Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media

Alexandra Olteanu
alexandra@aolteanu.com

Onur Varol
Indiana University
ovarol@indiana.edu

Emre Kıcıman
Microsoft Research
emrek@microsoft.com

ABSTRACT

Millions of people regularly report the details of their real-world experiences on social media. This provides an opportunity to observe the outcomes of common and critical situations. Identifying and quantifying these outcomes may provide better decision-support and goal-achievement for individuals, and help policy-makers and scientists better understand important societal phenomena. We address several open questions about using social media data for open-domain outcome identification: Are the words people are more likely to use after some experience relevant to this experience? How well do these words cover the breadth of outcomes likely to occur for an experience? What kinds of outcomes are discovered? Studying 3-months of Twitter data capturing people who experienced 39 distinct situations across a variety of domains, we find that these outcomes are generally found to be relevant (55–100% on average) and that causally related concepts are more likely to be discovered than conceptual or semantically related concepts.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human information processing; H.4.2 Types of Systems: Decision support

Author Keywords

Experience Outcomes; Information Extraction; Social Media

1. INTRODUCTION

Many people, for many reasons, are interested in investigating unfamiliar or poorly understood situations. Individuals find themselves needing to make sense of an unfamiliar situation and understand how events and actions might unfold. Someone diagnosed with a medical condition might be interested in learning about the likelihood of specific symptoms. A person training for a marathon may wonder which training regime has the best outcomes. Someone making a major life decision—to go to law school, to join the navy, or to move across the country—may wonder how their life may change as a result. And a person making an everyday decision—whether to go to the gym, eat an ice cream, or walk in the park—might benefit from knowledge about the aggregated implications of such decisions.

Of course, investigating poorly understood scenarios is not limited to individuals exploring their own situations. Policy makers and scientists ask similar questions about situations of societal importance: What happens to a child after being bullied [3]? What factors put people at risk of considering suicide [28]? And, what happens to individuals when they lose their jobs [81]? From social and public health to financial and many other domains, phenomenon of interest to scientists are as wide-ranging as the questions of interest to individuals.

With computing and sensing devices embedded in our everyday lives and mediating an increasing degree of our interactions with both the digital and physical world, services that help people understand unfamiliar situations and possible actions will have broad, increasing impact in enabling better decision-making and goal-achievement. Research in fields such as social psychology, medicine, computer supported collaborative work, and human computer interaction has shown that information such as action plans, task and goal reminders, and reviews can have a significant positive impact on goal achievement of individuals [2, 45, 55, 80]. Possible uses include coaching apps that help review past and upcoming actions and their likely future impact; personalized recommendations based on individual's short and long-term goals; pros/cons lists and other decision aids; as well as general purpose, situational exploration and information sense - making tools; and many other personal assistant systems [24]. Analogous applications for sense-making, visualization and exploration of a given phenomenon are similarly critical for policy-makers and scientists interested in monitoring and designing interventions to address societal problems.

Addressing informational needs in potentially *any* unfamiliar situation requires an open and domain-agnostic knowledge of scenarios people may find themselves in [11, 14, 22, 36]. Many of the existing solutions, however, tend to focus either on common needs, or on specific domains—being hand-crafted or leveraging domain specific patterns that enable them to achieve high precision due to limited scope [35, 65]. However, most people have both common and long-tail information needs [43], and such approaches do not scale to cover the long-tail of unpopular needs and situations people find themselves in [9, 14, 22]. While many of these needs will surface hundreds to thousands of times over long time periods on large-scale platforms (e.g., cloud-based virtual assistants, web search engines), they may still be considered rare and far

CSCW 2017, February 25–March 1, 2017, Portland, OR, USA.

ACM ISBN 978-1-4503-4189-0/17/03 ...$15.00.
http://dx.doi.org/10.1145/2998181.2998353

in the tail of things people inquire about to justify the cost of creating and maintaining answers for them [14].

Social Media as a Window into People's Experiences

To characterize and better understand such a wide variety of experiences and scenarios, we turn to social media as a data resource. In social media, hundreds of millions of people regularly and publicly report on their experiences, including the actions they take, the things that happen to them, and the experiences they have afterwards. They talk about work or relations [33, 41], health and dietary practices [1, 93], and even log information about their illnesses and coping strategies [23, 34]. People do this for many reasons [30, 51, 57, 71, 74]: keeping in touch with friends, gaining social capital, or even helping others. And with the increasing use of personal sensors and devices, from exercise trackers to health monitors, such social streams are becoming more regular, more detailed and more reliable [8, 67, 79]. Regardless of why people share this information, such *social media posts can be leveraged to better understand common and critical situations and their outcomes.*

We investigate a quantitative analysis of social media timelines on Twitter that extracts relationships between experiences that people mention having, and the words or experiences they are more likely to mention in the future: if a person mentions some experience Z, they are a times more likely to mention some outcome w in the future. For each relationship $\{Z, a, w\}$, we also extract qualitative information (social media messages) that can be used to better understand the nuance and the context of each specific relationship. While building specific applications is outside the scope of this paper, we believe that extracting these semi-structured relationships provides a potential building block for many applications.

Contributions and Study Overview

In this paper, we address several open questions as we try to exploit social media data for general-purpose outcome identification tasks. First, are the words people are more likely to use after some experience clearly relevant to a target experience? Second, how well do these words cover the breadth of outcomes likely to occur after an experience? And finally, what kind of words do we discover? Given that we cannot claim to perform a causal analysis, how often are the words conceptually related? How often do they capture actual causal relationships? To address these questions, our study is organized as follows:

Selection of experiences (§3): Using a large corpus of web search queries, we select 39 experiences covering a variety of domains. These experiences include consequential *actions* like taking a strong prescription drug or getting a divorce—where people are interested in the outcomes to aid their decisions—but, also, personal *situations* such as suffering from physical or mental ailments like gout, high cholesterol or anxiety.

Statistical analysis of social media timelines (§4): Then, we analyze 3-months of social media timelines to identify users reporting on one of these experiences, and to extract the words that they are more likely to use after these reports. The analysis methodology we apply to social media data is conceptually straightforward, comparing the timelines of users who have experienced some specific situation event to the timelines of users who did not, using a stratified propensity score analysis[1].

Results coverage & relevance (§5): Finally, to assess the quality of the outcome words we discover, we estimate their *precision* by crowdsourcing their human-judged relevance, and their *coverage* by contrasting them to search queries and related concepts in a large knowledge base, ConceptNet5 [91]. We also use this knowledge base to characterize the kinds of outcomes we detect (e.g., motivations, direct causes, or properties).

Results Overview. Overall, we find that the words people are more likely to use after reporting an experience are on average perceived to be 55–100% relevant across semantic domains; have a coverage of 41-52% over related concepts in a baseline knowledge base; and up to 60% for causal-like relationships such as motivations and consequences. However, we do find that the temporal relationships between experiences and outcomes are often reversed in social media timelines.

Prior work has begun to analyze social media and other temporal activity traces to build models related to specific experiences [28, 39, 81, 82, 98]. While propensity score analysis itself is not novel, this is, to our knowledge, the first adaptation of the method to extract outcomes of experiences from a large-scale, high-dimensional textual data set (social media messages) and the first to characterize the general relevance and kinds of experiences people are more likely to mention across a broad set of situations and domains.

2. PROBLEM SETTING & PRIOR WORK

In our quest to understand the potential of social media as a data resource for mining the outcomes of people experiences for decision support, our work is inspired by and builds upon prior efforts: to understand the nature and the use of social media, to apply causal inference techniques to social media data, and to construct large-scale knowledge bases for supporting users' informational needs. We also discuss towards the end of the paper how our work can complement efforts to design and build decision-support applications (§6.2).

Self-Disclosure on Social Media: Although social media is a multi-purpose communication medium, messages about users' own experiences account for a notable fraction of content: for instance, "me now" messages about personal state and experiences were found to constitute 37–51% of all messages [71] and about 26% of tweets were found to be experiential [53]. This tendency to disclose information about oneself on social media is part of a broader phenomenon [96]. Research estimates that self-disclosure represents 30–40% of human speech output [32, 59, 92]. This is a *key* feature of social media that promises to enable our work. At the same time, it is also important to note that social media data may not capture all users' experiences or all aspects surrounding

[1]Note that, while we borrow techniques from the causal inference literature, we cannot claim to meet the assumptions required for causal inference without additional domain knowledge that is unavailable in the general case.

these experiences, as users may often chose to remain "silent" on various topics [46]. In addition, users might also mention their experiences or related aspects out of order. Such idiosyncrasies of social media influence the kind of insights (outcomes of a given experience in our case) that one can draw from it, which we aim to understand and characterize.

Mining User Timelines: Aggregating user traces into user profiles or timelines has proved effective in understanding the behavior of various sub-populations. Using search logs, Paul et al. [77] characterized the information seeking behavior during various phases of prostate cancer, while Fourney et al. [39] aligned it with the natural clock of gestational physiology for pregnant users. More generally, Richardson [82] showed that such long-term search logs provide useful insights about topical and temporal real-world relationships. Similarly, by mining social media, De Choudhury et al. [27] found behavioral cues useful to predict the risk of depression before onset, while others studied the behavioral changes of users sharing personal health and fitness information [75], or even the characteristics of users supporting terror groups like ISIS [63]. This indicates that although data on many types of user experiences may be sparse (both within a social media, as well as a user timeline), we may still be able to draw insights about these experiences and their outcomes by aggregating cues from multiple users' timelines.

Observational Studies of Social Behavior: By leveraging this kind of data, prior work examined how dietary habits vary across locations [1], and the links between diseases, drugs, and side-effects [70, 76]. Other studies, which have looked at economic and financial trends [16, 44], have framed the problem of learning about the world as a prediction problem: given a historical known measure, predict its current or future values from current social media signals [4].

Controlling for Confounding Bias: While it has been emphasized that controlling for confounding bias in observational studies on social media is important [42], this is rare: many analyses are only co-occurrence based and assume that co-occurring items share some true relation. For instance, links among disease carriers and new infections based on co-visited locations were found by Sadilek et al. [85], while Paul and Dredze [76] identified links between mentioned ailments and the geographies where they occur.

Recent studies looking at migration patterns [99], shifts in suicidal ideation [28], at the effect of exercise on mental health [31], or of community feedback on individual user behavior [21] try to improve on correlational analyses by applying causal inference techniques that have come into extensive use in medicine, economics, and other sciences. Such techniques include differences-in-differences models [5], the potential outcomes framework of the Rubin causal model [88] and the structural equation model [78, 83]. Relatedly, Landeiro and Culotta apply causal modeling techniques to build more robust classifiers for social media texts [58].

To complement these efforts, our goal is to develop generalizable techniques that separate domain-agnostic mechanics of such analyses from the semantic interpretation of results that often requires domain knowledge. We choose to use a high-dimensional stratified propensity score analysis (§4.2). We use a stratified analysis to avoid matching issues outlined by King and Nielsen [54]. Our propensity score estimation considers all words used in the past by individuals. This high-dimensional analysis accounts for as many covariates as are available that could predict the likelihood of a user to have the experience, making the assumption of unconfoundedness more plausible [6]. While these terms are unlikely to capture all variables correlated with the confounding variables (as it is hard to argue that all relevant aspects of users' lives are captured in their social media streams), word use is known to correlate with various psycho-socio-economic factors including gender, age and personality [37, 87].

Knowledge Bases and Online Q&A sites: Finally, our work is inspired by prior efforts to generate, aggregate, and organize knowledge for helping users with their informational needs. Curated large knowledge bases, like Wikipedia or Freebase, help accurately answer questions about encyclopedic topics, from locations to celebrities [97], as they allow systems to reason over high-precision knowledge [35]. Yet, they have limited recall, which is typical to curated resources; and, often, even simple information about common actions—such as the effect of eating pasta before running a marathon, or the consequences of adopting a puppy—are missing. Question-answering sites are also important venue for knowledge generation and decision-support [72]. While question-and-answer sites are useful for some explorations, there is time and effort associated to individuals tracing theirs (or others) posted questions [62], as well as to judging the applicability of existing answers (if any); free text answers are not good building blocks for applications that require a statistical and temporal representation of situations and actions; and, moreover, they often provide subjective answers based on the opinions of a small number of people and of varying quality [72]. In contrast, our goal is to scale and automatize such explorations on behalf of interested individuals by aggregating relevant reports from many users on social media.

3. DATA COLLECTION

We begin with the description of the datasets we assemble for this study. For this, we first explain how we selected a set of experiences across multiple semantic domains in a way that balances between their topical diversity and their prevalence within a large corpus of web search queries. Then, we describe how we identify users that had each of these experiences on a popular social media platform, Twitter, which we refer to as *treated* users.

While our evaluation covers only one social media platform, Twitter is a large source of publicly available social media data, recognized as promising for many domains such as e.g. public health and well-being [25, 26], disaster relief [74], or microeconomic trends [81]. Twitter is also frequently used by users to report on their personal daily happenings [1, 25, 53, 71], which, along with their timestamps, may provide insights into temporal relations among personal events. This characteristic makes it suitable for our purpose. In addition, our analysis framework is designed to be data source and domain-

Category		Experience / Event	Treatment Users	Msgs
Business	Construct. and Maintenan.	*Building stairs	24	8.3K
		*Cleaning countertops	8	3.7K
		*Installing a garbage disposal	29	8.2K
		Painting the deck	592	164K
	Financial Services	Owning a good credit card	2291	920K
		Paying credit card debts	414	233K
		Buying life insurance	1881	561K
		Having pension	2344	796K
		*Incorporating one's business	28	9.1K
	Investing	Becoming a broker	855	355K
		Investing money	23981	9.8M
		Total	32447	12.8M
Health	Diseases	Having high blood pressure	5279	1.9M
		Having gout	364	118K
		Having high cholesterol	1384	522K
		Having kidney stone	727	259K
		*Having high triglycerides lev.	27	12K
	Mental	Suffering from depression	25207	10.5M
		Suffering from OCD	11429	4.8M
		Being a sociopath	1491	676K
		Being a psychopath	2895	1.3M
		Suffering from anxiety	53983	22.6M
		Suffering from bipolar disord.	13723	6.3M
	Pharmacy	Taking Prozac	617	222K
		Taking Lorazepam	47	19.4K
		Taking Promethazine	242	118K
		Taking Tramadol	397	161K
		Taking Xanax	3300	1.4M
		Total	121112	50.9M
Society	Issues	Losing belly fat	93	24.8K
		Increasing gross income	135	93.2K
	Law	Getting divorced	2717	1.2M
		Becoming a notary	65	22.7K
		Applying for social security	6172	2.3M
		Filing for bankruptcy	921	347K
		*Having a living trust	18	7.5K
	Relationsh.	Finding true love	1885	654K
		*Recovering after adultery	9	4.4K
		Filing divorce	422	178K
		Dealing with jealousy issues	789	370K
		Changing last name	1019	403K
		Total	14245	5.7M

Table 1. The domains and experiences we use in our analysis, along with the number of users mentioning them, and the number of tweets, on any topic, written by these users. We note in gray (*) cases in which we found less than 30 treated users. While we expect such low-volume analyses to fail, we retained the results throughout the evaluation to capture a fuller understanding of the relationship of data volume to result quality.

agnostic, with only a few elements requiring adaptation—which we outline towards the end of the paper.

List of Experiences

Given that we want to understand the feasibility of our analysis for a broad variety of situations, we selected a diverse set of experiences about which people actively seek information to make decisions. To determine this, we used a large corpus of anonymized web search queries from a major search engine, Bing.com—covering the first week of every month in 2014—and extracted those queries containing question-phrases (e.g., *what, why, should, how*) or comparison terms (*or*, *vs*). This resulted in 200 million distinct queries from which we randomly sampled 11 million based on their popularity,[2] and categorized them along the semantic domains and subdomains from the Open Directory Project[3] using an existing query classifier [12]. We then selected the top 20 queries from the top 5 sub-domains of the 6 most popular domains, and kept the *decision*-related queries and those subdomains with at least 2 such queries.[4] Table 1 lists the experiences we considered in our study, and the basic figures of the corresponding datasets we assembled for them from Twitter, which we detail next.

Identifying Treated Users

For each of the experiences in our set, we construct a data collection consisting of corresponding treated users and their tweets. In doing so, our user identification strategy is geared towards precision to minimize the noise in the outcomes we distill later in the analysis. To this end, we start our analysis with a Twitter data archive corresponding to *all public tweets* in English from March 1 to May 31, 2014[5]—selected to avoid strong seasonal phenomena, typically observed in winter, summer and early fall [26, 68].

To identify tweets of users that had a given experience—in April 2014—we search for a disjunction (logical OR) of *multiple* queries each including a conjunction of phrases written to avoid possible ambiguity in the experience terms, and identify with reasonable accuracy tweets of users that have had the experience, as opposed to using the word in different contexts. For instance, for drug names, we wish to avoid including advertisement tweets like "*buy Xanax online*", and we instead search for tweets containing "*I took Xanax*", "*I was on Xanax*" or "*I tried Xanax*". Similarly, for legal actions, to avoid advertisement tweets like "*get legal advice to file for divorce*", we instead search for tweets containing e.g., "*I am filling for divorce*" or "*I filled divorce*". Specifically, for each experience we looked for combinations of 1st person personal or possessive pronoun (e.g., I, me, my), the experience object (e.g., Xanax, divorce, kidney stone, pension), and verbs or verbal phrases typically used in conjunction with this object (e.g., for *depression* we used verbs like *had* or *diagnosed*; for *belly fat*, we used verbs like *lost* or *burned*). As a result, we obtained tens of queries per experience.

Once we identified such tweets, we return to our *full* Twitter archive to extract a 3-month timeline for their corresponding users. We discard users with fewer than 5 or more than 1,000 tweets in this period to filter out possible bot or organization accounts, as well as users with too little data. Table 1 shows the resulting dataset figures for each experience. We notice important variations in the number of users across experiences: many users talk about depression and anxiety on social media, but only a few seem to discuss construction tasks. We discuss later in the paper how low numbers of users for some situations impact the quality of results (§5.2).

[2]By query popularity we refer to how often the query was executed.

[3]http://www.dmoz.org

[4]The filtering of *decision*-related queries was done by two annotators independently with a $\kappa = 0.50$ (95% CI: $[0.43, 0.57]$).

[5]Public Twitter stats from this interval show 255 million monthly active users and over 500 million tweets per day: see e.g., https://blog.twitter.com/2014/the-2014-yearontwitter, https://investor.twitterinc.com/releasedetail.cfm?ReleaseID=843245

Original (O) and Preprocessed (P) Text
O: Took a pill to help me sleep at 10:00 and I just woke up **P:** take pill help me sleep just wake up
O: OMG I've miss opera singing lessons, bleah **P:** oh_my_god have miss opera sing lesson bleah
O: can't believe my homework ap class is coloring. ccooollll **P:** cannot believe my homework ap class color cool

Table 2. Original (O) and processed (P) tweets. These examples have been carefully paraphrased for anonymity.

Data Preprocessing

Then, as a pre-processing step, before running our analysis, we clean and normalize the tweets to filter out non-informative content and collapse variations of the same dictionary term to a single base form. This step consists of basic lemmatization (removal of plurals, common verb suffixes, converting verbs to present tense and other word inflections), replacing URL and @user mentions, expanding known abbreviations, removing stopwords and normalizing common slang forms, such as repeated letters [17] (e.g., "wooowww" and "coool" become "wow" and "cool"). Table 2 shows examples of tweets before and after cleaning.

4. METHODOLOGICAL FRAMEWORK

To infer the outcomes of a target experience from a large corpus of social media messages, we first convert the messages into timeline representations of personal events per user (§4.1). Some of these events may be actions explicitly taken by the individual. Other events may describe outcomes that came about because of such an action, or situations that arose independent of user actions. We group these timelines into *treatment* and *control* groups based on whether they include or not a *target event*[6] and stratify them into comparable subpopulations through a propensity score analysis (§4.3 and §4.2). We then measure the binary-valued outcomes of the target experience within each stratum (§4.4).

Here, we focus on the extraction and evaluation of outcomes that follow personal events in social media timelines. Other aspects of a broader system, such as query formulation heuristics, visualization and exploration of outcomes, integration with various applications, or enabling web-scale processing are outside the scope of this paper.

4.1 Building User Timelines

We begin with a corpus of social media messages including a short text message, a user identifier and a timestamp. We represent each message as a list of arbitrarily sorted unigrams and bigrams which we will also refer to as *events* in the rest of the paper.[7] We organize these lists according to the user identifier, and then for each user we lay them out per their corresponding timestamp in a *per-user timeline*. We further summarize this complete per-user timeline, representing the information needed for our analysis as two partial timelines per user: one timeline tracking the *first* occurrence T_f of each unique event in a user's timeline, and one timeline tracking the *last* occurrence T_l of each unique event. This minimal summary allows us to efficiently calculate the set of all events that have occurred *before* any point in time, used for our propensity score estimation (§4.3); as well as the set of all events that have occurred *after* any point in time, used for outcome measurement (§4.4). Finally, we search through the timelines that contain a target event to identify the treatment and control groups.

Other work has suggested alternative methods of building experiential times from social media data. Kıcıman and Richardson [53], for example, advocates using a classifier to identify messages that describe personal experiences. We skip this step as we are also interested in outcomes occurring in non-experiential messages, such as changes in conversational language usage.

4.2 Propensity Score Analysis

The goal of our analysis is to understand whether a user mention of some experience makes her more or less likely to mention certain events in the future. This is a causal question. While we do not believe we can achieve the ideal identification of causal relationships, we can use techniques borrowed from causal inference literature that generally bring us closer to this ideal than simple correlational analyses.

Ideally, to determine whether some experience "causes" an outcome, we would be able to observe and compare two potential outcomes: one outcome $Y_i(X=1)$ after a person i has the target experience X, and another outcome $Y_i(X=0)$ when the same person in an identical situation does not have the experience. The causal effect of the experience on person i is then $Y_i(X=1) - Y_i(X=0)$. Of course, it is impossible to observe both $Y_i(X=1)$ and $Y_i(X=0)$ for the same i. Once we observe i having the experience or not, we cannot observe the other counterfactual outcome.

Instead, the causal effect is typically estimated as $E(Y(X=1)) - E(Y(X=0))$ by measuring $E(Y(X=1))$ and $E(Y(X=0))$ in two distinct but *comparable* (i.e., statistically identical) populations: a *treatment* group that has had the experience and a *control* group that has not. A simple method for constructing two such comparable populations is the randomized experiment. In non-randomized or observational studies, however, there may be systematic biases among the two treatment and control groups. Stratified propensity score (PS) matching attempts to address these biases by subdividing the *treatment* and *control* groups into comparable strata based on their estimated likelihood (or propensity) to have the target experience (estimated based on all observed covariates) [84]. Here, we use a high-dimensional propensity score analysis to estimate the "causal" effects among the words people use in their social media posts.[8] That is, if a person uses some word (mentioning a situation or action), are they more or less likely to use other words (mentioning outcomes) in the future, accounting for the covariates (past words).

The assignment into *treatment* or *control* groups is independent of the covariates when conditioned on the propensity

[6] In the context of timelines representation, the target experience is also represented as an event.

[7] We note that this broad inclusion of unigrams and bigrams results in as many as 100M unique events.

[8] Note that we do not claim true causality due to unmet assumptions, as discussed in (§6).

score function [84], which is a function of the covariates observed in the past indicating which users have a similar propensity to have the treatment. This conditional independence implies that the potential outcome of taking the action or not is independent, and thus the population average within each stratum can be estimated based on observed outcomes.

4.3 PS Estimation and Stratification

To isolate the effect of a *target event* Z from observed confounds—to the degree this is possible—we wish to condition our analysis of the effects of Z on our knowledge of the users, particularly our knowledge of their past events (e.g., past unigram and bigram usage). Ideally, as noted before, we would compare any treated user i mentioning Z to an identical user $\tilde{i}$ who has not mentioned Z—impractical in a high-dimensional space. An alternative is to implement this conditioning by comparing groups of treated $\mathscr{I}$ and non-treated $\tilde{\mathscr{I}}$ users that have a similar propensity to mention Z. To this end, we represent a user i's past events as a sparse binary vector of the *first occurrence*[9] of the top K most-popular unigrams and bigrams across our entire corpus.[10] In the *treatment group*, a user history vector is $H_{T_f} = e_1, ...e_K$ where e_j is 1 if the user has mentioned token j before mentioning the target event Z, and 0 otherwise. The history vector for users in the *control group* is defined analogously, with the difference that we consider only tokens mentioned before an arbitrary token in the timeline (selected randomly per-user, as there is no mention of target event Z for this group). In other words, the propensity of a user to mention Z as a function of her past experiences H is given by $p(H_{T_f}) = P(Z|H_{T_f})$.

In practice, for a given Z, we learn the propensity score estimator using an averaged perceptron learning algorithm [40]. We use the data from our treatment and control groups (which are represented as H_{T_f}) as training data. During training, we split our available data and perform a 10-fold stratified cross-validation that preserves the user distribution across treatment and control groups across folds. Using this propensity score, we then stratify the treatment and control groups to subdivide them into comparable groups (or strata), for which we test several approaches, including iterative splitting on the propensity score, binning and quantile division. For the analysis in this paper we stratify users into 10 quantile-based strata,[11] this being sufficient to obtain comparable treatment and control groups [19].

We assess the balance of covariates across stratified control and treatment using a 2-way analysis of variance model to measure the statistical significance of each covariate considering the propensity scored decile. We compare the F-statistics for each covariate before and after stratification. Our ANOVA test finds that, across all our experiments, we reduce the number of statistically significant differences by 33.4%, indicating that our analysis substantially reduces, though does not eliminate, differences in comparison between the treatment and control groups. Note that, in high-dimensional settings, complete balancing of all covariates is not always possible nor it is always necessary [6].

[9]This ensures that the propensity score estimation uses only tokens occurring before *any* mention of Z in each timeline.

[10]Our PS estimation uses all tokens occurring at least 50 times across users timelines.

[11]This strategy ensured better data support across strata.

4.4 Outcome Measurement

Next, we measure the effects of event Z by iteratively examining each post-hoc event (potential outcome) reported by users for systematic differences among the treatment and control groups in their expression of this event after Z.

At this step, we use the timeline representation that includes the *last occurrence* of each event to easily compute the set of outcomes mentioned at least once after the first mention of Z. Note that in this binary representation of outcome occurrences, measuring the difference in outcome occurrences between the treatment and control groups is equivalent to the difference in the likelihood of the binary outcome among the two groups. For each Z and possible outcome, we calculate this average treatment effect and statistical significance over the region of common support—essentially, strata with sufficient treatment and control users—as well as the treatment effect and statistical significance for individual strata.

In addition, we also augment our analysis results with visualizations of the temporal dynamics of outcomes after target event Z, as shown in Figure 1. For qualitative support, we extract pairs of messages demonstrating the target event and outcome effect, as shown in Table 4. While not evaluated directly, we use this information to inform and contextualize our results when we ask our human judges to determine the relevance of our results (see §5.2).

5. EMPIRICAL EVALUATION

To evaluate our outcomes extraction for a given experience, broadly, our experiments have three goals: (1) estimate the precision of the extracted outcomes by assessing how relevant they are to the experience; (2) assess the extent to which our results cover the breadth of outcomes likely to occur after an experience by contrasting them to existing knowledge bases (where available) and search logs; (3) characterize the breadth of applicability of our framework to extract outcomes of experiences across a broad variety of semantic domains, as well as the kinds of outcomes it is most likely to identify.

To run the experiments we present here, we group the datasets (and the experiences) by their sub-domain. We do so to draw the control group from users identified as having a different experience, yet in the same domain (e.g., we compare users suffering from gout with other users with experiences in the health/diseases domain like users suffering kidney stones).

5.1 Exploratory Analysis

We first review the raw results generated by our analysis framework. Table 3 shows the top-most statistically significant outcomes extracted for a sample of experiences. We can easily see that the extracted outcomes are topically related to the target experience: users mentioning that they are *suffering from gout* are significantly more likely to mention *flare ups* of their symptoms, *uric acid*, and the physical locations of their symptoms; users mentioning *losing belly fat*

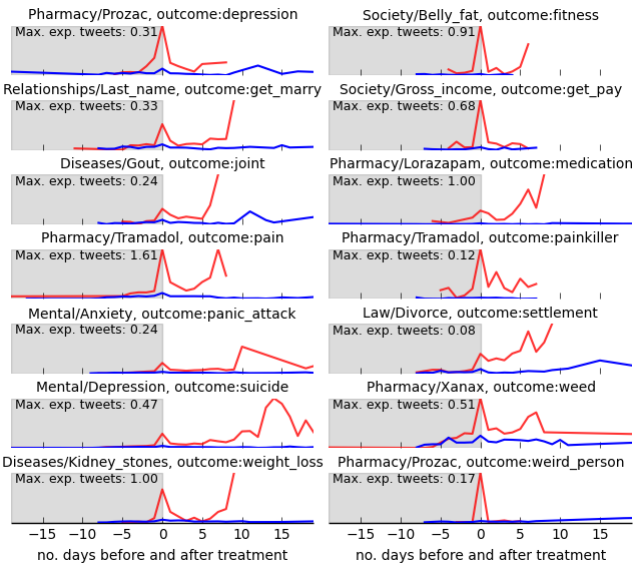


Figure 1. Comparison of temporal evolution of outcomes in treatment (red) and control groups (blue). The volume indicates the expected number of tweets per user (max value highlighted). Best seen in color.

are more likely to talk about their *fitness* programs and about adding new *videos* to their *playlist*; while users mentioning *high triglyceride levels* later discuss *statins*, *cardiovascular* issues, and *dietary* changes. Similarly, we see topically relevant outcomes for other scenarios (e.g., investment-related words such as the stock market, or varieties of anxieties and manifestations such as panic attacks).

Temporal Aspects

In addition, we looked at the temporal evolution of outcomes which provides additional context for characterizing the outcomes of an experience. For this, we use the timeline representation of events occurrences to see when outcomes occur for users in the treatment and control groups. Figure 1 shows the temporal evolution of a sample of outcomes after an experience is mentioned (or at a random time for control users).

We see that e.g., the likelihood of someone mentioning their *depression* increases around when she admits *taking Prozac*, with other outcomes following a similar pattern such as *get paid* when *increasing the gross income* or *weird person* when *taking Prozac*. In other cases, the outcomes become more prominent several days later, like discussing about *medication* when *taking Lorazepam*, *settlement* after *filling for divorce*, or *joint* when *suffering from gout* (known to lead to swelling joints). Other outcomes are more likely to occur both with the target event, as well as days later, e.g., the use of *weed* after *taking Xanax* (also taken for recreational use), or the mention of *suicide* after admitting to *suffering from depression* (although the likelihood is significantly higher later). There are also interesting interplays among outcomes of the same experience: we see the mention of *painkiller* peaking around the time users mention *taking Tramadol*, while *pain* seems to recur after several days.

To further understand these results, we do a qualitative pull-out of users who first mention the target experience and later mention the extracted outcome. These paired messages can provide important support to understanding the context of the mentions and interpret the semantic relationship between the two events (the experience and the outcome). Table 4 shows a sample of such cases. We find these pairs of messages to be critical in gathering judgements of the quality of results, which we describe in the next section.

5.2 Outcome Precision

Here, we describe the design of our crowdsourcing task to annotate the perceived precision of the extracted outcomes.

Crowdsourced Annotation

Detecting false positives among the extracted outcomes is akin to a binary classification task where the positive class corresponds to the outcomes perceived as relevant to the target experience. For this, we employed crowd-workers to manually annotate the outcomes. We showed workers two pairs of tweet examples each including treatment and outcome tweets by the same users (as in Table 4), and clickable links to corresponding search results returned by two major search engines, `Bing.com` and `Google.com`. The query issued to the search engines contains the treatment phrase and the outcome e.g., *taking Lorazepam public speaking*. We then ask them to review a statement like:

> *Someone taking Lorazepam will later on be more likely to talk about public speaking.*

and annotate this statement as: A. **Correct**—if the information available within the search results and tweet examples clearly confirm the statement; B. **Likely to be correct**—if they somewhat confirm the statement; C. **Likely to be incorrect**—if they somewhat contradict the statement; D. **Incorrect**—if they clearly contradict the statement; E. **Other issues**—if there is no evidence to judge the statement, the information is unrelated or incomprehensible.

Given our aim to filter out outcomes clearly not related to the target experience, the task is formulated to encourage workers to be inclusive. For purposes of our evaluation we collapse the first two categories identifying treatment/outcome pairs as correct or likely to be correct given the annotation context we provide (e.g., tweet examples, search queries) into a single positive class; and the rest of categories into a single negative class. For each experience in our list, we annotate the top 50 outcomes as ranked by their statistical significance, measured by two-tailed z-score. We limit our analysis to results that are statistically significant at $p < 0.0001$. For each outcome, we generate 1 to 3 tasks (with distinct pairs of treatment/outcome tweets) for labeling, and have each task judged by 3 workers (resulting in up to 9 annotations per treatment/outcome pair).

Overall, 95% of the annotations were collected from crowd-workers with 100% life-time approval rate on Amazon's Mechanical Turk, representing 82% of our tasks' annotators. For 88% of our treatment/outcome pairs we observe a clear majority of over 66% (typically, out of 9 distinct workers) towards one of the classes, with a 73% average pairwise percent agreement—deemed appropriate for exploratory evalua-

Outcome	Count	Effect%	Z-Score
Business\Investing: Investment			
market	15457	14.9	111.22
investor	3253	6.1	104.77
stock	15847	15.2	94.94
stock_market	2867	5.6	93.38
property	6315	7.1	83.84
business	27272	15.7	82.03
profit	4880	6.2	81.07
financial	5860	6.3	78.90
company	18897	12.7	78.72
fund	6979	7.4	78.66
Health\Diseases: Gout			
flare_up	35	4.1	12.33
uric_acid	27	2.9	10.36
uric	28	2.9	10.11
flare	81	4.9	9.92
big_toe	38	2.9	9.86
joint	301	7.2	7.22
aged	32	1.7	6.51
correlation	45	2.8	6.11
bollock	53	2.5	5.96
shite	108	3.4	5.93
Health\Diseases: Triglycerides			
your_risk	46	24.8	18.12
statin	48	23.1	17.69
lower	120	35.9	17.18
cardiovascular	54	23.0	16.72
healthy_diet	55	19.3	16.54
fatty_acid	29	18.3	16.37
help_prevent	73	26.9	16.01
risk_factor	33	18.3	15.55
fish_oil	48	24.4	15.42
inflammation	78	25.1	15.30

Outcome	Count	Effect%	Z-Score
Health\Pharmacy: Lorazepam			
ativan	24	17.4	14.11
depression_mention	23	11.8	11.02
take_med	24	12.8	10.18
one_morning	24	14.2	9.94
anti_depressant	28	12.3	9.29
doctor_prescribe	21	10.2	8.99
depressant	30	12.3	8.91
take_mg	29	13.7	8.91
benzo	21	10.6	8.73
help_have	33	10.9	8.58
Society\Issues: Belly fat			
burn	156	62.2	8.96
ab_workout	13	8.5	5.82
workout_lose	13	8.5	5.82
help_burn	8	11.1	5.82
add_video	26	14.0	5.75
url_playlist	26	14.0	5.75
fitness	39	18.6	5.51
ab	43	19.1	5.51
playlist_mention	30	15.3	5.39
biceps	7	4.7	4.74
Society\Law: Notary			
please_see	147	9.6	10.44
property_mention	89	7.0	10.22
my_answer	246	14.6	9.82
url_legal	68	6.2	9.54
refinance	74	6.5	9.22
can_file	78	5.8	8.80
please_check	192	9.3	8.80
attorney	449	12.6	8.47
mortgage	329	12.6	8.33
obtain	243	9.1	7.97

Outcome	Count	Effect%	Z-Score
Society\Relationships: Divorce			
your_summer	50	10.1	14.13
ex_husband	22	4.5	12.53
summer_get	51	9.0	12.51
url_divorce	23	4.8	12.24
attorney	37	4.8	10.97
marriage	286	13.1	10.46
marry	851	22.5	10.34
tunnel_come	15	3.8	10.09
lawyer	81	6.8	9.98
yr	257	11.6	9.8
Health\Mental: Anxiety			
attack	23650	17.3	55.14
separation	5026	6.8	43.37
have	86166	3.0	32.42
have_social	2381	3.7	28.57
panic_attack	4027	4.4	28.56
anxious	5034	4.7	27.03
panic	8269	5.6	26.77
mom	46259	4.2	25.09
give_me	37173	6.6	25.04
literally	37692	4.3	24.36
Business\Financial: Pension			
tax	1334	18.9	18.89
retire	675	15.6	17.97
budget	762	14.0	17.05
benefit	920	14.7	15.82
vote	1278	13.9	15.57
government	876	11.5	15.47
financial	673	13.7	15.22
income	619	12.3	14.87
report	1125	13.7	14.7
investment	490	10.4	14.28

Table 3. Most significant 10 outcomes following selected events.

Treatment	Example tweet	Outcome	Example tweet
Dealing with jealousy issues	*ironically, u ask why I have jealousy issues*	wake up	*@user I need u to wake up because im bored*
Suffering from depression	*if u think depression is eccentric or cute u can have mine bc i dont wanna deal with it*	self harm	*@user dont self harm, remember yr worth so much better, u dont deserve this pain, stay safe*
Suffering from anxiety	*if hadnt spent years dealing with anxiety, I wouldnt have my sense of humor*	yelling	*dont have anger issues at all, really happy when yelling at people*
Paying credit card debts	*seriously, my soul was deep hurt when I paid that credit card bill*	apartment	*Im checking some apartments in NYC lol*
Having high blood pressure	*@user but I do have really high blood pressure!*	stress	*had a heart mri and the news are as good as they can be. much stress is now removed from my life*
Having gout	*@user I know I have gout...*	uric	*is the increase in uric acid production in blood, forming crystals and depositing them in joints*

Table 4. Paired treatment and outcome messages for selected users, carefully paraphrased for anonymity.

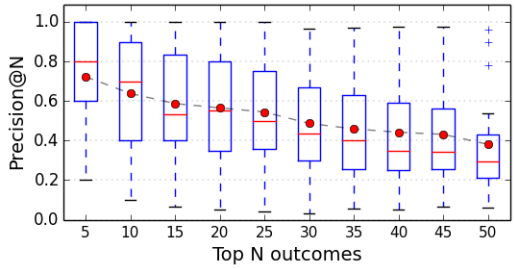
tions such as ours [73]—and a Krippendorff Alpha agreement coefficient of 0.45—indicating a moderate agreement [86].[12] We surmise that crowd-workers collectively provide reliable labels at a volume that it would be otherwise costly to obtain, concurring with Olteanu et al. [74].

Measuring Precision

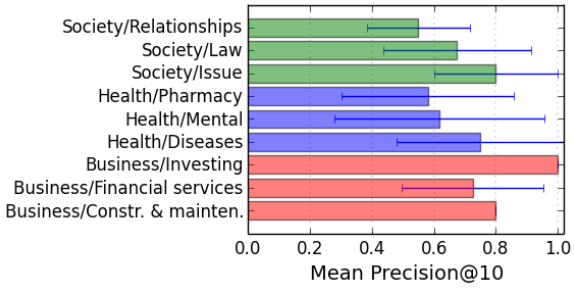
With these annotations, we define the *precision* of our results as the fraction of outcomes *perceived as* relevant to the experience: $P = \frac{\{discovered\,outcomes\} \bigcap \{relevant\,outcomes\}}{\{discovered\,outcomes\}}$. While our system outputs up to 20k likely outcomes, most applications would only consider the topmost results. Thus, we measure precision at a given cut-off rank N (precision at N, *P@N*).

Results: Figure 2(a) shows the precision variation at different cut-offs across experiments. We notice a drop of 10–20% in precision from the top 5 to the top 20 outcomes—with the median precision dropping from close 80% to about 50%, followed by a slower overall decay. Yet, even after the top 30, the discover outcomes attain an average perceived precision of over 50%. These results have two main takeaways: overall, the discovered outcomes tend to attain good precisions scores across experiences, which correlate with their effect size. Figure 2(b) shows how P@10 varies across domains—ranging from over 55% to 100% on average per

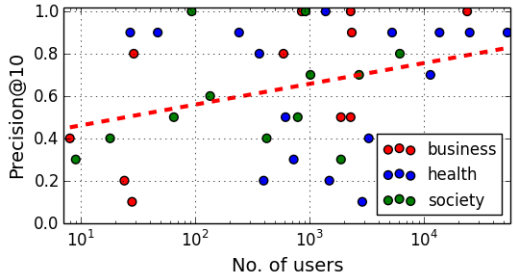
[12]Note that the Krippendorff Alpha coefficient yields lower agreement scores than the more popular Cohen Kappa cofficient (when applicable) [86]. Further, both Krippendorff Alpha coefficient and pairwise percent agreement scores significantly penalize disagreement, even when there is a clear majority that agrees on a label.



(a) Distribution of P@N values. The boxplots summarize the precision@N across all experiences in our list. Red lines represent the median, while dots the mean.



(b) P@10 across domains. Only experiences with over 30 treated users (grayed out in Table. 1) are considered.



(c) P@10 vs the number of treated users for each experience in our data collections.

Figure 2. Variations in precision (a) across top N outcomes; (b) across events and domains; and (c) with data volume. Outcomes are ranked by their z-scores.

domain, with higher P@10 being obtained for events in the business domain; and lower for the pharmacy and relationships subdomains. Figure 2(c) shows that the perceived precision varies with the data volume it was computed on. We find that data volume is a strong factor in the resultant quality of extracted outcomes. This partially explains the variance of P@10 across domains. However, other factors, such as errors in the semantic interpretation of words and domain-specific biases in the likelihood of users to mention certain outcomes might also play a factor in this domain variance.

To calibrate the quality of our results, we also compute a within-individual analysis as a baseline, comparing the words that are more likely to be mentioned after treatment as compared to before treatment by the treated users. Across all our domains, this analysis produced results with low statistical significance (only 56 outcomes across all 39 situations had a z-score ≥ 2.5). Our crowdsourced workers found the results to be of low precision as well. The top 10 results of the situations within the health-diseases subdomain, for example, had a precision of 0.19 (compared to 0.71). Inspecting the errors made by this baseline technique across domains, we found that many were due to temporally-correlated events (e.g., Mother's Day) or trending memes, that happened to occur after most individuals had mentioned the treatment experience. Our analysis, by comparing to a control group from the same period, can discount such temporal confounds.

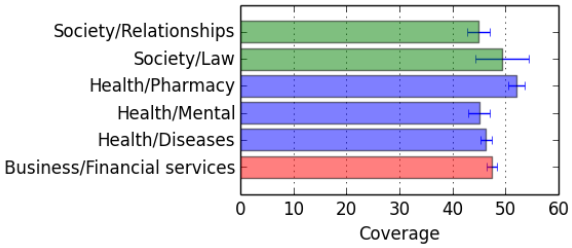
Error Analysis: The cases in which the outcomes terms were perceived as incorrect by our crowd-workers appear to be related to various extrinsic elements shaping the social media communications of different groups of users, such as the release of songs, books or movies, or other newsworthy events. For instance, our system extracted "*chiraq*" (a film, a song and Chicago's nickname) as an outcome for *taking promethazine* due to the overlap of our collection with the release of a song remix Chiraq[13] which lyrics say "*I be slowed down off promethazine*". Other outcomes perceived as incorrect include: (1) lexical ambiguous terms such as the outcome "*joint*" for users *having gout*—while people with gout can experience swelling joints, some tweets made reference to, e.g., restaurant joints or shared activities; (2) irrelevant terms such as the outcome "*vampire*" derived for users that have *installed a garbage disposal* where the outcome tweets seem unrelated to the target event, or (3) when the target event attracts the attention of organization accounts typically tweeting about related products, which appear as treated users in the corresponding collections (e.g., accounts of investment companies are included among the users that have *invested their money*).

Thereby, we note that while our use of a qualitative context to understand the specific meaning of an ambiguous term—including major splits when multiple meanings are used—helps with the interpretation and validation of the results, it does not help the statistical analysis itself. To further improve the precision of our results, additional steps to remove non-human accounts by e.g., training a classifier to detect organization accounts [66], or to apply existing natural language processing techniques that carefully resolve lexical ambiguities by e.g., considering the context in which they are used or by leveraging existing networks of concepts [13, 49], are needed before the statistical processing.

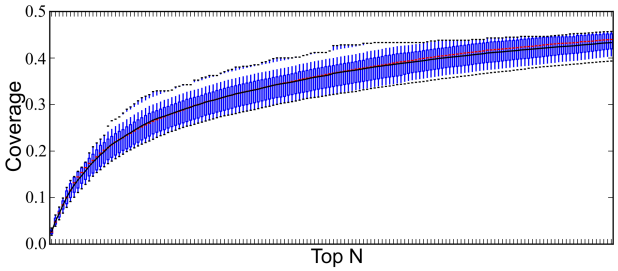
5.3 Outcome Coverage

To measure how well our extraction of treatment/outcome pairs covers known conceptual and causal relations, we contrast them with (a) available concepts and the relations they share from a large knowledge base, ConceptNet5 [91], and with (b) phrases that frequently co-occur with the experience terms in web search queries from Bing.com. This section evaluates the overall coverage of outcomes, with the coverage of relations being deferred to (§5.4).

[13] http://genius.com/Prico-x-lil-mister-x-billionaire-black-x-swagg-dinero-chiraq-remix-lyrics



(a) The percentage of covered concepts across domains. We include only the sub-domains with at least 3 experiences for which our analysis succeeded.



(b) The fraction of covered concepts according to the treatment effect for the Health/Mental subdomain (cumulative). All domains follow a similar distribution.

Figure 3. Coverage of related concepts in ConceptNet. The *concept-outcome similarity* has been set to 0.5**.**

ConceptNet5[14] is a source of general human knowledge modeled as a large semantic graph [61] including 3.9 million concepts linked by 12.5 million edges [91]. While the first version mainly relied on data from the Open Mind Common Sense project [90], now ConceptNet5's sources range "from dictionaries to online games" [91], including DBPedia [7], English Wiktionary[15], and WordNet 3.0 [38]. Each edge (or relation) in ConceptNet5 graph is associated with an *edge weight* that indicates the (quality and) confidence in the relation (the bigger the more knowledge sources confirms it).

To appraise the coverage of ConceptNet5 concepts, we limit our analysis to concepts found within a radius of 3 from the corresponding concepts to the target experience in the graph, which we refer to as *related concepts*. We do so as while one outcome (here concept) can lead to another, as we go further away from a given concept in the graph, qualitatively, fewer of the newly discovered concepts are obviously related to it (this is visible in our comparison with a higher fraction of concepts covered at a radius of 1 vs. 3, with the coverage dropping up to 40%). Furthermore, since some of the automatically inferred outcomes in ConceptNet can be spurious [91], we also limit our evaluation to those relations with a corresponding *edge weight* higher than 1. The resulting number of *related concepts* for each target experience varies from a few tens (for narrow concepts like *living trust* and *notary*) to thousands (for broader concepts like *jealousy* or *depression*).

[14]http://conceptnet5.media.mit.edu/

[15]http://en.wikitionary.org

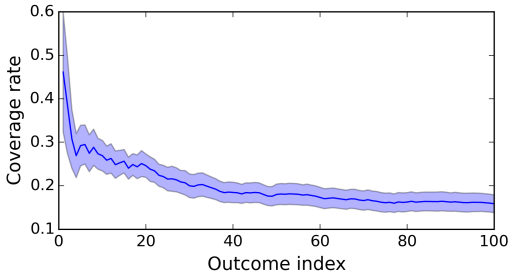


Figure 4. Fraction of outcomes that appear in the top 1000 most popular queries referring to the target experience by treatment effect rank (or index) across domains.

Coverage of concepts: Using this data we first measure the fraction of *related concepts* similar to at least one of the discovered outcomes. To assess the *concept-outcome similarity* between a discovered outcome and a *related concept* we use a tf-idf based similarity, which weights the overlapping terms by their tf-idf score treating each ConceptNet concept (typically containing several terms) as a document.

Figure 3(a) shows how the coverage of related concepts varies across domains, with the highest coverage being attained for the Pharmacy subdomain (>50%) and Society/Law domain; while the lowest is obtained for the Society/Relationships (42%–47%) and Mental health (41%–47%) subdomains—also the subdomains with the highest number of related concepts. Figure 3(b) shows we achieve higher discovery rates for outcomes with larger treatment effects, discovery rates that slow down as the treatment effect is less visible—the curve starts to flatten out after about 3000 outcomes by treatment effect, and at about 30% of related concepts covered. This trend is representative for all tested domains.

Outcomes Occurrence in Search Logs: Popular web search queries can be seen as a proxy for the general knowledge or beliefs about given topics. In other words, many users searching for, e.g., "*prostate cancer hormone therapy*" indicates that hormone therapy methods are known to be related to prostate cancer, or that users have, at least, heard about them before [77]. To understand if and how much the discovered outcomes capture common knowledge, we search them in phrases frequently co-occurring with the target experience phrase in popular web search queries, and measure the fraction of discovered outcomes that users search for along with the treatment phrase: $K = \frac{\{discovered\ outcomes\} \bigcap \{known\ outcomes\}}{\{discovered\ outcomes\}}$, where the *known* outcomes occur in the most frequent 1000 search queries that reference the experience.

Figure 4 shows how this score varies with the treatment effect. The top outcomes by treatment effect co-occur most frequently with the target experience within the top 1000 search queries by popularity (32%–60% of outcomes being covered across domains). However, we also notice a sudden drop followed by a brief increase and a slow decay. This might indicate that while a few top outcomes with the highest treatment effects are likely to be known by users, this is not the case for the rest of the discovered outcomes.

5.4 What Kinds of Outcomes are Discovered?

In ConceptNet5, the relations between concepts are categorized across a variety of types, capturing both conceptual and descriptive relations, e.g., *IsA, DerivedFrom* or *SimilarTo*, as well as more causal like relations, e.g., *Causes, HasSubEvent* or *MotivatedByGoal*. Example relations include [*xanax IsA prescription drug*], [*xanax UsedFor anxiety*] and [*divorce CausesDesire drink*]. Though all knowledge bases are incomplete, ConceptNet5's taxonomy is useful to gain insights into the kinds of outcomes we uncover.

Across all domains, the relations that get the highest coverage—typically with 40% to 65% of the ConceptNet's relations being covered—are: *HasFirstSubevent, HasSubevent, MotivatedByGoal, HasPrerequisite, CapableOf, Desires, Causes*. Figure 5 shows the distribution of coverage for the Pharmacy subdomain across relation types. In general, our results are more likely to cover causal relations, including implementation steps (*HasSubEvent, HasFirstSubEvent, HasLastSubEvent*), motivations and prerequisites (*MotivatedByGoal, HasPrerequisite*), and implications (*Desires, NotDesires, CapableOf, UsedFor*). In contrast, our results do not cover more conceptual and descriptive relationships as well, including things that cannot be done (*NotCapableOf*), and alternate names and similar actions (*DefinedAs, RelatedTo, IsA, SimilarTo*). Further, we notice that the relations for which this distinction is most prominent (with a gap of up to 6 percentage points between the original distribution of relations in ConceptNet5 vs. the one of the covered relations), e.g., *CapableOf, Desires, NotDesires*, also benefit from higher coverage rates of more than 50%.

These results indicate that our framework can distill outcomes that share a mixture of relations with the target experience. Note that different relation types might be of interest to different stakeholders. For instance, an individual diagnosed with anxiety might be interested in learning about likely, yet *not desired*, symptoms (e.g., *panic attacks* or a *nervous breakdown*), while someone diagnosed with gout may want to know that this is *related to* high levels of *uric acid* and his *joints* may be affected as a result. On the other hand, a policy-maker may rather be interested in learning about real-world use cases for various drugs—e.g., our results for Xanax indicate that while this drug is typically *used for* medicinal treatment of *anxiety*, others mention *smoking weed* and *getting drunk* around the time they take Xanax, indicating recreational usage. (emphasis added to outcomes and relations extracted by our framework)

6. DISCUSSION AND FUTURE WORK

6.1 Future Work, Limitations & Challenges

Our systematic examination of a diverse set of situations on which people report on Twitter shows that social media is a promising source of data for general-purpose outcome identification. The general patterns we found here lay the foundations for future studies exploring in greater detail specific types of situations such as related to mental health, drug prescription, or relationship issues. However, we note that there are a number of challenges that remain to be addressed, which we elaborate next and hope to address in future work:

User Timelines & Outcome Inference

Our analysis aims to identify outcomes that are more likely to be mentioned following personal experiences. However, it is important to note that while we borrow propensity score analysis from the causal inference literature, our application of this technique is not a causal analysis, as two key assumptions may not hold: First, all confounding variables must be included in the observed covariates (the terms used by a user). Yet, while high-dimensional propensity score analyses, such as ours, are more likely to capture those variables correlated with confounding variables, it is difficult to argue that all relevant aspects of individuals' lives are captured in their Twitter streams. Second, the stable unit treatment value assumption (SUTVA) must hold—that is, one person's outcome must be independent of whether another person had the target experience. However, a typical conversation on some topic may, e.g., contain retweets or use same hashtags. It is, thus, plausible that one person's use of a term could indeed have some effect on others in the community. Without additional domain knowledge to assert these assumptions, we cannot in general make causal interpretations of our results.

Further, given that our study focuses on a predefined period of three month, our evaluation is likely to miss medium and longer-term outcomes that emerge towards the end of this period or long-after. On the other side, since our analysis relies on months-long statistics, we are also likely to miss more fine granular phenomena such as at the level of hours, minutes or seconds. To distill such outcomes—entailed when interested in the effect of short-term-impact events such as drinking a coffee or taking a nap—shorter time periods should be considered for the statistical analysis.

Other future work includes extending our techniques to represent and analyze continuous-valued events and actions (for example, events where people report jogging or biking for n miles) as well as properly representing the structure and importance of repetition and duration of long-lasting experiences. Accounting for other factors known to influence outcomes, such as social support and environmental factors is also important. Better characterization of abstract outcomes, such as changes in a person's behaviors, mood, language, likelihood to engage with others or to disclose sensitive information is another rich avenue for future research. Deeper exploration of heterogeneous treatment effects is also necessary to help individuals understand the implications of situations and actions.

Data & Population Biases

Our study relies on social media timelines to depict personal experiences. This ignores population and various selection biases [47, 52]. As future research succeeds to characterize the propensity of individuals to variably report information, we may improve our analysis by incorporating this heteroskedasticity within a weighted propensity score analysis.

In addition, while we focus our analysis on a single source of social media data, we strongly believe that fusing multiple data sources is important to building a more complete picture of the outcomes of situations and actions. The reason is that different social media tend to provide a different range of in-

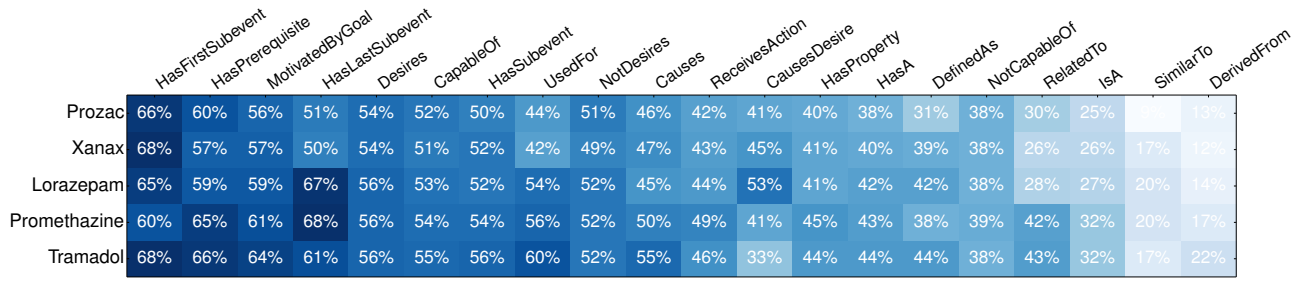


Figure 5. Distribution of coverage across the types of relations available in ConceptNet5 (Health/Pharmacy).

sights, which depends on the overall ecology of each platform (e.g., norms or functional affordances) [94]. For instance, on relatively anonymous sites like Reddit, users are more likely to make sensitive and personal disclosures, compared to more public spaces such as Twitter [89]; while data from specialty sites like LinkedIn may provide better insights about professional development after taking a certain job [95]. Our framework can accommodate and integrate alternative or multiple data sources that are temporal in nature (i.e., timestamped messages with user identifiers). The two main elements that may require adaptation are the treated users' identification and data pre-processing heuristics—depending on the characteristics of the messages shared on each platform (e.g., use of language, length, or eloquence).

Scalability & Interpretability

Furthermore, it is worth noting that while we have not discussed scalability and performance aspects of our work, ensuring that our techniques can be applied to a web-scale corpus is also critical to many application scenarios.

Depending on how they are presented, results of automated inferences such as ours may be perceived as authoritative. Thus, another important challenge is how to avoid spreading any misleading information, or perpetuating existing misconceptions—two important research questions on their own. To minimize such risks, future work should consider integrating solutions that locate and filter out social media messages deemed as untrustworthy before applying the statistical analysis (see e.g., [20]); assigning different weights to data coming from different users according to their overall trustworthiness or their topical authority [48]; or augmenting our results with additional information about the users that mention certain outcomes [69].

Finally, as others have also noted [36], alone, data-driven inferences about human behavior (even when correct[16]) are often insufficient: simply observing patterns in the data will not make them immediately useful or interpretable. An important area left unexplored in this paper is how information about outcomes can best be used and presented to aid people in specific application scenarios, and the implications of these application patterns for the analysis framework itself. A key factor in visualizing the results of our analysis will be to clearly present supporting context (e.g., qualitative sampling and contextual summarization) that support individuals and policy-makers with semantic knowledge of the domain in interpreting the discovered outcomes. Future work should also consider highlighting within strata outcomes as well (e.g., outcomes occurring for users with a similar propensity for being in a situation), as they may help understanding potentially divergent and heterogeneous treatment effects.

[16]E.g., they correctly distill true relations from spurious correlations or relations due to coincidence or lurking factors.

6.2 Potential Applications of Outcomes

Social learning theory emphasizes that learning often occurs through the "*observation of other people's behavior and its consequences for them*" [10]. Economists have also noted that when individuals face somewhat similar decisions (e.g. similar alternatives, information availability and payoffs), they look to learning from the decisions and actions of others [15]. Research on social media use has also found evidence of social learning [18]: users monitor and adapt their behavior to what other users are doing. In addition, observing the decisions of others can also explain various human phenomena such as when and why mass convergent behavior is prone to fads and errors, and can, thus, offer important insights to economists, policy-makers, and others [15, 28, 31].

This points us to an interesting juxtaposition: The potential outcomes we uncover can support individuals in investigating and learning about the situations they are in. It can provide insights about unknown or poorly understood situations, can come in support of existing decisions, or can provide different perspectives in case of familiar situations. Yet, leveraging such insights is not limited to individuals; it can also provide important cues into collective behavior and tendencies in the context of various situations, and this can be beneficial for those asking questions of societal importance.

Next, we outline these directions through a couple of examples from *our* results (§5.1). We note that while we characterize the cues that can be extracted from social media about the outcomes of different experiences, and discuss possible applications, *our work does not make financial, legal or medical claims about the experiences we examine*.

Application for Individuals

First, we believe that individuals may benefit from the kind of outcomes we uncover. For instance, prior work on online health communities indicates that new patients seek experience-based information from others in similar situations for advice, or to validate their feeling or life decisions [34, 50, 64]. In such a scenario, our work can support

users in exploring the type of issues others in similar situations are likely to be concerned with such as physical location and flare ups of symptoms when *suffering from gout*, or *cardiovascular* issues and *dietary* choices when having high triglyceride levels. Users may be interested in similar explorations for other important life events as well: e.g., people planning their retirement may be interested in knowing that others having a *pension* are more likely to mention about *taxes*, *benefits* and *health care*.

Further, even when the outcomes of an action or situation are known, aggregated statistics about their likelihood can prove informative for those seeking information about them: someone *taking Prozac* to treat a depressive episode might feel relieved to know that while the likelihood of mentioning *depression* is high among others *taking Prozac*, the incidence of these mentions tends to quickly fade away after the treatment starts. Similarly, those considering *taking Tramadol* (a powerful pain killer for around-the-clock pain treatment) may benefit from knowing that the mentions of *pain* reoccur at about one week after the mentions of *taking Tramadol*.

Apart from helping individuals understand new situations, information about potential outcomes can also be used to support them in achieving goals or making decisions. To aid goal achievement, prior research has leveraged crowdsourcing and friend-sourcing to create action plans, showing that it can help improve behavior [2, 60]. Yet, involving others in creating action plans can be taxing due to worries about disclosing information or being judged [2]. We see such techniques as complementary to our work, as mining action-outcomes from social media can reduce the manual effort required to scale their generation of action plans for a broader set of scenarios, as well as the amount of information individuals would need to disclose. Crowdsourcing has also been used to elicit common-sense contexts that can aid in social media interpretation [56]. Such mechanisms, modified for scalability, could aid our identification and interpretation of situation, actions and outcomes in social media.

Application for Policy-makers & Others

While our work is motivated primarily by the desire to help individuals understand their situations and the possible implications of their actions on a need basis, there is also an opportunity to use this kind of analysis to better understand behavioral phenomena of societal importance, third-party interventions and other policy questions. For instance, learning about the concerns (and their likelihood) of people *having a pension* within a given time period is not only informative for individuals, but it is also an important source of information for policy-makers [29]. In addition, insights such as when someone is more likely to talk about *suicide* after admitting to *suffering from depression* can be used to trigger requests for support from clinical experts [27]. Other example, for pharmaceutical and public health research, such a source of information can help with understanding drug uses that fall beyond the drug prescription (as we have found about *Xanax*).

Further, large, quantitative analyses such as ours can complement small-scale qualitative or survey-based studies of social phenomena (e.g., see [23, 34]), and vice-versa. For instance, both the design and the findings of a survey on how people cope with being diagnosed with a disease (or with any other critical life event), may benefit from insights about what topics patients are more likely to talk about after a diagnosis (e.g., *weight loss* after reporting to have *kidney stone*), as well as when they are more likely to talk about them (e.g., the mentions about *weight loss* start to become prominent after about a week). Insights about topics of interest may inform what questions are being asked, while insights on temporal dynamics may be used to align patients answers with time-dependent-episodes [39].

7. CONCLUSIONS

Through the analysis of the combined experiences of hundreds of millions of people, as reported on social media, we can gain insights into the long-tail of critical and everyday situations that individuals, policy-makers and scientists are interested in more fully understanding. We believe there are opportunities here to build a wide variety of applications that aid individuals' decision-making, goal achievement, and sense-making of unfamiliar situations, as well as complementing existing methods available to policy makers attempting to understand societally important situations as well.

In this paper, we focused on open questions related to the quality and kinds of experiences that people are more likely to mention on social media after reporting a targeted experience. To answer these questions, we studied outcomes discovered from social media after people reported situations selected from a broad variety of domains. Our results showed that the discovered outcomes attain a high precision (65–88% relevance), which correlates with their measured treatment effect, and that the overall quality of results is tied to the initial data volume, where fewer than 100s of users experiencing an outcome provides poorer quality results. We also find that causally related concepts are more likely to be discovered than conceptual or semantically related concepts. These results indicate that such outcome information is meaningful and relevant, and that social media timelines are indeed a valuable resource for understanding how a broad variety of common and critical situations unfold over time.

Ethical Considerations: Some of our analyzed domains are sensitive (e.g., mental health issues), and even efforts to help (as ours) what could be considered *vulnerable groups* should be carefully scrutinized for ethical challenges and other risks. At no point in our analysis did we identify or attempt to identify the real identities of these users—we worked with aggregated results and saw only hashed user IDs when we extracted messages for annotation. In addition, we carefully paraphrased for anonymity the text of all tweets we give as examples in this paper.

Reproducibility: For tweet IDs and the list of phrases used to locate relevant users for each of the personal experiences included in our study, please see https://www.microsoft.com/en-us/research/publication/distilling-outcomes-personal-experiences-propensity-scored-analysis-social-media.

Acknowledgments: We are grateful to Matthew Richardson for fruitful discussions, insight and expertise that greatly assisted this research.

8. REFERENCES

1. Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 3197–3206.
2. Elena Agapie, Lucas Colusso, Sean A Munson, and Gary Hsieh. 2016. PlanSourcing: Generating Behavior Change Plans with Friends and Crowds. In *Proc. of the ACM Conf. on Computer-Supported Cooperative Work & Social Computing*. 119–133.
3. Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 3895–3905.
4. Sitaram Asur and Bernardo A Huberman. 2010. Predicting the future with social media. In *IEEE/WIC/ACM Conf. on Web Intelligence and Intelligent Agent Technology*. 492–499.
5. Susan Athey and Guido W Imbens. 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74, 2 (2006), 431–497.
6. Susan Athey, Guido W Imbens, and Stefan Wager. 2016. Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing. *arXiv preprint arXiv:1604.07125* (2016).
7. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *Dbpedia: A nucleus for a web of open data*.
8. Kat Austen. 2015. What could derail the wearables revolution? *Nature* 525 (2015).
9. Ricardo Baeza-Yates and Diego Saez-Trumper. 2015. Wisdom of the Crowd or Wisdom of a Few?: An Analysis of Users' Content Generation. In *Proc. of the ACM Conf. on Hypertext & Social Media*. 69–74.
10. Albert Bandura and David C McClelland. 1977. Social learning theory. (1977).
11. Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction for the web. In *Proc. of International Joint Conference on Artificial Intelligence*, Vol. 7. 2670–2676.
12. Paul N Bennett, Krysta Svore, and Susan T Dumais. 2010. Classification-enhanced ranking. In *Proc. of the World wide web Conf.* ACM, 111–120.
13. Shane Bergsma, Dekang Lin, and Randy Goebel. 2009. Web-Scale N-gram Models for Lexical Disambiguation. In *Proc of. Joint Conf. on Artificial Intelligence*.
14. Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 237–246.
15. Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1998. Learning from the behavior of others: Conformity, fads, and informational cascades. *The Journal of Economic Perspectives* 12, 3 (1998), 151–170.
16. Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.
17. Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proc. of the Conf. on Empirical Methods in natural language processing*. Assoc. for Computational Linguistics, 562–570.
18. Moira Burke, Cameron Marlow, and Thomas Lento. 2009. Feed me: motivating newcomer contribution in social network sites. In *Proc. of the ACM Conf. on human factors in computing systems*. 945–954.
19. Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22, 1 (2008), 31–72.
20. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proc. of the World Wide Web Conf.* 675–684.
21. Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How Community Feedback Shapes User Behavior. In *Proc of. AAAI Conf. on Weblogs and Social Media*.
22. Lydia B Chilton and Jaime Teevan. 2011. Addressing people's information needs directly in a web search result page. In *Proc. of the World wide web Conf.* ACM, 27–36.
23. Wen-Ying Sylvia Chou, Yvonne Hunt, Anna Folkers, and Erik Augustson. 2011. Cancer survivorship in the age of YouTube and social media: a narrative analysis. *Journal of medical Internet research* 13, 1 (2011).
24. Phil Cohen, Adam Cheyer, Eric Horvitz, Rana El Kaliouby, and Steve Whittaker. 2016. On the Future of Personal Assistants. In *Proc. of the ACM Extended Abstracts on Human Factors in Computing Systems*. 1032–1037.
25. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Major life changes and behavioral markers in social media: case of childbirth. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work*. 1431–1442.
26. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Social media as a measurement tool of depression in populations. In *Proc. of the ACM Web Science Conf.* 47–56.

27. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013c. Predicting Depression via Social Media. In *Proc. of AAAI Conf. on Weblogs and Social Media*.
28. Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 2098–2110.
29. Ed Diener. 2006. Guidelines for national indicators of subjective well-being and ill-being. *Applied Research in Quality of Life* (2006).
30. Joan DiMicco, David R Millen, Werner Geyer, Casey Dugan, Beth Brownholtz, and Michael Muller. 2008. Motivations for social networking at work. In *Proc. of the ACM Conf. on Computer-Supported Cooperative Work*. 711–720.
31. Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proc. of the AAAI Conf. on Artificial Intelligence*. 182–188.
32. Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. 1997. Human conversational behavior. *Human Nature* 8, 3 (1997), 231–246.
33. Kate Ehrlich and N Sadat Shami. 2010. Microblogging Inside and Outside the Workplace. In *AAAI Conf. on Weblogs and Social Media*.
34. Jordan Eschler, Zakariya Dehlawi, and Wanda Pratt. 2015. Self-Characterized Illness Phase and Information Needs of Participants in an Online Cancer Forum. In *Proc of. AAAI Conf. on Web and Social Media*.
35. Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proc. of the ACM Conf. on Knowledge Discovery and Data Mining*. 1156–1165.
36. Ethan Fast, William McGrath, Pranav Rajpurkar, and Michael S Bernstein. 2016. Augur: Mining Human Behaviors from Fiction to Power Interactive Systems. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 237–247.
37. Lisa A Fast and David C Funder. 2008. Personality as manifest in word use: correlations with self-report, acquaintance report, and behavior. *Journal of personality and social psychology* (2008).
38. Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.
39. Adam Fourney, Ryen W White, and Eric Horvitz. 2015. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 737–746.
40. Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning* 37, 3 (1999), 277–296.
41. Venkata Rama Kiran Garimella, Ingmar Weber, and Sonya Dal Cin. 2014. From "I love you babe" to "leave me alone"-Romantic Relationship Breakups on Twitter. In *Conf. on Social Informatics*. Springer, 199–215.
42. Daniel Gayo-Avello. 2011. Don't turn social media into another 'Literary Digest' poll. *Commun. ACM* (2011).
43. Sharad Goel, Andrei Broder, Evgeniy Gabrilovich, and Bo Pang. 2010a. Anatomy of the long tail: ordinary people with extraordinary tastes. In *Proc. of the ACM Conf. on Web Search and Data Mining*. 201–210.
44. Sharad Goel, Jake M Hofman, Sébastien Lahaie, David M Pennock, and Duncan J Watts. 2010b. What can search predict. In *WWW*.
45. Peter M Gollwitzer and Paschal Sheeran. 2006. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology* 38 (2006), 69–119.
46. Wei Gong, Ee-Peng Lim, and Feida Zhu. 2015. Characterizing silent users in social media communities. In *Proc of. AAAI Conf. on Web and Social Media*.
47. Pedro Calais Guerra, Wagner Meira Jr, and Claire Cardie. 2014. Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proc. of the ACM Conf. on Web Search and Data Mining*. 443–452.
48. Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. 2009. Ranking comments on the social web. In *Proc. of Conf. on Computational Science and Engineering*, Vol. 4. IEEE, 90–97.
49. Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In *Proc of. IEEE Conf. on Data Engineering*. 495–506.
50. Jina Huh and Mark S Ackerman. 2012. Collaborative help in chronic disease management: supporting individualized problems. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work*. 853–862.
51. Adam N Joinson. 2008. Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 1027–1036.
52. Emre Kıcıman. 2012. OMG, i have to tweet that! a study of factors that influence tweet rates. In *Proc of. AAAI Conf. on Web and Social Media*.
53. Emre Kıcıman and Matthew Richardson. 2015. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *Proc. of the ACM Conf. on Knowledge Discovery and Data Mining*. 547–556.

54. Gary King and Richard Nielsen. Working Paper. Why Propensity Scores Should Not Be Used for Matching. (Working Paper).
55. Nicolas Kokkalis, Thomas Köhn, Johannes Huebner, Moontae Lee, Florian Schulze, and Scott R Klemmer. 2013. Taskgenies: Automatically providing action plans helps people complete tasks. *TOCHI* 20, 5 (2013), 27.
56. Yen-Ling Kuo, J Hsu, and Fuming Shih. 2012. Contextual commonsense knowledge acquisition from social content by crowd-sourcing explanations. In *4th AAAI Workshop on Human Computation*.
57. Cliff Lampe, Nicole Ellison, and Charles Steinfield. 2006. A Face (book) in the crowd: Social searching vs. social browsing. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work*. 167–170.
58. Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. In *Proc of. AAAI Conf. on Artificial Intelligence*.
59. MH Landis and Harold E Burtt. 1924. A Study of Conversations. *Journal of Comparative Psychology* 4, 1 (1924), 81.
60. Edith Law and Haoqi Zhang. 2011. Towards Large-Scale Collaborative Planning: Answering High-Level Search Queries Using Human Computation. In *AAAI Conf. on Artificial Intelligence*.
61. Hugo Liu and Push Singh. 2004. ConceptNeta practical commonsense reasoning tool-kit. *BT technology journal* 22, 4 (2004), 211–226.
62. Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. When web search fails, searchers become askers: understanding the transition. In *Proc. of the ACM Conf. on Research and Development in Information Retrieval*. 801–810.
63. Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. #FailedRevolutions: Using Twitter to Study the Antecedents of ISIS Support. *arXiv preprint arXiv:1503.02401* (2015).
64. Michael Massimi, Jackie L Bender, Holly O Witteman, and Osman H Ahmed. 2014. Life transitions and online health communities: reflecting on adoption, use, and disengagement. In *Proc. of the ACM conference on Computer Supported Cooperative Work & Social Computing*. 1491–1501.
65. Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 1999. A machine learning approach to building domain-specific search engines. In *Proc. of International Joint Conference on Artificial Intelligence*.
66. James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *Proc of. AAAI Conf. on Web and Social Media*.
67. Andrew Meola. 2016. Wearables and mobile health app usage has surged by 50% since 2014. http://www.businessinsider.com/fitbit-mobile-health-app-adoption-doubles-in-two-years-2016-3. (2016). [Online; Accessed 27-July-2016].
68. Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one* 8, 4 (2013).
69. Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work*. 441–450.
70. Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15, 8 (2013).
71. Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work*. 189–192.
72. Kevin Kyung Nam, Mark S Ackerman, and Lada A Adamic. 2009. Questions in, knowledge in?: a study of naver's question answering community. In *Proc. of the ACM Conf. on Human Factors in Computing Systems*. 779–788.
73. Kimberly A Neuendorf. 2002. *The content analysis guidebook*. Sage.
74. Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proc. of the ACM Conf. on Computer Supported Cooperative Work & Social Computing*. 994–1009.
75. Kunwoo Park, Ingmar Weber, Meeyoung Cha, and Chul Lee. 2015. Persistent sharing of fitness app status on twitter. *arXiv preprint arXiv:1510.04049* (2015).
76. Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health.. In *Proc of. AAAI Conf. on Web and Social Media*. 265–272.
77. Michael J Paul, Ryen W White, and Eric Horvitz. 2015. Diagnoses, decisions, and outcomes: Web search as decision support for cancer. In *Proc. of the World Wide Web Conf.* ACM, 831–841.
78. Judea Pearl. 2000. *Causality: models, reasoning and inference*. Vol. 29. Cambridge Univ Press.
79. Andrew Perrin. 2015. Social media usage: 2005-2015. (2015).
80. Andrew Prestwich, Marco Perugini, and Robert Hurling. 2010. Can implementation intentions and text messages promote brisk walking? A randomized trial. *Health Psychology* 29, 1 (2010), 40.
81. Davide Proserpio, Scott Counts, and Apurv Jain. 2016. The psychology of job loss: using social media data to characterize and predict unemployment. In *Proc. of the 8th ACM Conf. on Web Science*. 223–232.

82. Matthew Richardson. 2008. Learning about the world through long-term query logs. *ACM Transactions on the Web* 2, 4 (2008), 21.
83. James M Robins and Larry Wasserman. 1999. On the impossibility of inferring causation from association without background knowledge. *Computation, causation, and discovery* (1999), 305–321.
84. Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Stat. Assoc.* 79, 387 (1984), 516–524.
85. Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. In *Proc. of AAAI Conf. on Artificial Intelligence*.
86. Philipp Schaer. 2012. Better than their reputation? on the reliability of relevance assessments with students. In *Conf. of the Cross-Language Evaluation Forum for European Languages*. Springer, 124–135.
87. H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and others. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* (2013).
88. Jasjeet S Sekhon. 2007. The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods. In *Oxford handbook of political methodology*.
89. Martin Shelton, Katherine Lo, and Bonnie Nardi. 2015. Online Media Forums as Separate Social Lives: A Qualitative Study of Disclosure Within and Beyond Reddit. *iConf. Proceedings* (2015).
90. Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems 2002: Coopis, doa, and odbase*. 1223–1237.
91. Robert Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proc. of Language Resources and Evaluation Conference*. 3679–3686.
92. Diana I Tamir and Jason P Mitchell. 2012. Disclosing information about the self is intrinsically rewarding. *National Academy of Sciences* (2012).
93. Rannie Teodoro and Mor Naaman. 2013. Fitter with Twitter: Understanding Personal Health and Fitness Activity in Social Media. In *AAAI Conf. on Weblogs and Social Media*.
94. Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *AAAI Conf. on Weblogs and Social Media*.
95. José Van Dijck. 2013. You have one identity: performing the self on Facebook and LinkedIn. *Media, Culture & Society* 35, 2 (2013), 199–215.
96. Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling Self-Disclosure in Social Networking Sites. In *Proc. of the ACM Conf. on Computer-Supported Cooperative Work & Social Computing*. 74–85.
97. Xuchen Yao and Benjamin Van Durme. 2014. Information Extraction over Structured Data: Question Answering with Freebase.. In *ACL (1)*. Citeseer, 956–966.
98. Elad Yom-Tov and Evgeniy Gabrilovich. 2013. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research* 15, 6 (2013).
99. Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. 2014. Inferring international and internal migration patterns from twitter data. In *Proc. of the World Wide Web Conf. Companion*.