Pipeline for modeling causal beliefs from natural language

J Hunter Priniski
Department of Psychology
UCLA
priniski@ucla.edu

Ishaan Verma
Information Sciences Institute
USC
iverma@usc.edu

Fred Morstatter
Information Sciences Institute
USC
fredmors@isi.edu

Abstract

Causal reasoning is a core cognitive function and is central to how people learn and update their beliefs. Causal information is also central to how people represent and use language. Natural Language Processing algorithms that detect people's causal representations can illuminate the considerations shaping their beliefs and reasoning. We present a causal language analysis pipeline that leverages a Large Language Model to identify causal claims in natural language documents, and aggregates claims across a corpus to produce a causal claim network. The pipeline then applies a clustering algorithm that groups causal claims according to their semantic topics. We demonstrate the pipeline by modeling causal belief systems surrounding the Covid-19 vaccine from tweets.

1 Introduction

Causal information facilitates learning (Holyoak and Cheng, 2011; Waldmann, 2007, 2017), and is crucial to how humans use and represent language (Mackie, 1980; Wolff et al., 2005; Lupyan, 2016). Causal relations are also ubiquitous in higher-level reasoning: they underlie our rich and flexible categories (Gelman and Legare, 2011), shape our explanatory preferences (Lombrozo and Vasilyeva, 2017), and structure our memories of events (Bower and Morrow, 1990).

Beliefs about causal relations can also have pernicious outcomes. For example, beliefs that vaccines cause autism are central to antivaccination attitudes (Horne et al., 2015; Powell et al., 2022), and the belief that liberal politicians have causal influence over the outcome of climate science research motivates climate change denialism (Cook and Lewandowsky, 2016). Because misinformation in online environments can spread rapidly to encourage these attitudes (Priniski et al., 2021; Priniski and Holyoak, 2022), new data science methods are necessary to combat these trends. However, data science algorithms generally struggle to advance a rigorous scientific understanding of psychological processes, as they provide correlational evidence that does not isolate cognitive mechanisms. Methodologists should aim to develop Natural Language Processing (NLP) algorithms that produce cognitively plausible data representations that researchers can utilize to guide explanatory understanding and motivate future interventions.

Because causal relations are the backbone of most higher-level reasoning processes in humans and are central to how we use language, developing systems that can isolate people's causal representations from language data is a natural place to start. However, NLP has historically struggled to identify instances of *psychological causality* (what a speaker *thinks* causes what) (Dunietz et al., 2017). This is because the variety of ways people communicate causality is immense (Talmy, 2000, 2011; Wolff, 2007), with most causal information latent in language inputs (Khoo et al., 2002; Blanco et al., 2008). Previously, methods that relied on hand labeling causal constructions to relate linguistic features to components of causal relations were extremely brittle and struggled to generalize to out-of-sample data (Yang et al., 2022). However, Large Language Models may help overcome this shortcoming as these models utilize rich semantic representations of language and sub-word tokenization that can help them identify instances of causal language not expressed in training (Devlin et al., 2018; Liu et al., 2019; Dunietz et al., 2017).

In addition to simply identifying instances of causal language, data representations should account for the breadth of people's conceptual systems in which a causal claim is made. For example, causal beliefs are not held in isolation; instead, people have rich interlocking belief systems that span multiple topics that shape the integration of evidence (Quine and Ullian, 1978; Priniski and Holyoak, 2022; Gelman and Legare, 2011). Previ-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 436–443
July 10-12, 2023 ©2023 Association for Computational Linguistics

ous methods for producing representations of people's belief systems rely on experiments and are slow to develop and may not generalize outside the lab (Powell, 2023; Powell et al., 2022). Because it is important to understand the full context of people's belief systems to reliably predict how people will interpret evidence and make decisions, tools must be designed that can identify the vast web of beliefs that people use to interpret information in the wild. NLP tools can take advantage of the proliferation of social data online to build these representations (Goldstone and Lupyan, 2016).

To this end, we introduce a pipeline based on the Large Language Model, RoBERTA-XL (Liu et al., 2019), that detects causal claims in text documents, and aggregates claims made across a corpus to produce a network of interlocking causal claims, or a *causal claim network* [1]. Causal claim networks can be used to approximate the beliefs and causal relations composing people's conceptual understanding of the entities and events discussed in a corpus. To guide future research, we host a pipeline that produces interactive visualizations of people's causal belief networks. We demonstrate this software by building causal belief systems surrounding Covid-19 vaccines from tweets.

2 How to build causal claim networks using our pipeline

The pipeline for extracting causal claim networks follows three main steps (see Figure 1). First, text documents are fed to a Large Language Model, a RoBERTa-XL transformer model (Liu et al., 2019), trained to extract causal claims made in a document (sentence to a paragraph in length). Second, the entities that compose causal claims are clustered according to their embeddings, clusters proxy *causal topics* (Grootendorst, 2022). Third, claims made across the corpus are coreferenced and strung together to make a network of cause and effect claims to be displayed in an interactive visualization.

We will now describe how a user could use our pipeline to build a causal claim network to visualize the causal claims made in a corpus of text documents. As shown in Figure 2, this follows two steps. First, a user uploads a .csv file containing the documents they wish to analyze. Documents should be a sentence to a paragraph in length and can range from tweets, journal entries, or news headlines. Next, the user selects which column in the dataframe contains the texts to be analyzed. A user can also specify if they want the pipeline to preprocess the documents and cluster the entities. It is worth noting that entity clustering works best when there are an abundance of causal claims about semantically distinct entities. If a user chooses to cluster claims and the pipeline does not produce an output, it does not mean that there are no causal claims present, but rather that there are no clear semantic clusters. In these cases, the users should deselected 'Cluster entities' and rerun the pipeline.

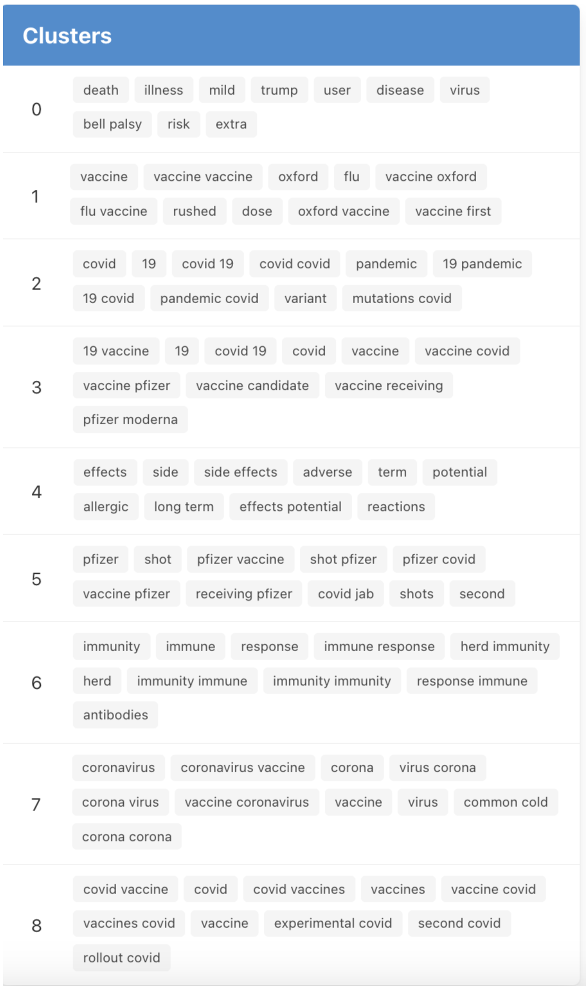
As seen in Figure 2, we analyze a data set of tweets about the Covid-19 vaccine with the file name *covid_tweets.csv*, and the column containing the tweet texts is titled *tweets*. We provide access to this dataset on the tool interface, which can be downloaded to replicate this tutorial.

Once the document file is uploaded and the user presses submit, the job is queued and causal claims will be extracted. As shown in Figure 3, a job status window will be populated and the user will be updated on the degree of completion. As a rough reference, extracting causal claims from about 6000 tweets takes about a minute to complete once the job begins.

Once the job is completed, the screen will be populated with the causal claim network, like the one in Figure 4. There are a few things worth highlighting here. First, each edge represents a single extracted causal claim in the corpus, and nodes are colored by their causal cluster, or topic (see Figure 5). Clusters proxy topics in the data set and can be interpreted as central *causal topics* in the data set. In the next section, we describe how we calculate clusters.

The causal claim network produced by the pipeline is interactive. A user can click on an edge to see the document and extracted causal claim that constitutes that edge (see Figure 6). Furthermore, as shown in Figure 7, a user can simplify the network by selecting to collapse the edges between the nodes. The edge thickness is proportional to the number of documents between those two clusters. To facilitate downstream analysis of causal claims (e.g., by analyzing sentiment or stance of causal claims), a user can download the edge list that produced the network as a .csv file. Columns in this .csv file include: cause word span, cause cluster, effect word span, effect cluster, text, and

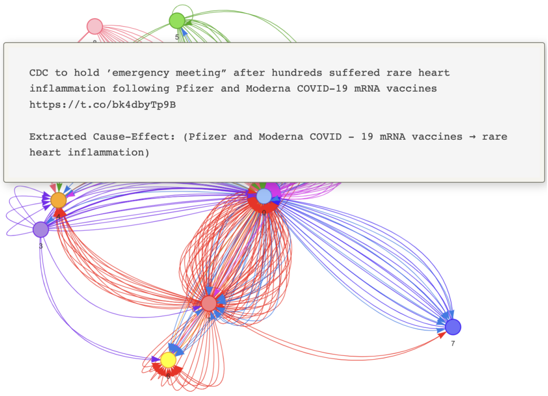
[1] We host the pipeline at the following link: https://mindsgpu02.isi.edu:5020. The code is available on GitHub: https://github.com/ishaanverma/causal-claims-pipeline.



Clusters

0	death, illness, mild, trump, user, disease, virus, bell palsy, risk, extra
1	vaccine, vaccine vaccine, oxford, flu, vaccine oxford, flu vaccine, rushed, dose, oxford vaccine, vaccine first
2	covid, 19, covid 19, covid covid, pandemic, 19 pandemic, 19 covid, pandemic covid, variant, mutations covid
3	19 vaccine, 19, covid 19, covid, vaccine, vaccine covid, vaccine pfizer, vaccine candidate, vaccine receiving, pfizer moderna
4	effects, side, side effects, adverse, term, potential, allergic, long term, effects potential, reactions
5	pfizer, shot, pfizer vaccine, shot pfizer, pfizer covid, vaccine pfizer, receiving pfizer, covid jab, shots, second
6	immunity, immune, response, immune response, herd immunity, herd, immunity immune, immunity immunity, response immune, antibodies
7	coronavirus, coronavirus vaccine, corona, virus corona, corona virus, vaccine coronavirus, vaccine, virus, common cold, corona corona
8	covid vaccine, covid, covid vaccines, vaccines, vaccine covid, vaccines covid, vaccine, experimental covid, second covid, rollout covid

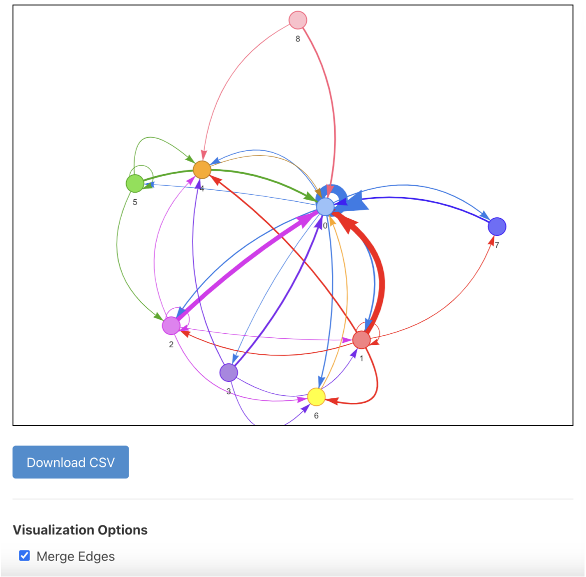
Figure 5: Causal clusters, or causal topics, are shown to the right of the produced causal claim network. Each topic consists of a set of keywords that describes the cluster.



CDC to hold ‘emergency meeting’ after hundreds suffered rare heart inflammation following Pfizer and Moderna COVID-19 mRNA vaccines https://t.co/bk4dbyTp9B

Extracted Cause-Effect: (Pfizer and Moderna COVID - 19 mRNA vaccines → rare heart inflammation)

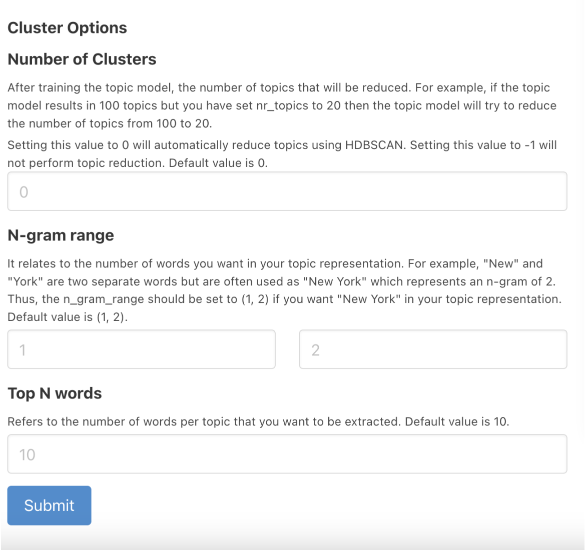
Figure 6: Hovering over an edge in the causal claim network displays the document and extracted causal claim that constitutes that edge. The document is shown at the top of the box, and the extracted cause claim is at the bottom.



Visualization Options

☑ Merge Edges

Figure 7: Causal claim network with merged edges, where edge weights equates to the number of documents linking two clusters. Merging edges is useful to quickly assess degree of linkage between causal clusters (nodes) in the network.



Cluster Options

Number of Clusters

After training the topic model, the number of topics that will be reduced. For example, if the topic model results in 100 topics but you have set nr_topics to 20 then the topic model will try to reduce the number of topics from 100 to 20.

Setting this value to 0 will automatically reduce topics using HDBSCAN. Setting this value to -1 will not perform topic reduction. Default value is 0.

0

N-gram range

It relates to the number of words you want in your topic representation. For example, "New" and "York" are two separate words but are often used as "New York" which represents an n-gram of 2. Thus, the n_gram_range should be set to (1, 2) if you want "New York" in your topic representation. Default value is (1, 2).

1 2

Top N words

Refers to the number of words per topic that you want to be extracted. Default value is 10.

10

Figure 8: A user can specify parameters when running the pipeline to engage with exploratory data analysis. Users can specify the number of clusters, the n-gram range used during processing, and set the number of words to describe each topic.

	Precision	Recall	F1
SCITE	0.833	0.858	0.845
BERT	0.824	0.858	0.841
RoBERTa-XL	**0.883**	**0.865**	**0.874**

Table 1: Model performance on the causal relation identification task (Hendrickx et al., 2010). The RoBERTa-XL model demonstrates increased performance over the smaller transformer BERT and previously reported state-of-the-art implementations (Li et al., 2021)

transformer (Devlin et al., 2018). We therefore used the RoBERTa-XL transformer in our pipeline.

Once the sentences have been tagged using the causality tagging scheme, we run the tag2triplet algorithm proposed by Li et al., 2021 to extract the cause-effect tuples from the tagged sequence. The algorithm operates by first identifying the in-degree and out-degree of causality in the tagged sequence. Here, if the entity is labeled as a "cause", then the out-degree is increased by 1; if the entity is labeled as an "effect" then the in-degree is incremented by 1; and if the entity is labeled as "embedded causality" then both the in-degree and out-degree are incremented by 1. The algorithm then tries to align the identified entities such that each entity that has an outgoing edge (i.e., the cause) is joined with the entity that has an incoming edge (i.e., the effect) while taking into consideration the distance between the entities in the document and whether they contain a coordinating conjunction.

3.2 Step 2: Finding causal topics by clustering embeddings

Clusters of causes and effects proxy topics in the causal claim network. We cluster the embeddings of the nodes by extending the tf-idf measure of the embeddings (Grootendorst, 2022). This method was originally developed to cluster BERT representations to uncover topics in a corpus, but we implemented the algorithm to cluster RoBERTa-XL embeddings. This allows users to assess latent structure in the causal claims expressed in a corpus and simplifies the resulting causal claim graph by mapping semantically similar claims to a common node.

3.3 Constructing a causal claim network

Extracted cause-effect tuples serve as directed edges in the causal claim network, which are strung together throughout the corpus to form a causal claim network. Nodes are the identified cause and effect wordspans and the weighted edges encode the number of instances in the corpus where node i was said to cause node j. The edge direction encodes the direction of the causal relation and can be supplied with additional semantic content (e.g., relational vectors, sentiment).

4 Case study: Building a network of causal claims about the Covid-19 vaccine from tweets

4.1 Data set of tweets

To test this pipeline, we build a causal claim network using a set of 6000 tweets about the Covid-19 vaccine (Poddar et al., 2022). The original dataset was curated by subsetting a larger sample of tweets from before and after the release of the Covid-19 vaccines.

4.2 Pipeline results

The pipeline returns 408 extracted causal claims belonging to nine distinct clusters (see Figure 5). The clusters are, as expected, about the various Covid-19 vaccines and their anticipated consequences. By aggregating the keywords for each cluster, we can define the set of causal topics returned by the pipeline. More specifically, cluster 0 contains keywords related to *Death*; 1: *Oxford vaccines*, 2: *Covid-19 pandemic*; 3: *Pfizer vaccine*, 4: *Side-effects*, 5: *Pfizer shot*, 6: *Immunity and antibodies*, 7: *Coronavirus*, and 8: *Covid vaccine*. As shown in Table 2, we see that the clusters are about a range of topics with varying semantics and valence, which suggests that the pipeline can help us understand the breadth of considerations guiding Twitter discussions about the Covid-19 vaccine.

4.3 Secondary analyses of extracted causal claims

By analyzing the causal claims returned by the pipeline (which is also available for download as a .csv file), we can explore how these causal clusters are linked to one another. For example, as shown in 2, some of the clusters are more commonly composed of word spans denoting cause events, while others are more composed of word spans denoting effect events.

5 Related work

Although our approach is domain-general (documents do not need to belong to a single issue or topic for the pipeline to return a set of causal

Cluster	Topic	Causes	Effects
0	Death	154	299
1	Oxford vaccine	108	15
2	Pandemic	56	16
3	Pfizer vaccine	25	1
4	Side-effects	1	25
5	Pfzier shot	22	2
6	Immunity	4	29
7	Coronavirus	13	8
8	Covid vaccine	17	0

Table 2: Number of identified word spans per each causal cluster. The topic label is determined by assessing the top keywords in each causal cluster. Each cluster has a different distribution of cause and effect spans.

claims), we demonstrate the use of our pipeline modeling causal claims about the Covid-19 vaccine. Previous work has developed systems specifically designed to analyze claims about Covid-19. For example, Li et al. (2022) built a system specifically designed to monitor claims made about Covid-19. This system identifies claims and arguments made in the corpus, and sources additional Wikidata information to put the claims in a richer content.

Mining causal claim networks requires isolating causal claims which oftentimes constitute arguments expressed in a text document: a reasoner makes a causal claim when explaining a mechanism (Lombrozo and Vasilyeva, 2017) or argument. Claim detection is an active area of research (Palau and Moens, 2009; Goudas et al., 2014), as is the detection of the components of arguments in text (Sardianos et al., 2015). Because causal reasoning is central to the way people construct arguments (Abend et al., 2013), understanding how people posit causal claims can shed light on the types of arguments people will endorse related to that issue.

Most claim detection algorithms work on the level of single documents, but approaches such as those of Levy et al. 2014 propose corpus-wide claim detection. Our pipeline utilizes a mixing of the two: claims are detected within a document and then aggregated across the documents in the corpus to provide a corpus-level representation.

6 Conclusions and future work

Interactive data visualization is an effective way for people to make sense of complex data (Janvrin et al., 2014) and can be an effective tool to guide scientific thinking (Franconeri et al., 2021). Our pipeline is designed to help researchers explore the causal claims expressed in a corpus through interactive exploration.

There are therefore many applications of this tool to the study of human reasoning and belief change, and future work will test the efficacy of these use cases. For example, researchers in cognitive science have worked on developing methods to measure people's rich conceptual systems about vaccines (Powell, 2023). These methods often require the development of surveys that measure people's attitudes toward a variety of related issues. Causal claim networks can give researchers a starting place to know what measures they should include in these surveys.

In future work, we will work on expanding the visualization tool to include features that allow for richer forms of interaction. For example, by allowing users to build subnetworks based on another data attribute (e.g., stance of the document, expressed sentiment), to allow for comparisons across networks. Related to this, future work will also develop quantitative measures of divergence across networks.

Acknowledgements

This work was funded in part by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract No. W911NF-21-C-0002 and AFOSR MURI grant number FA9550-22-1-0380.

References

Gabriel Abend, Caitlin Petre, and Michael Sauder. 2013. Styles of causal thought: An empirical investigation. *American Journal of Sociology*, 119(3):602–654.

Eduardo Blanco, Nuria Castell, and Dan Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Gordon H Bower and Daniel G Morrow. 1990. Mental models in narrative comprehension. *Science*, 247(4938):44–48.

John Cook and Stephan Lewandowsky. 2016. Rational irrationality: Modeling climate change belief polarization using bayesian networks. *Topics in Cognitive Science*, 8(1):160–179.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.