Extraction of Causal Narratives from News Articles

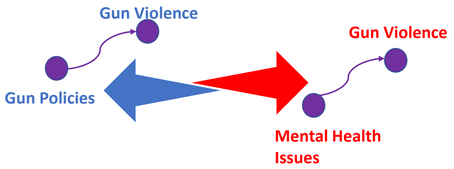
NYU Center for Data Science

Group Members: Varsha Vattikonda, Vanessa Xu, Victor Cui

Mentors: Marco Morucci, Guillaume Frechette, Sevgi Yuksel, Dingyue Liu

Abstract

This project aims to automate the process of discovering Causal Narratives in News articles from major US outlets. The motivation behind this is to understand whether different media outlets propagate different narratives about the same sets of facts. Specifically, we are interested in whether left-leaning and right-leaning news outlets adopt and propagate different causal stories for the same outcome.

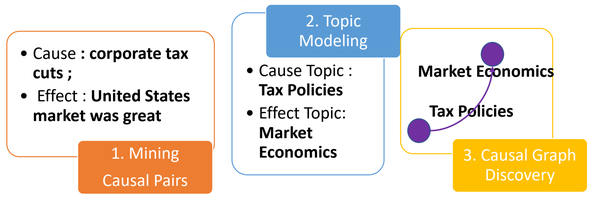


Introduction

INTRODUCTION : We have converted this into a Data Science problem constituting the following steps:

- Identify and extract cause-effect phrase pairs.
- Identify the topics extracted cause and effect phrases belong to.
- Generate causal graphs using topic-labeled cause-effect pairs.

INPUT : In January, with corporate tax cuts in place, the outlook for the United States market was great.



DATA : The POLUSA dataset constructed by L. Gebhard and F. Hamborg is our major source of data.[1] It contains 0.9 million news articles over a span of two and a half years from 18 major news outlets in the U.S. Outlets are labeled with political leaning, which is necessary to study political polarization, so it well suits our project goal. Unfortunately, it is not constructed specifically for causal analysis purposes, so it does not have any labels required in our following experiments and evaluations, being a major drawback.

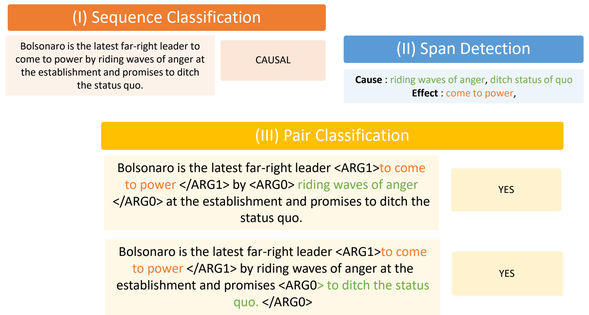
Date	Outlet	Headline	Body	Leaning
1/1 /2017 0:00	Los Angeles Times	Afghan refugees coming to California struggle with PTSD	California's capital has emerged as a leading destination for Afghan refugees who were awarded special visas because of their service to coalition forces in the war.	LEFT

PRE-PROCESSING : Because the causal mining models we are using only work on texts that are 2-3 sentences long, we have split each article into chunks of 1,2 and 3 sentences and then fed into the models as input

Mining Causal Pairs

Causal text Mining is a 3 stepped approach:[2]

1. **Sequence Classification**: Given an example, does it contain any causal relations? Based on BERT and predicts logits for the two labels
2. **Span Detection**: Given a causal example, which words in the sentence correspond to the Cause and Effect arguments? It might identify multiple Causes/Effects in each sample. Based on BERT and predicts BIO-CE tags for each token
3. **Pair Classification**: Pass all the combinations of generated Cause and Effect pairs to model for classifying if each pair is causal. Based on BERT and predicts logits for the two labels



Input	(I) Sequence Classification	(II) Span Detection	(III) Pair Classification
Baghdad has seen near-daily attacks blamed on Islamic militants since 2003.	CAUSAL	**Cause**: 2003 **Effect**: Baghdad has seen near-daily attacks blamed	NO
The raises are especially notable when compared to the federal minimum wage, which has stayed constant at $7.25 since 2009.	NOT CAUSAL	N/A	N/A

Topic Modeling

Nearest Neighbor Match

"A safe abortion"

What BERT already does:

Lexical Overlap

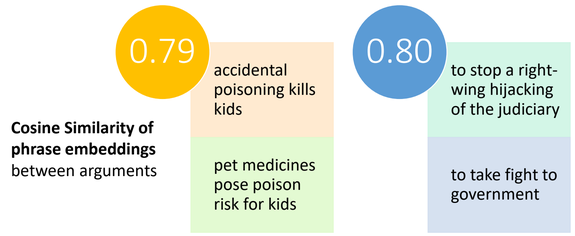
"Getting an abortion safely"

What **Phrase BERT** can do:

Semantically similar & lexically diverse

"**Ending a pregnancy by a trained professional who follows the WHO guidelines**"

We trained a Phrase-BERT [4] based topic model on all of the extracted causal arguments, including both causes and effects, and clustered them into 50 most relevant topic groups. We then manually labelled each topic based on their top 10 most relevant arguments, according to topic embeddings.



Cosine Similarity of phrase embeddings between arguments

Topic examples generated from Phrase BERT based Topic Models

Foreign Politics

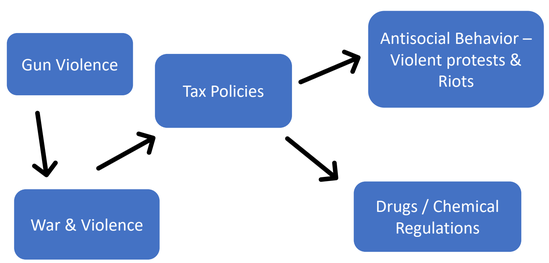
- to intimidate China
- to thwart illegal border crossings
- Iran step away a potential conflict with the United States
- counter China's conversion of contested islands into military bases
- block arms sales to Saudi Arabia; to halt Iran's aggressions
- the State Department is looking for unusual allies to tighten the screws on Iran

Gun Violence

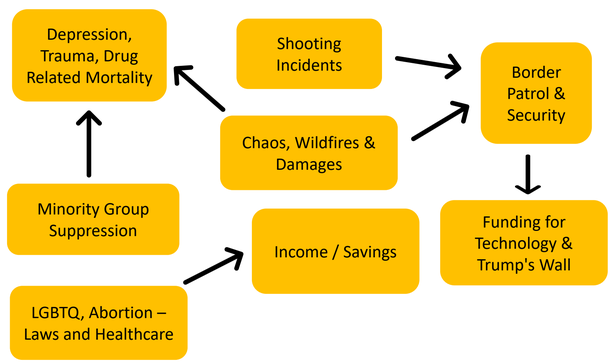
- who chose to kill children with a revolver instead of a high-caliber shotgun
- shot and killed in front of his sister by a hooded gunman on Friday night
- when Jesus Mesa Jr., a border guard, shot a fleeing 15-year-old boy in the head
- killing a man who had picked her up when she was a teenage trafficking victim
- the gunman was holding people hostage and claimed that he had a bomb.

Results

Left Aligned : The New York Times (NYT)



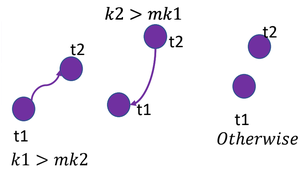
Center Aligned : The National Broadcasting Company (NBC)



Causal Graph Discovery

We have hence adopted a simple approach [3] based on counts from the population which proved to be quite effective. This approach - Selects causal edges that appear m-times more than non-causal edges between the same topics i.e.

CAUSE EFFECT	Topic : t1	Topic: t2
Topic: t1		k2
Topic: t2	k1	



Here, while attempting to reduce noise in the graph by filtering out connections that have stronger non-causal support in the textual repository we are also attempting to filter away the topics that are just corelated and hence co-occur.

Future Work

- All the models we have used for Causal Pairs Mining are not fine-tuned on POLUSA, and hence the results have only been sub-optimal. Additional labelled data is expected to bring substantial improvements.
- One other key area that needs to be explored is the method of evaluation for an unsupervised problem like this. In addition, Causal Pairs extraction was only done for 2-3 sentence chunks and hence using state-of-the-art models that can mine causal pairs from long texts will be a great addition to the solution.
- Topic modeling methods that do not require manual labelling.
- Look into suitable advanced causal graph constructing algorithms.

Acknowledgement

Big thanks to our mentors Marco Morruci, Guillaume Frechette, Sevgi Yuksel, and Dingyue Liu for their help along the way, as well as instructor Najoung Kim for the regular check-in and timely suggestions.

References

[1] Lukas Gebhard, Felix Hamborg.
The POLUSA Dataset: 0.9M Political News Articles Balanced by Time and Outlet Popularity, 2020.
[2] Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng.
Unicausal: Unified benchmark and model for causal text mining, 2022.
[3] Galia Nordon, Gideon Koren, Varda Shalev, Benny Kimelfeld, Uri Shalit, & Kira Radinsky.
Building causal graphs from medical literature and electronic medical records
Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):1102–1109, Jul. 2019
[4] Shufan Wang, Laure Thompson, & Mohit Iyyer
Phrase-bert: Improved phrase embeddings from bert with anapplication to corpus exploration.
In Empirical Methods in Natural Language Processing, 2021