Generating and Identifying Propaganda With Machine Learning

Originally published at https://www.unite.ai/generating-and-identifying-propaganda-with-machine-learning/

New research from the United States and Qatar offers a novel method for identifying fake news that has been written in the way that humans *actually write* fake news – by embedding inaccurate statements into a largely truthful context, and by the use of popular propaganda techniques such as *appeals to authority* and *loaded language*.

The project has resulted in the creation of a new fake news detection training dataset called *PropaNews*, which incorporates these techniques. The study's authors have found that detectors trained on the new dataset are 7.3-12% more accurate in detecting human-written disinformation than prior state-of-the-art approaches.

Technique	Generated Disinformation and Propaganda
Appeal to Authority	Cairo's Tahrir Square was the scene of clashes between protesters and police on Wednesday. " At least three people were killed and more than 600 were injured in the clashes, " said Egypt's President.
Loaded Language	Cairo's Tahrir Square was the scene of deadly clashes between protesters and police on Wednesday.

From the new paper, examples of 'appeal to authority' and 'loaded language'. Source: https://arxiv.org/pdf/2203.05386.pdf

The authors claim that to the best of their knowledge, the project is the first to incorporate propaganda techniques (rather than straightforward factual inaccuracy) into machine-generated text examples intended to fuel fake news detectors.

Most recent work in this field, they contend, has studied bias, or else reframed 'propaganda' data in the context of bias (arguably because bias became a highly fundable machine learning sector in the post-Analytica era).

The authors state:

> *'In contrast, our work generates fake news by incorporating propaganda techniques and preserving the majority of the correct information. Hence, our approach is more suitable for studying defense against human-written fake news.'*

They further illustrate the growing urgency of more sophisticated propaganda-detection techniques*:

> *'[Human-written] disinformation, which is often used to manipulate certain populations, had catastrophic impact on multiple events, such as the 2016 US Presidential Election, Brexit, the COVID-19 pandemic, and the recent Russia's assault on Ukraine. Hence, we are in urgent need of a defending mechanism against human-written disinformation.'*

The paper is titled *Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation*, and comes from five researchers at the University of Illinois Urbana-Champaign, Columbia University, Hamad Bin Khalifa University at Qatar, the University of Washington, and the Allen Institute for AI.

Defining Untruth

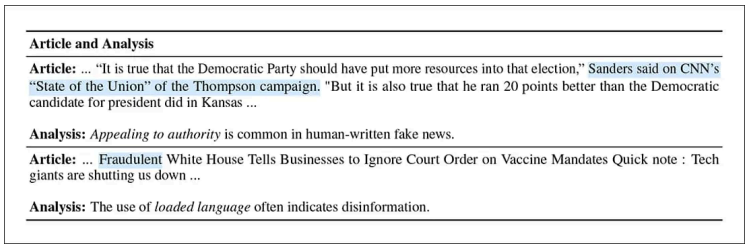
The challenge of quantifying propaganda is largely a logistical one: it is very expensive to hire humans to recognize and annotate real-world material with propaganda-like characteristics for inclusion in a training dataset, and potentially far cheaper to extract and utilize high-level features that are likely to work on 'unseen', future data.

In service of a more scalable solution, the researchers initially gathered human-created disinformation articles from news sources deemed to be low in factual accuracy, via the Media Bias Fact Check site.

They found that 33% of the articles studied used disingenuous propaganda techniques, including *emotion-triggering terms*, *logical fallacies*, and *appeal to authorities*. An additional 55% of the articles contained inaccurate information mixed in with accurate information.

Generating Appeals to Authority

The *appeal to authority* approach has two use-cases: the citation of inaccurate statements, and the citation of completely fictitious statements. The research focuses on the second use case.



From the new project, the Natural Language Inference framework RoBERTa identifies two further examples of appealing to authority and loaded language.

With the objective of creating machine-generated propaganda for the new dataset, the researchers used the pretrained seq2seq architecture BART to identify salient sentences that could later be altered into propaganda. Since there was no publicly available dataset related to this task, the authors used an extractive summarization model proposed in 2019 to estimate sentence saliency.

For one article from each news outlet studied, the researchers substituted these 'marked' sentences with fake arguments from 'authorities' derived both from the Wikidata Query Service and from authorities mentioned in the articles (i.e. people and/or organizations).

Generating Loaded Language

Loaded language includes words, often sensationalized adverbs and adjectives (as in the above-illustrated example), that contain implicit value judgements enmeshed in the context of delivering a fact.

To derive data regarding loaded language, the authors used a dataset from a 2019 study containing 2,547 *loaded language* instances. Since not all the examples in the 2019 data included emotion-triggering adverbs or adjectives, the researchers used SpaCy to perform dependency parsing and Part of Speech (PoS) tagging, retaining only apposite examples for inclusion in the framework.

The filtering process resulted in 1,017 samples of valid *loaded language*. Another instance of BART was used to mask and replace salient sentences in the source documents with loaded language.

PropaNews Dataset

After intermediate model training conducted on the 2015 CNN/DM dataset from Google Deep Mind and Oxford University, the researchers generated the PropaNews dataset, converting non-trivial articles from 'trustworthy' sources such as *The New York Times* and *The Guardian* into 'amended' versions containing crafted algorithmic propaganda.

The experiment was modeled on a 2013 study from Hanover, which automatically generated timeline summaries of news stories across 17 news events, and a total of 4,535 stories.

The generated disinformation was submitted to 400 unique workers at Amazon Mechanical Turk (AMT), spanning 2000 Human Intelligence Tasks (HITs). Only the propaganda-laden articles deemed *accurate* by the workers were included in the final version of PropaNews. Adjudication on disagreements were scored by the Worker Agreement With Aggregate (WAWA) method.

The final version of PropaNews contains 2,256 articles, balanced between fake and real output, 30% of which leverage *appeal to authority*, with a further 30% using *loaded language*. The remainder simply contains inaccurate information of the type which has largely populated prior datasets in this research field.

The data was split 1,256:500:500 across training, testing and validation distributions.

HumanNews Dataset

To evaluate the effectiveness of the trained propaganda detection routines, the researchers compiled 200 human-written news articles, including articles debunked by Politifact, and published between 2015-2020.

This data was augmented with additional debunked articles from untrustworthy news media outlets, and the sum total fact-checked by a computer science major graduate student.

The final dataset, titled HumanNews, also includes 100 articles from the *Los Angeles Times*.

Tests

The detection process was pitted against prior frameworks in two forms: *PN-Silver*, which disregards AMT annotator validation, and *PN-Gold*, which includes the validation as a criteria.

Competing frameworks included the 2019 offering Grover-GEN, 2020's Fact-GEN, and *FakeEvent*, wherein articles from PN-Silver are substituted with documents generated by these older methods.

Training Data	Detector: GROVER-LARGE	Detector: ROBERTA-LARGE
Without human validation (silver)		
GROVER-GEN	59.5 (±3.0)	54.9 (±2.6)
FAKEEVENT	44.7 (±1.8)	42.4 (±2.5)
FACTGEN	45.8 (±7.7)	43.9 (±3.2)
PN-SILVER	**62.1** (±4.6)	**61.8** (±1.3)
With human validation (gold)		
PROPANEWS	**66.8** (±2.0)	**66.9** (±2.1)
w/o AA	64.2 (±2.1)	64.8 (±2.9)
w/o LL	65.6 (±2.9)	65.2 (±1.8)
w/o AA & LL	63.8 (±4.8)	63.8 (±4.4)

Variants of Grover and RoBERTa proved to be most effective when trained on the new PropaNews dataset, with the researchers concluding that *'detectors trained on PROPANEWS perform better in identifying human-written disinformation compared to training on other datasets'*.

The researchers also observe that even the semi-crippled ablation dataset PN-Silver outperforms older methods on other datasets.

Out of Date?

The authors reiterate the lack of research to date regarding the automated generation and identification of propaganda-centric fake news, and warn that the use of models trained on data prior to critical events (such as COVID, or, arguably, the current situation in eastern Europe) cannot be expected to perform optimally:

> *'Around 48% of the misclassified human-written disinformation are caused by the inability to acquire dynamic knowledge from new news sources. For instance, COVID-related articles are usually published after 2020, while ROBERTA was pre-trained on news articles released before 2019. It is very challenging for ROBERTA to detect disinformation of such topics unless the detector is equipped with the capabilities of acquiring dynamic knowledge from news articles.'*

The authors further note that RoBERTa achieves a 69.0% accuracy for the detection of fake news articles where the material is published prior to 2019, but drops down to 51.9% accuracy when applied against news articles published after this date.

Paltering and Context

Though the study does not directly address it, it's possible that this kind of deep dive into semantic affect could eventually address more subtle weaponization of language, such as paltering – the self-serving and selective use of truthful statements in order to obtain a desired result that may oppose the perceived spirit and intent of the supporting evidence used.

A related and slightly more developed line of research in NLP, computer vision and multimodal research is the study of context as an adjunct of meaning, where selective and self-serving reordering or re-contextualizing of true facts becomes equivalent to an attempt to evince a different reaction than the facts might ordinarily effect, had they been presented in a clearer and more linear fashion.

** My conversion of the authors' inline citations to direct hyperlinks.*

First published 11th March 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai