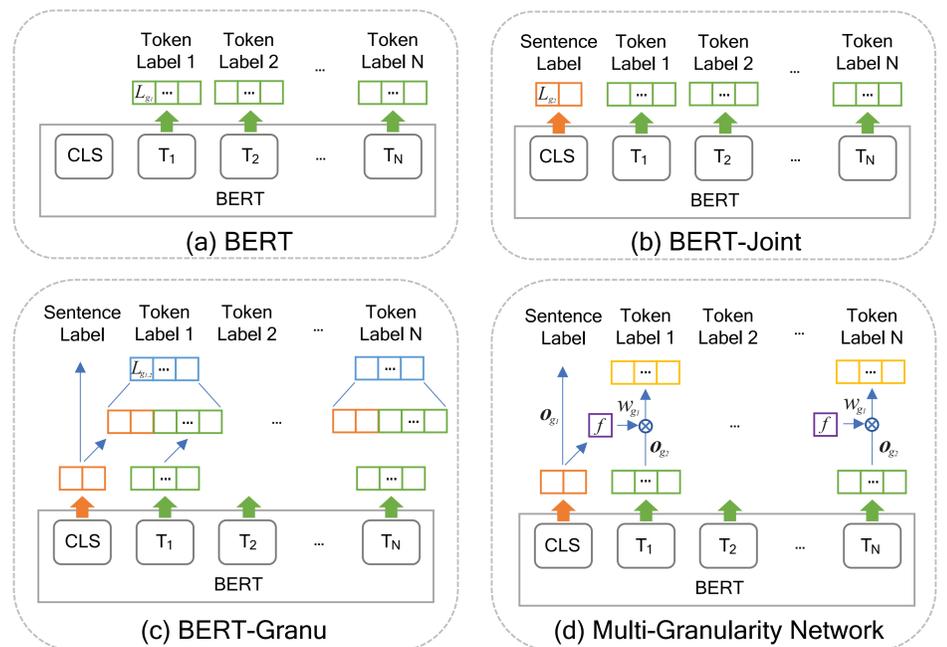


Figure 4: The architecture of the baseline models (a-c), and of our proposed multi-granularity network (d).

## Experiments

Model	FLC Task			Spans			SLC Task		
	P	R	$F_1$	P	R	$F_1$	P	R	$F_1$
BERT	21.48	<b>21.39</b>	21.39	39.57	36.42	37.90	<b>63.20</b>	53.16	57.74
Joint	20.11	19.74	19.92	39.26	35.48	37.25	62.84	55.46	58.91
Granu	23.85	20.14	21.80	43.08	33.98	37.93	62.80	55.24	58.76
Multi-Granularity									
ReLU	23.98	20.33	21.82	43.29	34.74	38.28	60.41	<b>61.58</b>	<b>60.98</b>
Sigmoid	<b>24.42</b>	21.05	<b>22.58</b>	<b>44.12</b>	<b>35.01</b>	<b>38.98</b>	62.27	59.56	60.71

Table 1: Evaluation of the models for Spans, FLC and SLC tasks. The proposed models improve over the baselines.

## What We Are Up To

- SemEval 2020 Task 11 on Fine Grained Propaganda Detection: <https://propaganda.qcri.org/semEval2020-task11>
- Our Propaganda Analysis Project (where you can find this paper): <https://propaganda.qcri.org>
- The Tanbih Project, which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading: <http://tanbih.qcri.org>