What we talk about when we talk about causality: Features of causal statements across large-scale social discourse

Thomas C. McAndrew*†, Joshua C. Bongard‡†, Christopher M. Danforth*†, Peter S. Dodds*†,
Paul D. H. Hines§† and James P. Bagrow*†
*Department of Mathematics and Statistics, University of Vermont, Burlington, VT, United States
†Vermont Complex Systems Center, University of Vermont
‡Department of Computer Science, University of Vermont
§School of Engineering, University of Vermont

***Abstract*—Identifying and communicating relationships between causes and effects is important for understanding our world, but is affected by language structure, cognitive and emotional biases, and the properties of the communication medium. Despite the increasing importance of social media, much remains unknown about causal statements made online. To study real-world causal attribution, we extract a large-scale corpus of causal statements made on the Twitter social network platform as well as a comparable random control corpus. We compare causal and control statements using statistical language and sentiment analysis tools. We find that causal statements have a number of significant lexical and grammatical differences compared with controls and tend to be more negative in sentiment than controls. Causal statements made online tend to focus on news and current events, medicine and health, or interpersonal relationships, as shown by topic models. By quantifying the features and potential biases of causality communication, this study improves our understanding of the accuracy of information and opinions found online.**

***Keywords*—social media; online social network; causal attribution; natural language processing**

I. Introduction

Social media and online social networks now provide vast amounts of data on human online discourse and other activities [1], [2], [3], [4], [5], [6], [7]. With so much communication taking place online and with social media being capable of hosting powerful misinformation campaigns [8] such as those claiming vaccines cause autism [9], [10], it is more important than ever to better understand the discourse of causality and the interplay between online communication and the statement of cause and effect.

Causal inference is a crucial way that humans comprehend the world, and it has been a major focus of philosophy, statistics, mathematics, psychology, and the cognitive sciences. Philosophers such as Hume and Kant have long argued whether causality is a human-centric illusion or the discovery of a priori truth [11], [12]. Causal inference in science is incredibly important, and researchers have developed statistical measures such as Granger causality [13], mathematical and probabilistic frameworks [14], [15], [16], [17], and text mining procedures [18], [19], [20] to better infer causal influence from data. In the cognitive sciences, the famous perception experiments of Michotte *et al.* led to a long line of research exploring the cognitive biases that humans possess when attempting to link cause and effect [21], [22], [23].

How humans understand and communicate cause and effect relationships is complicated, and is influenced by language structure [24], [25], [26], [27] and sentiment or valence [28]. A key finding is that the perceived emphasis or causal weight changes between the agent (the grammatical construct responsible for a cause) and the patient (the construct effected by the cause) depending on the types of verbs used to describe the cause and effect. Researchers have hypothesized [29] that this is because of the innate weighting property of the verbs in the English language that humans use to attribute causes and effects. Another finding is the role of a valence bias: the volume and intensity of causal reasoning may increase due to negative feedback or negative events [28].

Despite these long lines of research, causal attributions made via social media or online social networks have not been well studied. The goal of this paper is to explore the language and topics of causal statements in a large corpus of social media taken from Twitter. We hypothesize that language and sentiment biases play a significant role in these statements, and that tools from natural language processing and computational linguistics can be used to study them. We do not attempt to study the factual correctness of these statements or offer any degree of verification, nor do we exhaustively identify and extract all causal statements from these data. Instead, here we focus on statements that are with high certainty causal statements, with the goal to better understand key characteristics about causal statements that differ from everyday online communication.

The rest of this paper is organized as follows: In Sec. II we discuss our materials and methods, including the dataset we studied, how we preprocessed that data and extracted a 'causal' corpus and a corresponding 'control' corpus, and the details of the statistical and language analysis tools we studied these corpora with. In Sec. III we present results using these tools to compare the causal statements to control statements.

We conclude with a discussion in Sec. IV.

II. Materials and Methods

Dataset, filtering, and corpus selection

Data was collected from a 10% uniform sample of Twitter posts made during 2013, specifically the Gardenhose API. Twitter activity consists of short posts called tweets which are limited to 140 characters. Retweets, where users repost a tweet to spread its content, were not considered. (The spread of causal statements will be considered in future work.) We considered only English-language tweets for this study. To avoid cross-language effects, we kept only tweets with a user-reported language of 'English' and, as a second constraint, individual tweets needed to match more English stopwords than any other language's set of stopwords. Stopwords considered for each language were determined using NLTK's database [30]. A tweet will be referred to as a 'document' for the rest of this work.

All document text was processed the same way. Punctuation, XML characters, and hyperlinks were removed, as were Twitter-specific "at-mentions" and "hashtags" (see also the Appendix). There is useful information here, but it is either not natural language text, or it is Twitter-specific, or both. Documents were broken into individual words (unigrams) on whitespace. Casing information was retained, as we will use it for our Named Entity analysis, but otherwise all words were considered lowercase only (see also the Appendix). Stemming [31] and lemmatization [32] were not performed.

Causal documents were chosen to contain one occurrence only of the exact unigrams: 'caused', 'causing', or 'causes'. The word 'cause' was not included due to its use as a popular contraction for 'because'. One 'cause-word' per document restricted the analysis to single relationships between two relata. Documents that contain *bidirectional* words ('associate', 'relate', 'connect', 'correlate', and any of their stems) were also not selected for analysis. This is because our focus is on causality, an inherently one-sided relationship between two objects. We also did not consider additional synonyms of these cause words, although that could be pursued for future work. *Control documents* were also selected. These documents did not contain any of 'caused', 'causing', or 'causes', nor any bidirectional words, and are further matched temporally to obtain the same number of control documents as causal documents in each fifteen-minute period during 2013. Control documents were otherwise selected randomly; causal synonyms may be present. The end result of this procedure identified 965,560 causal and 965,560 control documents. Each of the three "cause-words", 'caused', 'causes', and 'causing' appeared in 38.2%, 35.0%, and 26.8% of causal documents, respectively.

Tagging and corpus comparison

Documents were further studied by annotating their unigrams with **Parts-of-Speech** (POS) and **Named Entities** (NE) tags. POS tagging was done using NLTK v3.1 [30] which implements an averaged perceptron classifier [33] trained on the Brown Corpus [34]. (POS tagging is affected by punctuation; we show in the Appendix that our results are relatively robust to the removal of punctuation.) POS tags denote the nouns, verbs, and other grammatical constructs present in a document. Named Entity Recognition (NER) was performed using the 4-class, distributional similarity tagger provided as part of the Stanford CoreNLP v3.6.0 toolkit [35]. NER aims to identify and classify proper words in a text. The NE classifications considered were: Organization, Location, Person, and Misc. The Stanford NER tagger uses a conditional random field model [36] trained on diverse sets of manually-tagged English-language data (CoNLL-2003) [35]. Conditional random fields allow dependencies between words so that 'New York' and 'New York Times', for example, are classified separately as a location and organization, respectively. These taggers are commonly used and often provide reasonably accurate results, but there is always potential ambiguity in written text and improving upon these methods remains an active area of research.

Comparing corpora: Unigrams, POS, and NEs were compared between the cause and control corpora using **odds ratios** (ORs):

$$\mathrm{OR}(x) = \frac{p_C(x)/(1-p_C(x))}{p_N(x)/(1-p_N(x))}, \tag{1}$$

where $p_C(x)$ and $p_N(x)$ are the probabilities that a unigram, POS, or NE x occurs in the causal and control corpus, respectively. These probabilities were computed for each corpus separately as $p(x) = f(x)/\sum_{x' \in V} f(x')$, where $f(x)$ is the total number of occurrences of x in the corpus and V is the relevant set of unigrams, POS, or NEs. Confidence intervals for the ORs were computed using Wald's methodology [37].

As there are many unique unigrams in the text, when computing unigram ORs we focused on the most meaningful unigrams within each corpus by using the following filtering criteria: we considered only the ORs of the 1500 most frequent unigrams in that corpus that also have a term-frequency-inverse-document-frequency (tf-idf) score above the 90th percentile for that corpus [38]. The tf-idf was computed as

$$\text{tf-idf}(w) = \log f(w) \times \log\left(\frac{D}{df(w)}\right), \tag{2}$$

where D is the total number of documents in the corpus, and $df(w)$ is the number of documents in the corpus containing unigram w. Intuitively, unigrams with higher tf-idf scores appear frequently, but are not so frequent that they are ubiquitous through all documents. Filtering via tf-idf is standard practice in the information retrieval and data mining fields.

Cause-trees

For a better understanding of the higher-order language structure present in text phrases, *cause-trees* were constructed. A cause-tree starts with a root cause word (either 'caused', 'causing' or 'causes'), then the two most probable words following (preceding) the root are identified. Next, the root word plus one of the top probable words is combined into

a bigram and the top two most probable words following (preceding) this bigram are found. Repeatedly applying this process builds a binary tree representing the n-grams that begin with (terminate at) the root word. This process can continue until a certain n-gram length is reached or until there are no more documents long enough to search.

Sentiment analysis

Sentimental analysis was applied to estimate the emotional content of documents. Two levels of analysis were used: a method where individual unigrams were given crowdsourced numeric sentiment scores, and a second method involving a trained classifier that can incorporate document-level phrase information.

For the first sentiment analysis, each unigram w was assigned a crowdsourced "labMT" sentiment score $s(w)$ [6]. (Unlike [6], scores were recentered by subtracting the mean, $s(w) \leftarrow s(w) - \langle s \rangle$.) Unigrams determined by volunteer raters to have a negative emotional sentiment ('hate','death', etc.) have $s(w) < 0$, while unigrams determined to have a positive emotional sentiment ('love', 'happy', etc.) tend to have $s(w) > 0$. Unigrams that have labMT scores and are above the 90th percentile of tf-idf for the corpus form the set $\tilde{V}$. (Unigrams in $\tilde{V}$ need not be among the 1500 most frequent unigrams.) The set $\tilde{V}$ captures 87.9% (91.5%) of total unigrams in the causal (control) corpus. Crucially, the tf-idf filtering ensures that the words 'caused', 'causes', and 'causing', which have a slight negative sentiment, are not included and do not introduce a systematic bias when comparing the two corpora.

This sentiment measure works on a per-unigram basis, and is therefore best suited for large bodies of text, not short documents [6]. Instead of considering individual documents, the distributions of labMT scores over all unigrams for each corpus was used to compare the corpora. In addition, a **single sentiment score** for each corpus was computed as the average sentiment score over all unigrams in that corpus, weighed by unigram frequency: $\sum_{w \in \tilde{V}} f(w)s(w) \Big/ \sum_{w' \in \tilde{V}} f(w')$.

To supplement this sentiment analysis method, we applied a second method capable of estimating with reasonable accuracy the sentiment of individual documents. We used the sentiment classifier [39] included in the Stanford CoreNLP v3.6.0 toolkit to documents in each corpus. Documents were individually classified into one of five categories: very negative, negative, neutral, positive, very positive. The data used to train this classifier is taken from positive and negative reviews of movies (Stanford Sentiment Treebank v1.0) [39].

Topic modeling

Lastly, we applied topic modeling to the causal corpus to determine what are the topical foci most discussed in causal statements. Topics were built from the causal corpus using Latent Dirichlet Allocation (LDA) [40]. Under LDA each document is modeled as a bag-of-words or unordered collection of unigrams. Topics are considered as mixtures of unigrams by estimating conditional distributions over unigrams: $P(w|T)$, the probability of unigram w given topic T and documents are considered as mixtures of topics via $P(T|d)$, the probability of topic T given document d. These distributions are then found via statistical inference given the observed distributions of unigrams across documents. The total number of topics is a parameter chosen by the practitioner. For this study we used the MALLET v2.0.8RC3 topic modeling toolkit [41] for model inference. By inspecting the most probable unigrams per topic (according to $P(w|T)$), we found 10 topics provided meaningful and distinct topics.

III. Results

We have collected approximately 1M causal statements made on Twitter over the course of 2013, and for a control we gathered the same number of statements selected at random but controlling for time of year (see Methods). We applied **Parts-of-Speech** (POS) and **Named Entity** (NE) taggers to all these texts. Some post-processed and tagged example documents, both causal and control, are shown in Fig. 1A. We also applied sentiment analysis methods to these documents (Methods) and we have highlighted very positive and very negative words throughout Fig. 1.

In Fig. 1B we present odds ratios for how frequently unigrams (words), POS, or NE appear in causal documents relative to control documents. The three unigrams most strongly skewed towards causal documents were 'stress', 'problems', and 'trouble', while the three most skewed towards control documents were 'photo', 'ready', and 'cute'. While these are only a small number of the unigrams present, this does imply a negative sentiment bias among causal statements (we return to this point shortly).

Figure 1B also presents odds ratios for POS tags, to help us measure the differences in grammatical structure between causal and control documents (see also the Appendix for the effects of punctuation and casing on these odds ratios). The causal corpus showed greater odds for plural nouns (Penn Treebank tag: NNS), plural proper nouns (NNPS), Wh-determiners/pronouns (WDT, WP$) such as 'whichever','whatever', 'whose', or 'whosever', and predeterminers (PDT) such as 'all' or 'both'. Predeterminers quantify noun phrases such as 'all' in 'after *all* the events that caused you tears', showing that many causal statements, despite the potential brevity of social media, can encompass or delineate classes of agents and/or patients. On the other hand, the causal corpus has lower odds than the control corpus for list items (LS), proper singular nouns (NNP), and interjections (UH).

Lastly, Fig. 1B contains odds ratios for NE tags, allowing us to quantify the types of proper nouns that are more or less likely to appear in causal statements. Of the four tags, only the "Person" tag is less likely in the causal corpus than the control. (This matches the odds ratio for the proper singular noun discussed above.) Perhaps surprisingly, these results together imply that causal statements are less likely to involve individual persons than non-causal statements. There is considerable celebrity news and gossip on social media [5]; discussions of celebrities may not be especially focused on attributing causes

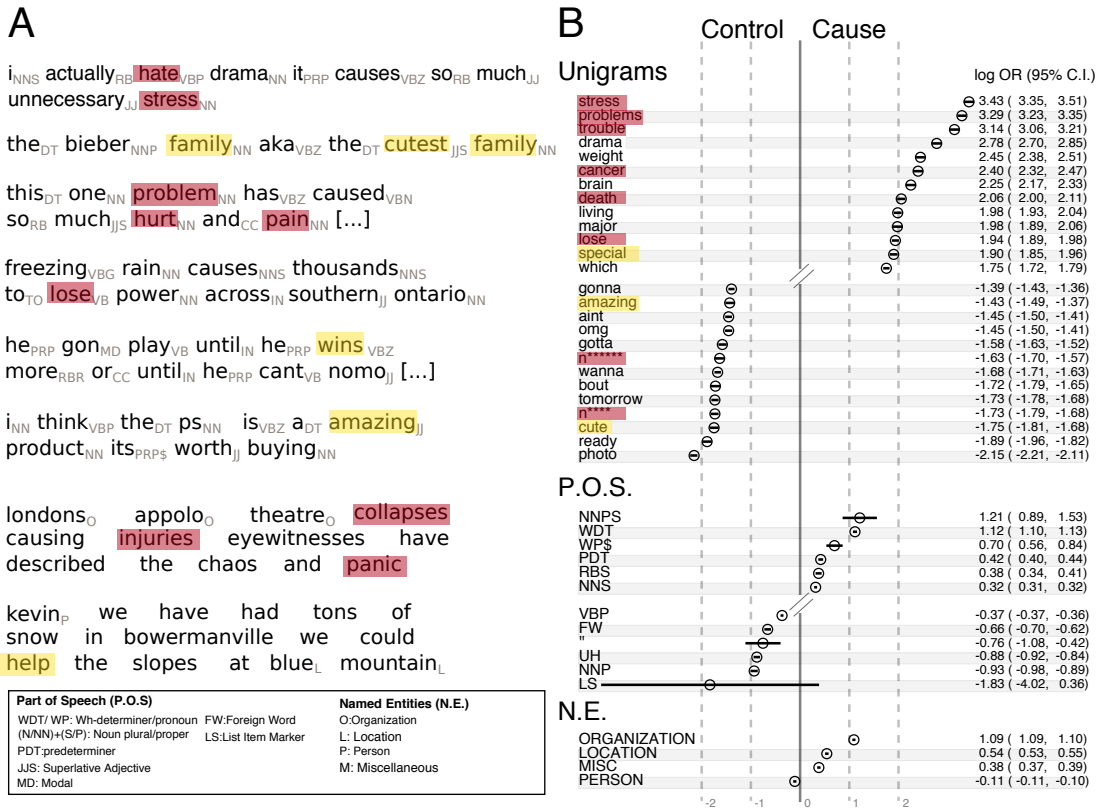

Fig. 1. Measuring the differences between causal and control documents. (**A**) Examples of processed documents tagged by Parts-of-Speech (POS) or Named Entities (NEs). Unigrams highlighted in red (yellow) are in the bottom 10% (top 10%) of the labMT sentiment scores. (**B**) Log Odds ratios with 95% Wald confidence intervals for the most heavily skewed unigrams, POS, and all NEs between the causal and control corpus. POS tags that are plural and use Wh-pronouns (that, what, which, ...) are more common in the causal corpus, while singular nouns and list items are more common in the controls. Finally, the 'Person' tag is the only NE less likely in the causal corpus. Certain unigrams were censored for presentation only, not analysis. All shown odds ratios were significant at the $\alpha = 0.05$ level except LS (List item markers). See also the Appendix.

to these celebrities. All other NE tags, Organization, Location, and Miscellaneous, occur more frequently in the causal corpus than the control. All the odds ratios in Fig. 1B were significant at the $\alpha = 0.05$ level except the List item marker (LS) POS tag.

The unigram analysis in Fig. 1 does not incorporate higher-order phrase structure present in written language. To explore these structures specifically in the causal corpus, we constructed "cause-trees", shown in Fig. 2. Inspired by association mining [42], a cause-tree is a binary tree rooted at either 'caused', 'causes', or 'causing', that illustrates the most frequently occurring n-grams that either begin or end with that root cause word (see Methods for details).

The "causes" tree shows the focused writing (sentence segments) that many people use to express either the relationship between their own actions and a cause-and-effect ("even if it causes"), or the uncontrollable effect a cause may have on themselves: "causes me to have" shows a person's inability to control a causal event ("[...] i have central heterochromia which causes me to have dual colors in both eyes"). The 'causing' tree reveals our ability to confine causal patterns to specific areas, and also our ability to be affected by others causal decisions. Phrases like "causing a scene in/at" and "causing a ruckus in/at" (from documents like "causing a ruckus in the hotel lobby typical [...]") show people commonly associate bounds on where causal actions take place. The causing tree also shows people's tendency to emphasize current negativity: Phrases like "pain this is causing" coming from documents like "cant you see the pain you are causing her" supports the sentiment bias that causal attribution is more likely for negative cause-effect associations. Finally, the 'caused' tree focuses heavily on negative events and indicates people are more likely to remember negative causal events. Documents with phrases from the caused tree ("[...] appalling tragedy [...] that caused the death", "[...] live with this pain that you caused when i was so young [...]") exemplify the negative events that are focused on are large-scale tragedies or very personal negative events in one's life.

Taken together, the popularity of negative sentiment unigrams (Fig. 1) and n-grams (Fig. 2) among causal documents shows that emotional sentiment or "valence" may play a role in how people perform causal attribution [28]. The "if it bleeds, it leads" mentality among news media, where violent and negative news are more heavily reported, may appeal to this innate causal association mechanism. (On the other hand, many news media themselves use social media for

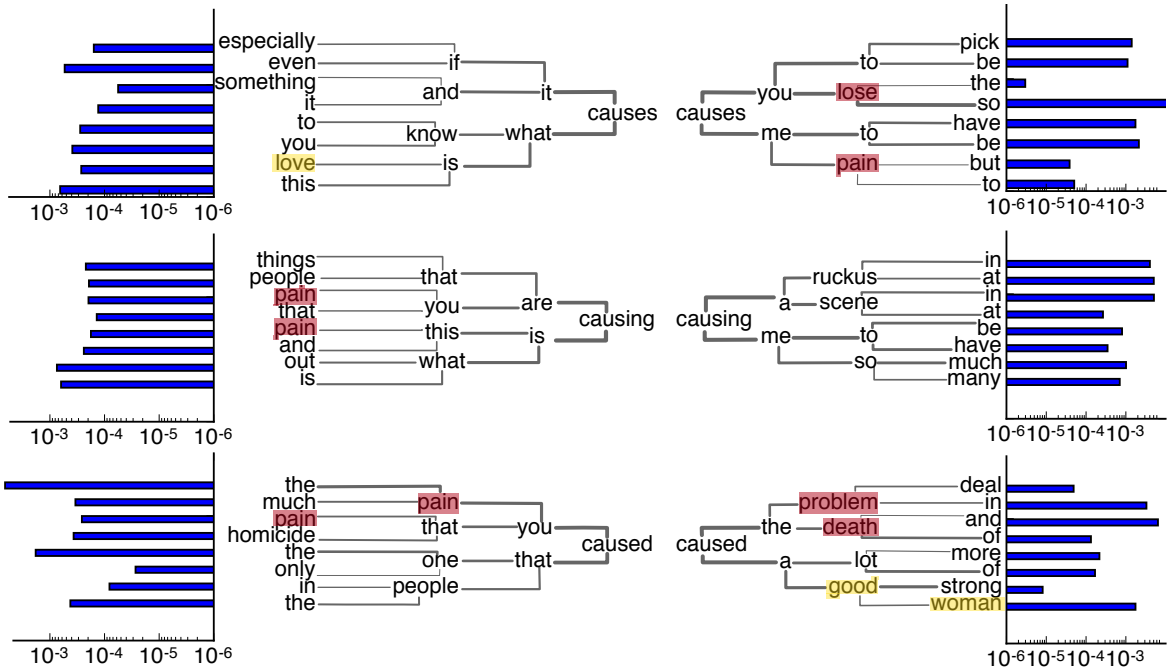

Fig. 2. "Cause-trees" containing the most probable n-grams terminating at (left) or beginning with (right) a chosen root cause-word (see Methods). Line widths are log proportional to their corresponding n-gram frequency and bar plots measure the 4-gram per-document rate $N(4\text{-gram})/D$. Most trees express negative sentiment consistent with the unigram analysis (Fig. 1). The 'causes' tree shows (i) people think in terms of causal probability ("you know what causes [...]"), and (ii) people use causal language when they are directly affected or being affected by another ("causes you", "causes me"). The 'causing' tree is more global ("causing a ruckus/scene") and ego-centric ("pain you are causing"). The 'caused' tree focuses on negative sentiment and alludes to humans retaining negative causal thoughts in the past.

reporting.) The prevalence of negative sentiment also contrasts with the "better angels of our nature" evidence of Pinker [43], illustrating one bias that shows why many find the results of Ref. [43] surprising.

Given this apparent sentiment skew, we further studied sentiment (Fig. 3). We compared the sentiment between the corpora in four different ways to investigate the observation (Figs. 1B and 2) that people focus more about negative concepts when they discuss causality. First, we computed the mean sentiment score of each corpus using crowdsourced "labMT" scores weighted by unigram frequency (see Methods). We also applied tf-idf filtering (Methods) to exclude very common words, including the three cause-words, from the mean sentiment score. The causal corpus text was slightly negative on average while the control corpus was slightly positive (Fig. 3A). The difference in mean sentiment score was significant (t-test: $p < 0.01$).

Second, we moved from the mean score to the distribution of sentiment across all (scored) unigrams in the causal and control corpora (Fig. 3B). The causal corpus contained a large group of negative sentiment unigrams, with labMT scores in the approximate range $-3 < s < -1/2$; the control corpus had significantly fewer unigrams in this score range.

Third, in Fig. 3C we used POS tags to categorize scored unigrams into nouns, verbs, and adjectives. Studying the distributions for each, we found that nouns explain much of the overall difference observed in Fig. 3B, with verbs showing a similar but smaller difference between the two corpora. Adjectives showed little difference. The distributions in Fig. 3C account for 87.8% of scored text in the causal corpus and 77.2% of the control corpus. The difference in sentiment between corpora was significant for all distributions (t-test: $p < 0.01$).

Fourth, to further confirm that the causal documents tend toward negative sentiment, we applied a separate, independent sentiment analysis using the Stanford NLP sentiment toolkit [39] to classify the sentiment of individual documents not unigrams (see Methods). Instead of a numeric sentiment score, this classifier assigns documents to one of five categories ranging from very negative to very positive. The classifier showed that the causal corpus contains more negative and very negative documents than the control corpus, while the control corpus contains more neutral, positive, and very positive documents (Fig. 3D).

We have found language (Figs. 1 and 2) and sentiment (Fig. 3) differences between causal statements made on social media compared with other social media statements. But *what* is being discussed? What are the topical foci of causal statements? To study this, for our last analysis we applied topic models to the causal statements. Topic modeling finds groups of related terms (unigrams) by considering similarities between how those terms co-occur across a set of documents.

We used the popular topic modeling method Latent Dirichlet Allocation (LDA) [40]. We ranked unigrams by how strongly associated they were with the topic. Inspecting these unigrams we found that a 10-topic model discovered meaningful topics.

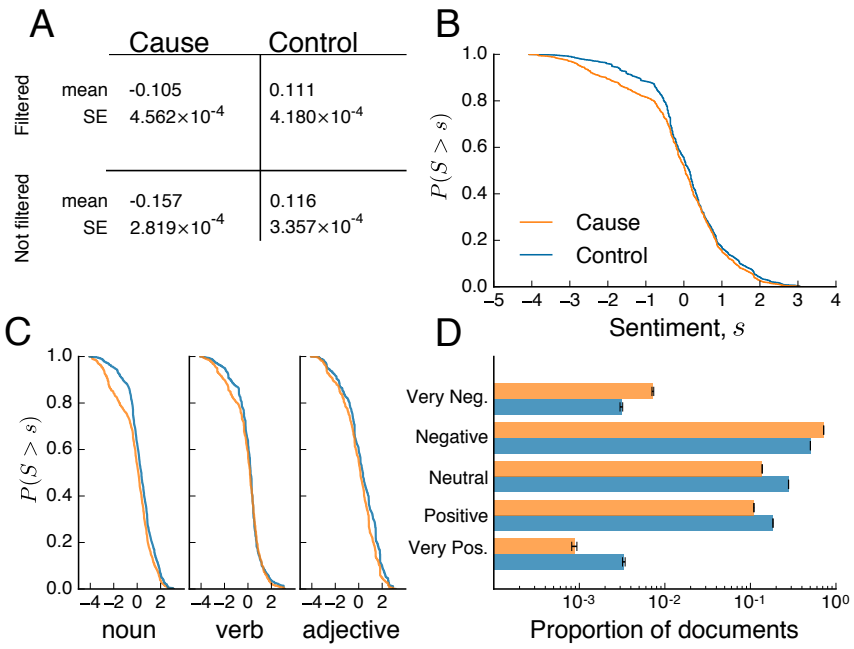

Fig. 3. Sentiment analysis revealed differences between the causal and control corpora. (**A**) The mean unigram sentiment score (see Methods), computed from crowdsourced "labMT" scores [6], was more negative for the causal corpus than for the control. This held whether or not tf-idf filtering was applied. (**B**) The distribution of unigram sentiment scores for the two corpora showed more negative unigrams (with scores in the approximate range $-3 < s < -1/2$) in the causal corpus compared with the control corpus. (**C**) Breaking the sentiment distribution down by Parts-of-Speech, nouns show the most pronounced difference in sentiment between cause and control; verbs and adjectives are also more negative in the causal corpus than the control but with less of a difference than nouns. POS tags corresponding to nouns, verbs, and adjectives together account for 87.8% and 77.2% of the causal and control corpus text, respectively. (**D**) Applying a different sentiment analysis tool—a trained sentiment classifier [39] that assigns individual documents to one of five categories—the causal corpus had an overabundance of negative sentiment documents and fewer positive sentiment documents than the control. This shift from very positive to very negative documents further supports the tendency for causal statements to be negative.

See Methods for full details. The top unigrams for each topic are shown in Tab. I.

Topics in the causal corpus tend to fall into three main categories: (i) news, covering current events, weather, etc.; (ii) medicine and health, covering cancer, obesity, stress, etc.; and (iii) relationships, covering problems, stress, crisis, drama, sorry, etc.

While the topics are quite different, they are all similar in their use of negative sentiment words. The negative/global features in the 'news' topic are captured in the most representative words: damage, fire, power, etc. Similar to news, the 'accident' topic balances the more frequent day-to-day minor frustrations with the less frequent but more severe impacts of car accidents. The words 'traffic' and 'delays' are the most probable words for this topic, and are common, low-impact occurrences. On the contrary, 'crash', 'car', 'accident' and 'death' are the next most probable words for the accident topic, and generally show a focus on less-common but higher-impact events.

The 'medical' topic also focused on negative words; highly probable words for this topic included 'cancer', 'break', 'disease', 'blood', etc. Meanwhile, the 'body' topic contained words like: 'stress', 'lose', and 'weight', giving a focus on on our more personal struggles with body image. Besides body image, the 'injuries' topic uses specific pronouns ('his', 'him', 'her') in references to a person's own injuries or the injuries of others such as athletes.

Aside from more factual information, social information is well represented in causal statements. The 'problems' topic shows people attribute their problems to many others with terms like: 'dont', 'people', 'they', 'them'. The 'stress' topic also uses general words such as 'more', 'than', or 'people' to link stress to all people, and in the same vein, the 'crisis' topic focuses on problems within organizations such as governments. The 'drama' and 'sorry' topics tend towards more specific causal statements. Drama used the words: 'like', 'she', and 'her' while documents in the sorry topic tended to address other people.

The topics of causal documents discovered by LDA showed that both general and specific statements are made regarding news, medicine, and relationships when individuals make causal attributions online.

IV. Discussion

The power of online communication is the speed and ease with which information can be propagated by potentially any connected users. Yet these strengths come at a cost: rumors and misinformation also spread easily. Causal misattribution is at the heart of many rumors, conspiracy theories, and misinformation campaigns.

Given the central role of causal statements, further studies of the interplay of information propagation and online causal attributions are crucial. Are causal statements more likely to

“News”	“Accident”	“Problems”	“Medical”	“Crisis”	“Sorry”	“Stress”	“Body”	“Drama”	“Injuries”
damage	traffic	dont	cancer	their	any	more	stress	like	his
fire	delays	people	break	our	never	than	lose	she	him
power	crash	they	some	from	been	being	weight	her	out
via	car	problems	men	how	sorry	over	stuff	lol	back
new	accident	why	can	about	there	person	living	out	her
news	death	about	disease	social	know	sleep	quickly	trouble	when
from	between	when	from	crisis	will	which	special	good	head
says	after	know	most	via	ive	people	proof	now	into
after	year	them	our	great	they	one	diets	sh*t	off
video	down	like	others	money	out	stress	excercise	life	well
global	there	drama	loss	many	problems	someone	f*ck	twitter	which
rain	man	who	heart	issues	can	makes	who	got	from
warming	due	one	health	should	now	think	giving	scene	down
water	from	youre	food	war	trouble	when	love	get	game
explosion	snow	get	symptoms	problems	see	most	god	too	fall
outage	road	stop	hair	government	one	thinking	people	girl	face
storm	old	think	women	true	how	brain	will	haha	then
change	over	how	blood	new	would	depression	our	needs	get
house	problems	sh*t	how	world	could	anxiety	those	see	injuries
may	chaos	want	skin	they	were	lack	one	walk	had
flooding	morning	cant	records	media	had	night	around	drama	sports
gas	two	because	adversity	other	ever	without	life	hes	stick
air	driving	too	high	obama	whats	love	his	woman	over
say	major	hate	helium	financial	again	mental	thats	some	while
stir	today	need	which	change	did	them	work	last	eyes
heavy	disruption	only	eating	violence	time	mind	out	strong	only
weather	train	really	may	will	think	fact	good	shes	hit
collapse	accidents	many	body	also	well	insomnia	sex	always	famous
climate	almost	even	smoking	shutdown	something	hand	come	him	hockey
death	into	then	own	issue	ill	even	great	ways	right
deaths	driver	someone	acne	support	still	feel	say	little	left
home	police	their	death	kids	about	physical	back	because	injury
oil	until	away	brain	problem	sure	emotional	when	said	got
massive	delay	always	alcohol	poor	hope	become	give	really	room
attack	congestion	feel	common	free	get	can	their	thats	involvement
blast	school	thats	deaths	says	youve	too	them	here	innumerable
two	late	say	news	pay	thats	less	things	man	time
city	weather	thing	treatment	against	day	same	goes	ass	play
into	been	something	unknown	party	some	often	comes	night	run
state	earlier	yourself	damage	confusion	good	keeps	too	ego	because

TABLE I

TOPICAL FOCI OF CAUSAL DOCUMENTS. EACH COLUMN LISTS THE UNIGRAMS MOST HIGHLY ASSOCIATED (IN DESCENDING ORDER) WITH A TOPIC, COMPUTED FROM A 10-TOPIC LATENT DIRICHLET ALLOCATION MODEL. THE TOPICS GENERALLY FALL INTO THREE BROAD CATEGORIES: NEWS, MEDICINE, AND RELATIONSHIPS. MANY TOPICS PLACE AN EMPHASIS ON NEGATIVE SENTIMENT TERMS. TOPIC NAMES WERE DETERMINED MANUALLY. WORDS ARE HIGHLIGHTED ACCORDING TO SENTIMENT SCORE AS IN FIG. 1.

spread online and, if so, in which ways? What types of social media users are more or less likely to make causal statements? Will a user be more likely to make a causal statement if they have recently been exposed to one or more causal statements from other users?

The topics of causal statements also bring forth important questions to be addressed: how timely are causal statements? Are certain topics always being discussed in causal statements? Are there causal topics that are very popular for only brief periods and then forgotten? Temporal dynamics of causal statements are also interesting: do time-of-day or time-of-year factors play a role in how causal statements are made?

Our work here focused on a limited subset of causal statements, but more generally, these results may inform new methods for automatically detecting causal statements from unstructured, natural language text [18]. Better computational tools focused on causal statements are an important step towards further understanding misinformation campaigns and other online activities. Lastly, an important but deeply challenging open question is how, if it is even possible, to validate the *accuracy* of causal statements. Can causal statements be ranked by some confidence metric(s)? We hope to pursue these and other questions in future research.

APPENDIX

Parts-of-speech tagging depends on punctuation and casing, which we filtered in our data, so a study of how robust the POS algorithm is to punctuation and casing removal is important. We computed POS tags for the corpora with and without casing as well as with and without punctuation (which includes hashtags, links and at-symbols). Two tags mentioned in Fig. 1B, NNPS and LS (which was not significant), were affected by punctuation removal. Otherwise, there is a strong correlation (Fig. 4) between Odds Ratios (causal vs. control) with punctuation and without punctuation, including casing and without casing ($\rho = 0.71$ and 0.80, respectively), indicating the POS differences between the corpora were primarily not due to the removal of punctuation or casing.

ACKNOWLEDGMENTS

We thank R. Gallagher for useful comments and gratefully acknowledge the resources provided by the Vermont Advanced Computing Core. This material is based upon work supported by the National Science Foundation under Grant No. ISS-1447634.

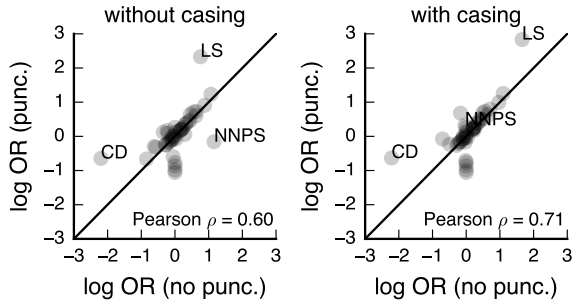

Fig. 4. Comparison of Odds Ratios for all Parts-of-Speech (POS) tags with punctuation retained and removed for documents with and without casing. Tags Cardinal number (CD), List item marker (LS), and Proper noun plural (NNPS) were most affected by removing punctuation.

REFERENCES

[1] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann *et al.*, "Life in the network: the coming age of computational social science," *Science (New York, NY)*, vol. 323, no. 5915, p. 721, 2009.

[2] S. Asur and B. A. Huberman, "Predicting the Future With Social Media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2010, pp. 492–499.

[3] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of Social Media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[4] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." in *LREc*, vol. 10, 2010, pp. 1320–1326.

[5] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who Says What to Whom on Twitter," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 705–714.

[6] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *PloS one*, vol. 6, no. 12, p. e26752, 2011.

[7] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth, "The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place," *PloS one*, vol. 8, no. 5, p. e64417, 2013.

[8] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and Tracking Political Abuse in Social Media," in *ICWSM*, 2011.

[9] M. Salathé and S. Khandelwal, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control," *PLoS Comput Biol*, vol. 7, no. 10, p. e1002199, 2011.

[10] H. J. Larson, L. Z. Cooper, J. Eskola, S. L. Katz, and S. Ratzan, "Addressing the vaccine confidence gap," *The Lancet*, vol. 378, no. 9790, pp. 526–535, 2011.

[11] D. Hume, *A Treatise of Human Nature*. Courier Corporation, 2012.

[12] I. Kant and P. Guyer, *Critique of Pure Reason*. Cambridge University Press, 1998.

[13] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[14] D. B. Rubin, "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions," *Journal of the American Statistical Association*, 2011.

[15] J. S. Sekhon, "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods," *The Oxford handbook of political methodology*, pp. 271–299, 2008.

[16] C. E. Frangakis and D. B. Rubin, "Principal Stratification in Causal Inference," *Biometrics*, vol. 58, no. 1, pp. 21–29, 2002.

[17] J. Pearl, *Causality*. Cambridge university press, 2009.

[18] R. Girju, D. Moldovan *et al.*, "Text Mining for Causal Relations," in *FLAIRS Conference*, 2002, pp. 360–364.

[19] C. Pechsiri and A. Kawtrakul, "Mining Causality from Texts for Question Answering System," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 10, pp. 1523–1533, 2007.

[20] H. D. Kim, M. Castellanos, M. Hsu, C. Zhai, T. Rietz, and D. Diermeier, "Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 885–890.

[21] M. Rolfs, M. Dambacher, and P. Cavanagh, "Visual Adaptation of the Perception of Causality," *Current Biology*, vol. 23, no. 3, pp. 250–254, 2013.

[22] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in cognitive sciences*, vol. 4, no. 8, pp. 299–309, 2000.

[23] R. Joynson, "Michotte's Experimental Methods," *British Journal of Psychology*, vol. 62, no. 3, pp. 293–302, 1971.

[24] H. H. Kelley, "Attribution Theory in Social Psychology." in *Nebraska symposium on motivation*. University of Nebraska Press, 1967.

[25] S. E. Taylor and S. T. Fiske, "Point of View and Perceptions of Causality." *Journal of Personality and Social Psychology*, vol. 32, no. 3, p. 439, 1975.

[26] H. H. Kelley and J. L. Michela, "Attribution Theory and Research," *Annual review of psychology*, vol. 31, no. 1, pp. 457–501, 1980.

[27] D. J. Hilton, "Conversational processes and causal explanation," *Psychological Bulletin*, vol. 107, no. 1, p. 65, 1990.

[28] G. Bohner, H. Bless, N. Schwarz, and F. Strack, "What triggers causal attributions? the impact of valence and subjective probability," *European Journal of Social Psychology*, vol. 18, no. 4, pp. 335–345, 1988.

[29] R. Brown and D. Fish, "The psychological causality implicit in language," *Cognition*, vol. 14, no. 3, pp. 237–273, 1983.

[30] S. Bird, "NLTK: the Natural Language Toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006, pp. 69–72.

[31] J. B. Lovins, *Development of a Stemming Algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[32] J. Plisson, N. Lavrac, D. Mladenic *et al.*, "A Rule based Approach to Word Lemmatization," *Proceedings of IS-2004*, pp. 83–86, 2004.

[33] D. M. Tax, M. van Breukelen, R. P. W. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by multiplying?" *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.

[34] W. N. Francis and H. Kucera, "Brown corpus manual," *Brown University*, 1979.

[35] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[36] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[37] A. Agresti and M. Kateri, *Categorical Data Analysis*. Springer, 2011.

[38] D. Klein and C. D. Manning, "Fast Exact Inference with a Factored Model for Natural Language Parsing," in *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, vol. 15. MIT Press, 2003, p. 3.

[39] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631. Citeseer, 2013, p. 1642.

[40] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[41] A. K. McCallum, "MALLETT: A Machine Learning for Language Toolkit," 2002.

[42] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, 1993.

[43] S. Pinker, *The Better Angels of Our Nature: Why Violence Has Declined*. Penguin, 2011.