Explaining Fine-Grained Detection of Propaganda in News Articles using NLP

Anurag Pant
UID: 705085298

This article works on simplifying the paper ‘Fine-Grained Analysis of Propaganda in News Articles’ published in EMNLP-2019. The work done in this paper attempts to find the different types of propaganda present in different news articles that may aim at influencing people's mindset with the purpose of advancing a specific agenda.

Background

Previous work done on detecting propaganda in news articles typically ended up labelling an entire news outlet as a source of propaganda news instead of locating instances of propaganda within a single article. This presents a problem since, at times, a news outlet known for producing propaganda news articles can produce objective non-propagandistic articles and vice-versa. Mislabeled articles can cause problems in the training of models that work on

detecting propaganda. At the same time, this blanket label prevents us from finding out the type of propaganda(s) present within the article that prompted it to be classified as such. This paper works on creating an expert annotated dataset of news articles which can be used in the development of explainable AI systems.

What is Propaganda?

Propaganda is the use of psychological and emotional techniques to appeal to the emotions of the audience in order to influence people's mindsets and advance a specific agenda. Propaganda techniques are often used to hide the logical fallacies in news reports while still persuading the audience about the arguments being made.

Forms of Propaganda

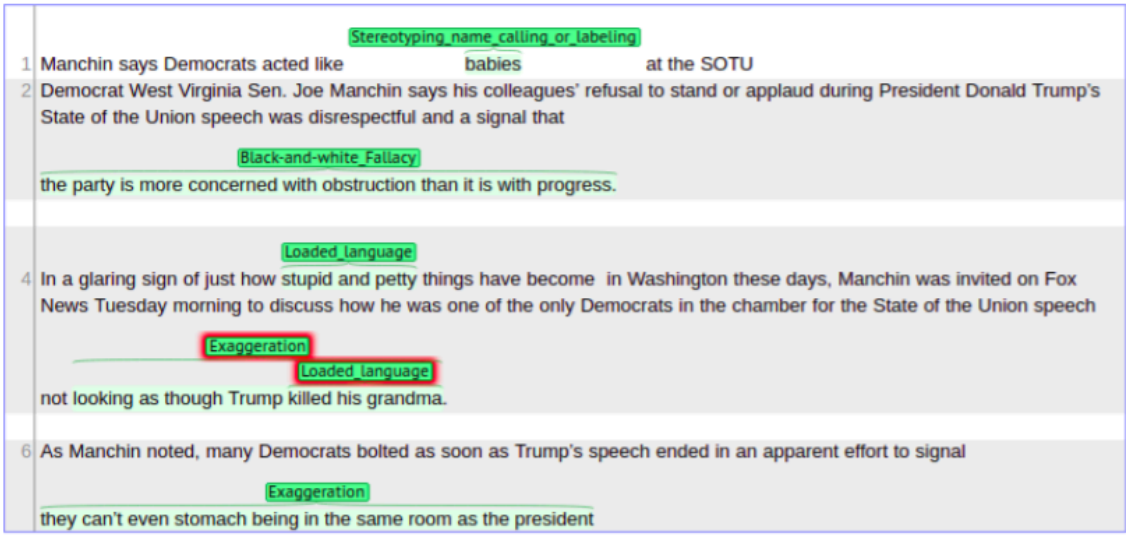
This paper discusses 18 forms of propaganda commonly found in news articles:

1. **Loaded Language.** Using words/phrases with strong emotional implications to influence an audience. E.g. "a lone lawmaker's childish shouting"
2. **Name-calling or labelling.** Labelling the object of the propaganda campaign as something the audience already has strong views about. E.g. "Republican congressweasels"
3. **Repetition.** Repeating the message multiple times to make the audience accept it.
4. **Exaggeration or minimization.** Representing something in an excessive manner or making something seem less important or smaller than it actually is. E.g. "the best of the best",

5. **Doubt.** Questioning the credibility of someone/something. E.g. “Is he ready to be the Mayor?”
6. **Appeal to fear/prejudice.** Using anxiety and/or panic to create support for an idea. E.g. “stop those refugees; they are terrorists.”
7. **Flag-waving.** Playing on strong national feeling to promote an action/idea. E.g. “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. E.g. “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 labelling/stereotyping. E.g. “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.
11. **Black-and-white fallacy, dictatorship.** Presenting two alternative options as the only possibilities. E.g. “There is no alternative to war.”
12. **Thought-terminating cliche.** Words or phrases that discourage critical thought and meaningful discussion. E.g. “it’s common sense”
13. **Whataboutism.** Discredit an opponent’s position by charging them with hypocrisy without directly disproving their argument.
14. **Reductio ad Hitlerum.** Disapprove an action/idea by suggesting that the idea is popular with groups hated by the target audience. E.g. “Only a communist can think this way”
15. **Red herring.** Introducing irrelevant material to the issue being discussed, so that everyone’s attention is diverted away from the points made.
16. **Bandwagon.** Persuade the audience to join in and take the course of action because “everyone else is taking the same action”. E.g. “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.
18. **Straw man.** When an opponent’s proposition is substituted with a similar one which is then refuted in place of the original.

Procuring Data and Manual Annotation

451 news articles from 48 news outlets (13 propagandistic and 36 nonpropagandistic) were annotated to create the dataset. This task wasn't well suited for crowdsourcing (since it required significant effort to memorize the different propaganda techniques), hence, the authors used a company that performs expert annotations.



Example given to the annotators

Given the initial poor inter-annotator agreement computed on the news articles, a consolidator worked with each pair of annotators (6 annotators and 3 consolidators). After working with the consolidator, the inter-annotator agreement increased significantly. The authors were able to conclude that the major reason behind the initial poor inter-annotator agreement was that mostly one of the annotators had missed some instances of propaganda in the initial stage.

Propaganda Technique	inst	avg. length
loaded language	2,547	23.70 ± 25.30
name calling, labeling	1,294	26.10 ± 19.88
repetition	767	16.90 ± 18.92
exaggeration, minimization	571	45.36 ± 35.55
doubt	562	123.21 ± 97.65
appeal to fear/prejudice	367	93.56 ± 74.59
flag-waving	330	61.88 ± 68.61
causal oversimplification	233	121.03 ± 71.66
slogans	172	25.30 ± 13.49
appeal to authority	169	131.23 ± 123.2
black-and-white fallacy	134	98.42 ± 73.66
thought-terminating cliches	95	34.85 ± 29.28
whataboutism	76	120.93 ± 69.62
reductio ad hitlerum	66	94.58 ± 64.16
red herring	48	63.79 ± 61.63
bandwagon	17	100.29 ± 97.05
obfusc., int. vagueness, confusion	17	107.88 ± 86.74
straw man	15	79.13 ± 50.72
all	7,485	46.99 ± 61.45

Dataset Statistics

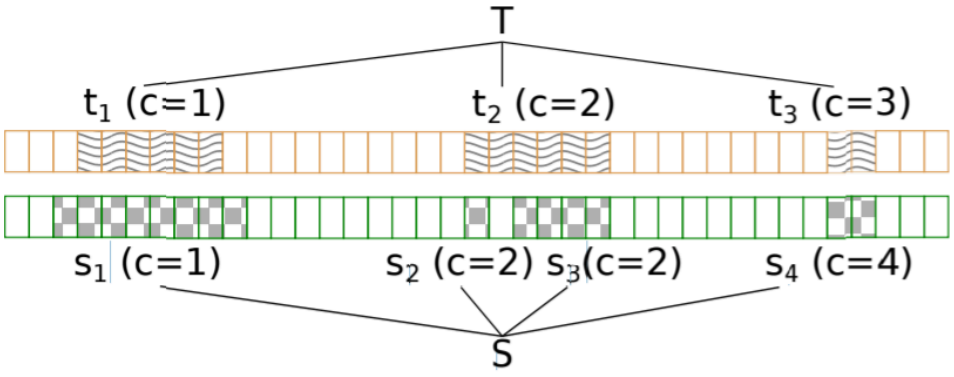
Evaluation Measures and Tasks

Two tasks were performed on the dataset created by the authors:

1. **SLC (Sentence-level Classification).** Predict whether a sentence contains at least one propaganda technique.
2. **FLC (Fragment-level Classification).** Identify both the spans and the type of propaganda technique.

A number of the spans might overlap in the text. In order to fairly evaluate the model, we need an evaluation measure that gives credits for partial overlaps.

Let the document be represented by *d.* A propagandistic text fragment in *d* is represented as *t*. A document includes a set of (possibly overlapping) fragments T. A model produces a set S with fragments *s* predicted on *d*. A labelling function *l(x)* associates *s* to 1 of the 18 propaganda techniques.



To handle partial overlaps between fragments with the same labels:

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

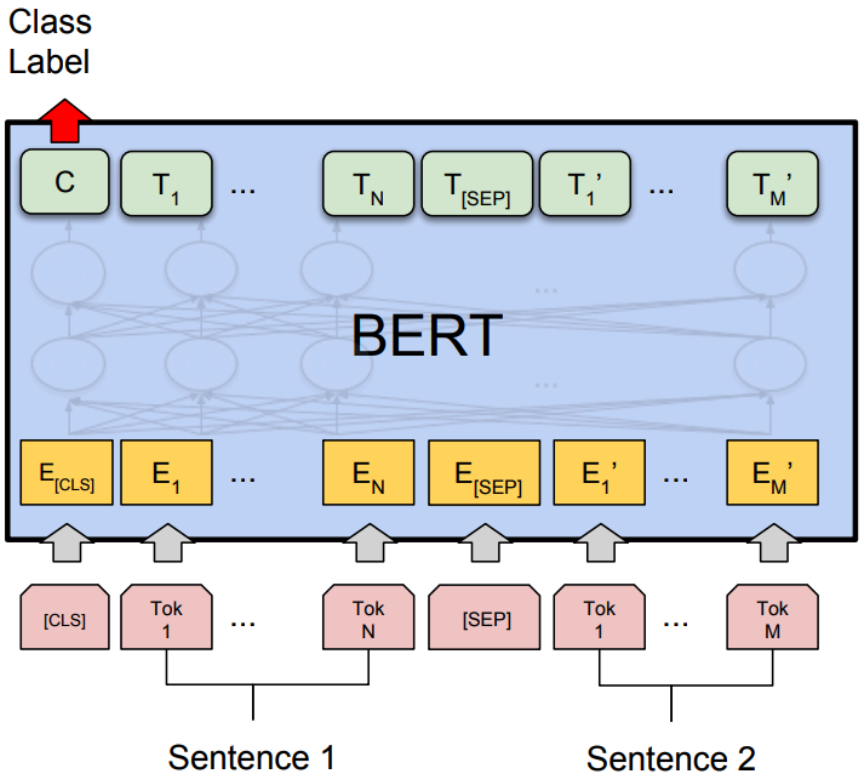
where *h* is a normalizing factor and δ(a, b) = 1 if a = b, else 0. The precision and recall of the model are then defined using the above equation as follows:

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

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

BERT: State-of-the-Art Language Model

Before we get into a discussion about the different language employed by the authors, we need to get familiarized with BERT which is a state-of-the-art language model created by researchers at Google AI Language. BERT applies bidirectional training to a Transformer, which is a self-attention mechanism. Self-attention allows a model to learn the context of a given word based on its surroundings. To get a more in-depth idea about BERT, you can read this article.

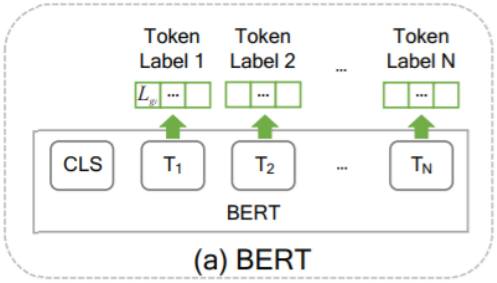


Baseline Models

The baseline models derive from BERT. The authors created 3 baseline models:

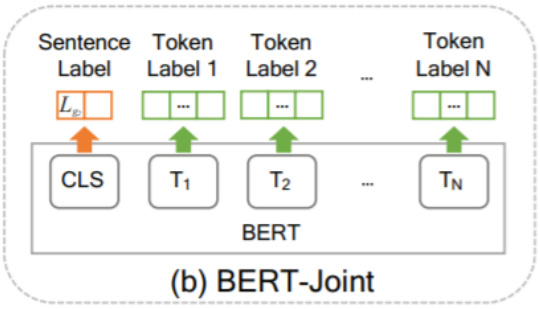
1. **BERT.** The authors add a linear layer on top of BERT and fine-tune it. For FLC, it is a 19-way classification task (L_{g2}), either 1 of the 18 propaganda techniques or none of

them. For SLC, the final hidden representation produced by BERT for the full sentence is passed to a binary classifier (L_{g1}).



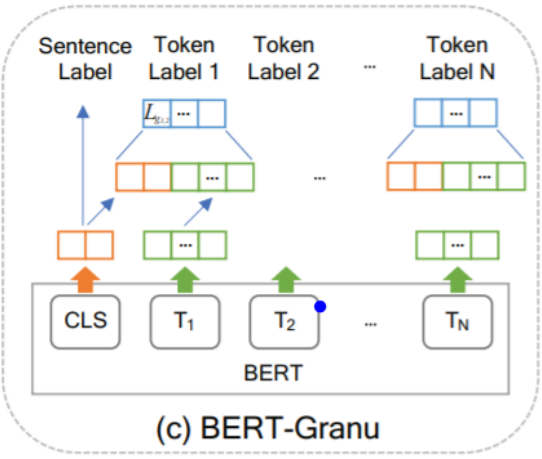
(a) BERT

2. **BERT-Joint.** The layers from the BERT baseline model, L_{g1} and L_{g2}, are used for both tasks and training for FLC and SLC occurs simultaneously.



(b) BERT-Joint

3. **BERT-Granularity.** Bert-Joint is modified to transfer information from SLC to FLC. L_{g1} and L_{g2} are concatenated and an extra 19-dimensional classification layer $L_{g1,2}$ is added on top to make the prediction for FLC.



(c) BERT-Granu

Multi-Granularity Network Model

The authors propose a model where the lower granularity task (SLC) drives the higher granularity task (FLC). The model uses contextualized embeddings produced by BERT.

Consider a general case, with *k* tasks of increasing granularity. Each task has a separated classification layer L_{gk} that receives features of granularity g_k and outputs o_{gk}. The output generates a weight for the next granularity task g_{k+1} using a trainable gate *f*:

$$w_{g_k} = f(\boldsymbol{o}_{g_k})$$

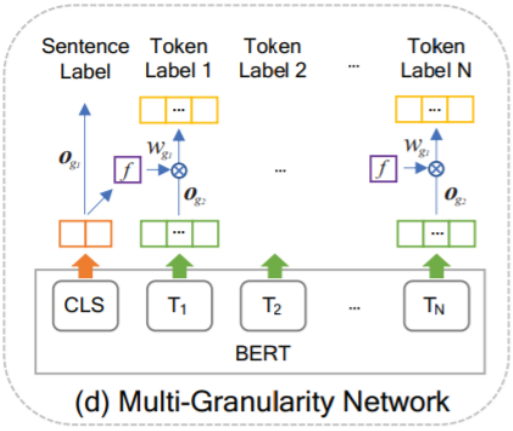
The resulting weight is multiplied by each element of the output of layer L_{gk+1} to produce the output for task g_{k+1}:

$$\boldsymbol{o}_{g_{k+1}} = w_{g_k} * \boldsymbol{o}_{g_{k+1}}$$

For this set of task, it means that if the sentence-level classifier is confident the sentence does not contain propaganda, $w_{gk} = 0$, then $o_{gk+1} = 0$ and there would be no propagandistic technique predicted for any span within that sentence.

For the loss function, the authors use sigmoid activation for L_{g1} and softmax activation for L_{g2}. A weighted sum of losses with hyperparameter α (taken experimentally as 0.9) was used:

$$\mathcal{L}_{\mathcal{J}} = \mathcal{L}_{g_1} * \alpha + \mathcal{L}_{g_2} * (1 - \alpha)$$



(d) Multi-Granularity Network

Results

The results for the FLC task using the baseline models and the multi-granularity network proposed by the authors are as follows:

Model	Spans			Full Task		
	P	R	F_1	P	R	F_1
BERT	39.57	36.42	37.90	21.48	**21.39**	21.39
Joint	39.26	35.48	37.25	20.11	19.74	19.92
Granu	43.08	33.98	37.93	23.85	20.14	21.80
Multi-Granularity						
ReLU	43.29	34.74	38.28	23.98	20.33	21.82
Sigmoid	**44.12**	**35.01**	**38.98**	**24.42**	21.05	**22.58**

Here, Spans checks only whether the model has identified the fragment spans correctly, while Full Task evaluates according to the task of identifying the spans and assigning the correct propaganda technique.

The results for the SLC task using the baseline models and the multi-granularity network proposed by the authors are as follows:

Model	Precision	Recall	F1
All-Propaganda	23.92	100.0	38.61
BERT	**63.20**	53.16	57.74
BERT-Granu	62.80	55.24	58.76
BERT-Joint	62.84	55.46	58.91
MGN Sigmoid	62.27	59.56	60.71
MGN ReLU	60.41	**61.58**	**60.98**

Here, All-Propaganda is a baseline that always outputs the propaganda class.

It is abundantly evident from the above results that the multi-granularity network proposed by the authors gives the best result for the detection of propaganda within news articles.

Related Work

The research paper titled ‘Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking’ by Rashkin et al. (2017) worked on creating a database of news articles that belong to the categories: Hoax, Propaganda, Trusted News and Satire. They used these articles to learn Linguistic Inquiry and Word Count (LIWC) lexicon, which when used as features for NLP models help to improve the performance in predicting truthfulness of the news. This dataset, however, works on an article level and doesn’t go into the finer granularity like the paper discussed.

Another research paper ‘Proppy: A System to Unmask Propaganda in Online News’ by. Barron-Cedeno et al. (2019) worked on solving a similar problem. They presented a publicly available real-world real-time propaganda detection tool for online available news. They used n-gram features, lexicon features like LIWC along with some other features to detect propaganda within news articles. They evaluated Proppy on the dataset created by Rashkin et al. (2017) and used a binarized version of the dataset: propaganda vs. the other three categories. Their work was also based on an article-level granularity.

The work done by Horne et al. (2018) in their paper titled ‘Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape’ also aims at creating a dataset of news articles. They source 136K news articles from 92 news sources. They choose well-established and mainstream sources, fake news sources, satire sources and biased political blogs to get a wide range of news articles. The news articles are automatically tagged with the label of the news source from where they are sourced, e.g. all articles from fake news sources are automatically adjudged to be fake, which introduces a lot of noise in the dataset. The research paper discussed in our article works towards alleviating this issue.

In ‘Before Name-calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation’ by Habernal et al. (2018), the authors work with finding fallacies in arguments made online (similar to propaganda in news articles). They concentrate on the ad hominem fallacy in particular which they further divided into 5 subtypes: abusive, tu quoque, circumstantial, bias and guilt by association. The authors created a dataset using a forum on subreddit meant for online debates. They performed an in-depth analysis of the dataset, even employing neural models to recognize ad hominem arguments and to guess the reasonableness and controversial nature of the original post. The paper deals with detecting the fallacies at a higher granularity (entire argument) and only talks about 5 subtypes while the research paper discussed in our article deals with 18 types of propaganda and lower granularity.

Habernal et al. (2017) have also previously worked on ‘Argotario: Computational Argumentation Meets Serious Games’. In this research paper, they worked on creating a game that works on helping people recognize fallacies in arguments and awards them with in-game rewards for performing well. No NLP techniques are used in this paper, however, the final aim of the game is to lead to the creation of a dataset (since all data is provided by the players themselves) which can be used for future research. Their work, while novel, deals with higher granularity and doesn’t divide the fallacies into many subtypes, like the paper discussed in our article.