Towards Decision Support and Goal Achievement: Identifying Action-Outcome Relationships From Social Media

Emre Kıcıman
Microsoft Research
emrek@microsoft.com

Matthew Richardson
Microsoft Research
mattri@microsoft.com

ABSTRACT

Every day, people take actions, trying to achieve their personal, high-order goals. People decide what actions to take based on their personal experience, knowledge and gut instinct. While this leads to positive outcomes for some people, many others do not have the necessary experience, knowledge and instinct to make good decisions. What if, rather than making decisions based solely on their own personal experience, people could take advantage of the reported experiences of hundreds of millions of other people?

In this paper, we investigate the feasibility of mining the relationship between actions and their outcomes from the aggregated timelines of individuals posting experiential microblog reports. Our contributions include an architecture for extracting action-outcome relationships from social media data, techniques for identifying experiential social media messages and converting them to event timelines, and an analysis and evaluation of action-outcome extraction in case studies.

1. INTRODUCTION

While current structured knowledge bases (e.g., Freebase) contain a sizeable collection of information about entities, from celebrities and locations to concepts and common objects, there is a class of knowledge that has minimal coverage: actions. 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. While some of this information may be found within the free text of Wikipedia articles, the lack of a structured or semi-structured representation make it largely unavailable for computational usage. With computing devices continuing to become more embedded in our everyday lives, and mediating an increasing degree of our interactions with both the digital and physical world, knowledge bases that can enable our computing devices to represent and evaluate actions and their likely outcomes can help individuals reason about actions and their consequences, make better decisions and be more likely to achieve their individual goals.

In today's digitally connected world, hundreds of millions of people regularly and publicly report their goals, actions and outcomes on social media, including Twitter, Facebook and other social web sites. Such detailed records of the events occurring in people's lives provide an opportunity to learn the relationships among everyday actions, their outcomes, and higher-level goals. While there are many data sources (including web documents, search queries, and a variety of wearable sensors) that potentially capture relationships between actions and outcomes, our initial focus is on social media data for several reasons. First, status messages naturally capture the temporal occurrences of events experienced by individuals, allowing our analysis to exploit temporal relationships among actions and outcomes. Secondly, status messages capture both the actions that people take as well as their outcomes across a wide variety of domains. Finally, social media messages are annotated with persistent user identifiers that allow us to condition our results on past actions and other relevant information.

A *knowledge base of actions* has many potential applications, such as direct user exploration to aid decisions; review of recent actions and their likely future impact; and personalization of automated recommendations based on user's medium- and long-term goals. Research in the fields such as social psychology, medicine 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 [35, 17, 23]. Scaling the generation of these aids across an open-ended domain of actions and goals, tailored appropriately across populations, and then delivering them at the right time and place has to date been infeasible. However, with our computing devices continuing to be integrated more tightly into our everyday lives, and mediating more of our actions (through discovery, recommendation, purchase, guidance, tracking, etc.), embedding a knowledge base that can link available and occurring actions with their long-term consequences could enable such positive impact on individual outcomes.

This paper describes our efforts to build such a knowledge base of actions. To realize the full value of the large-scale longitudinal records of actions and outcomes in social media archives, there are many potential technical challenges that must be addressed, from interpreting and aggregating the natural language text of social media texts, to accounting for biases inherent in the data. While these are grand issues, we wonder whether straightforward approaches to these techni-

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

ACM 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2783310 .

cal challenges might yet produce reasonable and useful, if limited, representations of actions and their outcomes from social media. In this paper, we investigate that basic question of current feasibility through two case studies analyzing action and outcome relationships extracted with a general purpose analysis methodology. Our contributions include:

- An analysis framework to extract action-outcome relationships from social media data (Section 3).
- Algorithmic and implementation details for each major component of the framework, including the identification of experiential social media messages, extraction of a timeline representation of events from raw messages, and extraction of precedent and subsequent actioncome relationships (Section 4).
- Two case studies applying the techniques to Twitter data: extracting positive and negative outcomes for decision support, and identifying precedent events for supporting goal achievement (Section 5).

Addressing many other important and related issues, including social media biases, algorithmic scalability, efficacy of intervention methods, and causal reasoning, is a non-goal of this paper. These issues are briefly discussed in Section 6.

2. BACKGROUND AND RELATED WORK

2.1 Social Media Content

With the wide-spread adoption of social networking services over the last 10-15 years, much research has focused on understanding people's motivations and participatory behavior on these sites, both from a qualitative as well as quantitative perspective [27, 14, 19, 21, 13, 32, 26, 36, 6, 42]. Across these studies, common findings are that individuals are motivated to participate in social networking for a variety of purposes, including communicating and keeping up with current friends, meeting new people, managing one's professional reputation, and learning interesting new things.

This diversity of purpose on social networking sites has lead to a broad a variety of content being found within social media messages. Even in this variety, however, status messages reporting on an individual's own experiences constitute a significant percentage of content. Naaman, Boase and Lai categorize tweets and find that such "me now" messages, describing personal state and current experiences constitute 40% of messages [32]. Ramage, Dumais and Liebling perform an large-scale latent dirichlet allocation (LDA) analysis of Twitter messages at a word-level, and find that on average, tweets are composed of 11% substance, 5% status, 16% style words, 10% social and 56% other (other includes non-English words, many numbers, dates and times) [36].

This well-documented behavior of individuals announcing and discussing a broad range of their current activities and status in social media is one of the key features of social media datasets that promises to enable the work in this paper.

2.2 Mining Social Media and Search

Much research has focused on extracting and validating information and relationships about the off-line world from social media, search queries and other digital traces of human activities. In the health domain, social media studies have looked at the relationships between diseases, medicines, side-effects, and symptoms [33, 31] as well as disease transmission [40]. Similar studies have been conducted in urban informatics [8], mental health [9, 16], natural disaster monitoring [11, 41], and other domains. Many of these analyses rely on a co-occurrence analysis: the assumption is that items that co-occur frequently may share some true relationship. For example, Sadilek et al.'s analysis of disease contagion infers relationships between disease carriers and new infections based on co-visited locations. Paul and Dredze studied the relationship between mentioned ailments and the geographies in which they occur. Becker et al. analyze social media data to surface information and insights about real-world events [3].

Studies with similar goals have been applied to search query logs and other data sources. Richardson uses long-term query logs to identify topical and temporal relationships about the world [37]; [45] and [44] extract relationships between drugs and possible consequences (adverse reactions) from search queries. A closely related body of work frames the problem of learning about the real-world from social media, search and other data sets as a prediction problem. Given a known (historical) signal about the world, the goal is to predict the current or future signal from current social media signals. This approach has been applied to prediction of economic, financial and other signals [4, 7, 15, 2, 1].

Our goals are to extend this prior work by focusing on extracting action-outcome information from individual-level timelines at relatively fine granularity. More importantly, our goal is to explore generalizable techniques that require minimal information about specific actions, domains and outcomes.

2.3 Actions and Plans

Recently, there have been several attempts at using crowd sourcing techniques to create action plans to aid goal achievement. Law and Zhang use crowdsourced workers to generate simple plans related to the "high-level missions" driving search queries, and evaluate the effect of replacing search engine results for the original query with web resources related to the various steps required by a plan [28]. They find that organizing web resources in this way is useful for helping users navigate the space of their problem.

Kuo, Hsu and Shih use crowdsourcing to elicit the commonsense context that can aid in social media interpretation [25]. Mechanisms such as this, perhaps modified for scalability, could aid our identification and interpretation of events, actions and goals in social media. Kokkalis et al. describe a system to provide individuals with actionable and reusable plans, to see if plans generated by others are as effective at improving goal achievement as plans generated by oneself [23]. They find that, indeed, system-provided plans do have a positive effect on goal achievement.

We find the effectiveness of these techniques to improve goal achievement to be promising. We see these techniques for crowdsourcing action plans as largely complementary to mining action-outcomes from social media data, and believe that an existing knowledge base of actions could reduce the required manual effort to scale out the generation of action plans for a broader set of scenarios.

3. KNOWLEDGE BASE OF ACTIONS

In this section, we define the problem of extracting action-outcome relationships. We present details about the implied

subproblems and discuss how this framework can be used to formulate a variety of interesting questions.

3.1 Choice Exploration and Goal Achievement

We consider two major types of questions one might want to ask: **choice exploration**, and **goal achievement**. For the former, we can help by advising the user what experiences to expect after taking a particular action (based on other people who have taken this action). For the latter, we can convey which actions are most likely to lead to the desired goal (based on other people who have accomplished the same goal). Since the social data is open-domain, these two topics cover a broad range of questions one might have.

One way to measure online users' desire to answer such questions is by looking at the queries they submit to a search engine. Many of these are decision questions beginning with "should I/you". The most common ones show their breadth of topic, including finance, relationships, and health: *should I refinance my mortgage*, *should I date a co-worker*, *should you marry your best friend*, *should I get a flu shot*, *should I file bankruptcy*, *should I upgrade to windows 8*. We also see many people asking for advice between two options, as in: *should I lease or buy a car*, *should I file married jointly or separately*, *should I eat before or after working out*, and *should I call him or wait for him to call me*. In both cases, we would like to provide people with the ability to see what experiences other people tend to have after taking one of the actions. For example, among those people who ate before working out vs. after working out, who was most likely to lose weight or get a side-ache, and what other unexpected effects might differ between the two populations?

Similarly, people show a desire for help in achieving goals. The most common question containing the word "marathon" is *how to train for a marathon*. Other common "how to" questions include *how to lose weight*, *how to draw*, *how to get pregnant*, and *how to speak Spanish*. As with decision support, we could provide people with the ability to see what actions were more commonly taken among those who accomplished their goal than those who didn't.

Though there may be online resources devoted to answering some of these questions, using social data has many distinct advantages. First, results are grounded in the real experiences of users who have taken an action, potentially leading to more reliable results than simply reading advice from web pages. Second, a question may be too rare for someone to have devoted writing advice about, but still have plenty of social data to answer via data mining. For example, someone may ask whether to move to one city vs. another. Web pages may exist to answer such a question for some city pairs, but surely not for any pair of cities that may be asked. In contrast, we need only look at social postings from people who have moved to one city vs. the other and compare their postings to see the relative benefits of each. Third, an answer may be contextually dependent on the asker. To the extent that we can infer demographic information for social media users [24], we can provide answers not just in the abstract, but specifically tailored to the asker: *people similar to you (urban male, age 25-35) have found that a low-carb diet works best for losing weight.*

3.2 Problem Definition

A key advantage of applying our techniques to social data is that it is fully open-domain. Social data contains experiences about anything that users wanted to post about, and as a result contains information on an incredibly wide range of topics. A sampling of the experiential tweets contained reports on love and relationships, food and alcohol, children, sleeping, weekends, weather, school, health, and so forth. A key goal in our problem definition and architecture is to ensure that our techniques match the open-domain nature of the data set and problem domain. Thus, our knowledge base of actions is simply an architecture for answering questions based on a large corpus of social data.

We formalize this core problem as follows: Given a corpus of social media messages and a query defined by two events, E^+ and E^-, our goal is to identify the precedent and subsequent relationships of an event E^+ that distinguish the social media timelines containing E^+ from timelines comparing some event E^-. Semantically, E^+ and E^- can be thought of as identifying either positive and negative outcomes or treatment and control classes. A class of events E^+ or E^- is specified as, for example, some specific observation, or a complex matching function.

Depending on the specific query we choose, we can ask different forms of high-level questions.

Choice Exploration: If we choose a query such that E^+ selects a specific action (and E^- selects an inverse action or null action), then the results from our analysis will identify what is likely to happen after taking the specified action.

Goal Achievement: If instead we choose a query such that E^+ selects the achievement of a specific goal (and E^- selects the non-achievement of that goal), then the precedents identified by our analysis will identify what is likely done and differentiates between people achieving the goal and not achieving it.

While this query setup is straightforward, there are subtleties in the selection of query specifiers. For example, if we which to explore how people achieve some goal E^+, we will find different results if we compare to an E^- that captures timelines of people who attempted but failed to achieve a goal; versus if we compare to an E^{-*} that captures timelines of people who never even tried to achieve the goal. The choice of E^- depends on the question that one wants to answer.

3.3 Architecture

Figure 1 shows the pipeline of data processing steps in our analysis. We begin with a corpus of social media messages. These messages consist of the original microblog text posted by individuals. We expect these messages to include at least a user identifier and a timestamp, but they may also include other metadata, such as includes geographic location, author details (name, brief biographical description, popularity statistics), as well as social network connections.

First, from this corpus of social media data, we extract a large set of timelines of event occurrences. Each timeline represents events occurring in a single individual's life. 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 background events that happened due to unrelated causes. These events may be directly extracted from individual social media messages, or inferred from the corpus as a whole.

By avoiding an explicit categorization of events as being actions or outcomes, we greatly simplify the task of generating timelines for individuals. Leaving this classification and

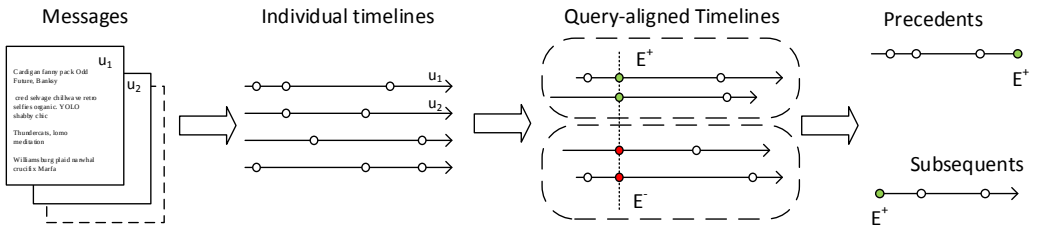

Figure 1: Steps of our general analysis

interpretation of actions and outcomes outside of the core data representation and analysis mechanics simplifies our task, at the cost of potentially requiring additional semantic understanding at higher-levels. We believe that this is likely to be a beneficial trade-off as adding additional semantics when grounded within a specific application context is often easier than building a general-purpose recognizer up-front.

In the next step, given a query, E^+ and E^-, we extract and temporally align a set of timelines that match the criteria E^+ and a set of timelines that match the criteria E^-. Representing a query as two distinct events, E^+ and E^-—as opposed to comparing a single event class against a background model of all timelines—provides significant flexibility to ask a broader range of questions of our collected data.

Finally, from these two sets of event timelines, we extract the precedent events and subsequent events that distinguish the E^+ and E^- timeline subsets from each other.

3.4 Subproblems

There are a number of implied subproblems within the key tasks of event timeline extraction, subselection of timelines according to a query, and identification of precedent and antecdent events, including:

Identification of experiential messages: When extracting a timeline of events experienced by a person, the first thing we must do is identify *experiential messages* which report on personal experiences of the author, whether past, current or (expected) future. Non-experiential messages include conversational texts, hearsay, pointers to news articles and current events, among others. We describe our method for identifying experiential tweets in Section 4.1.

Timestamping event occurrences: While many social media messages provide *in situ* reports of an individual's experiences, it is not uncommon for authors to also report on past experiences and anticipated future experiences. For this reason, it is important to identify the time period referred to in a message, and timestamp the recognized events. We describe our approach and findings in Section 4.2.

Recognition and canonicalization of events: A key step in the generation of a timeline of events is the extraction of events from the text of social media messages. These events may be extracted directly from the textual representation of a message, or inferred from multiple messages. We discuss the former in Section 4.3 and provide an example of the latter in our second case study, in Section 5.3.

Identification of precedent and subsequent events that distinguish the two sets of timelines from each other. Our framework allows for various implementations, from correlational to causal analyses. Note that even when calculated using causal analyses, such as propensity score matching, it is unlikely that the strong assumptions necessary for inferring causality would hold (i.e., assuming the observability of all potential causal factors). Section 4.4 describes our implementation.

Identification of positive and negative valence of events: Of course, some outcomes of actions are good and others bad. In social media, messages describing such outcomes are often augmented with clear emotional words that signal the current mood of the author. Detecting these moods or sentiments and associating them with outcomes can help with reasoning about their significance. We use a domain-agnostic affect extractor, described in [10], to extract the author's levels of joviality, sadness, fatigue, hostility, etc. While we do not describe details here, we demonstrate its application in Section 5.2.

4. ANALYSIS DETAILS

In this section, we present the details of our framework, its specific application to Twitter data, and how we adapt and apply existing algorithms to address the challenges of extracting action-outcome relationships. In addition, we highlight key descriptive statistics of Twitter social media relevant to our overall tasks, including the percentage of Twitter messages that are experiential tweets, and the prevalence of relative time references.

4.1 Experiential Tweets

Social media fulfills a diverse set of roles, including experiential tweets that report on actions and events occurring that individuals are experiencing first hand, but also includes the dissemination of information about broader news and other world events, chit chat with friends, and incitements to action and advocacy [32, 5, 26, 19]. To extract action-consequence relationships, we must be able to distinguish experiential tweets from other social media content.

We tackle this as a straightforward classification task. We label ≈ 10000 messages using crowdsourced workers, asking them to specify whether or not a message is a "personal experience", defined as

> A message where the author is describing or indicating their own personal experience, such as an action or situation that they are currently in, have experienced, or are concretely planning to take in the definite future.

We explicitly instruct workers not to mark messages as personal experiences if they describe or declare personal desires or intents unless describing a concrete plan or action.

Personal Experiences
Just completed a 15.72 km run with @RunKeeper. Check it out! <URL> #RunKeeper
Just to set the mood I brought some Marvin Gaye and Chardonnay.
lacrosse is so much fun why didn't I start earlier lol
Oh yeah guys we got a new puppy.
@Alice Tell me about it. Knee isn't hurting today, but it's also taped within an inch of its life.
Other (Personal desires and goals)
When i turn 16, i'm driving anywhere and everywhere.
Hope you enjoy England! Wish i could go :(
I wish I could cook
I've got real big plans and such bad thoughts
Other (news, 3rd-person, misc.)
New campaign to protect children from second hand smoke launched... <URL>
Whoa. The kid from Cincinnati just suffered a horrible injury. Not good.
@Bob I hear you.
@Charlie did you enjoy your night at the club?

Table 1: A sample of experiential and non-experiential tweets.

Label	Count	Pct
Personal Experience	2580	26%
Other (Personal Desire/Goal)	755	7.6%
Other (news, 3rd-person, misc.)	6583	66%
Total Tweets	9873	100%

Table 2: Experiential tweet labeling results

To reinforce this, we ask workers to label the non-personal-experience tweets as either being a personal goal or other. Table 1 shows example messages for each class of labels.

We train a naïve Bayes classifier on these labeled messages, using maximum likelihood estimation for the NB parameters. We tokenize the messages based on whitespace, removing all non-alphanumeric characters, but not applying any stemming. We generate a feature t for every pair of co-occurring tokens in a message.

As shown in Table 4.1, the great majority of tweets labeled by our workers are found to be non-personal, other tweets. 26% of messages describe personal experiences. The primary implication for this paper is the confirmation of prior research that a significant amount of the data in Twitter is describing the kind of personal experience that is relevant to our learning of actions and outcomes. To measure the difficulty of the labeling task, we also collect two additional labels for each tweet. The inter-annotator agreement, measured by Fleiss' kappa, is 0.325, which is regarded as "fair agreement". For the remainder of the paper, we ignore the distinction between desire/goal and other, since we care only about whether a tweet is a personal experience or not.

4.2 Temporal expressions

Personal experiences are not always reported on social media as they occur. Often, people will post about an upcoming event or experiences in anticipation ("I can't believe my marathon is coming up next week"), will reference a recent event ("We got a cat yesterday") or a long past event ("I ran my first marathon ten years ago"). As noted by Ritter et al., building up a true timeline of event occurrences requires resolution of the temporal expressions accompanying such non-concurrent personal experience reports [38].

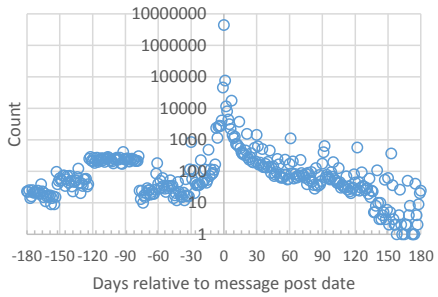

Figure 2: Distribution of relative time mentions

To do this, we built a simple rule based system, similar to TempEx [29], that can recognize and resolve basic expressions of relative offsets ("yesterday", "next weekend"), as well as references to nearby days and dates ("Tuesday" and "Feb 10th"). Figure 2 presents the distribution of relative times mentioned in our data sets. We see that most messages, by default, refer to the current date, and a large number refer to dates within a few days of the current date. As we look to dates further afield, we see more references to future events, and also spikes of references at week and month unit distances.

4.3 Event Extraction

Once we have identified a timeline of messages referring to the personal experiences of an individual, we wish to break apart each message into the component representations of the events (both actions and outcomes) that are being reported. This task is analogous to the task of named entity recognition [18], and shares many of its challenges, including candidate identification (what words in the message refer to an event of interest), disambiguation (when a candidate could mean multiple things, which does it mean) and canonicalization (can we recognize when two candidates with different forms are referring to the same underlying event).

Given how little information we have about what might constitute an action or an outcome and because our goal is an open-domain system, we make a design decision to simply extract all phrases of the message as potential events, without attempting to classify them as actions, outcomes, or neither. An advantage of using phrases instead of single words is the implicit sense disambiguation provided. For example, while the word spaghetti often refers to an Italian noodle dish, it sometimes is used as part of the name 'spaghetti squash'. Recognizing the phrase as the unit reduces the need for additional sense disambiguation.

We maintain an open-domain approach to phrase segmentation and the canonicalization of phrases into events:

Phrase Segmentation: We use a statistical modeling approach to infer the hidden phrase boundaries in a text. To efficiently find phrases, we use a phrase unigram language model, as described in Jin et al. [20]. Briefly, each token in a phrase unigram language model consists of one or more

Cluster name	Elements
cat_eats	bit_my_ear, bit_my_nose, bit_my_finger,...
woke_up_at_1	woke_up_at_3, woke_up_at_4,...
sleeping_on_my_bed	sleeping_on_my_lap, sleeping_on_my_chest
cheese_balls	cheese, cheese_pizza
loud_people	people_crazy, people_suck

Table 3: Example of phrase canonicalization. The most frequent element is selected as the cluster name.

white-space separated words. By encoding multiple words within a single unigram, the phrase language model is able to capture long distance relationships without requiring high Markov order statistics and concomitant large models. The phrase unigram language model itself is trained from a large corpus of text (in this case, from a complete archive of 16 days of tweets), using an EM process that iteratively segments a corpus into likely phrases and then retrains a new phrase unigram language model [1].

Given a phrase unigram language model, identifying phrase segmentations in a message is a matter of searching for the most probable combination of component phrase-unigrams. Below are segmentations of 2 sample messages:

It's gorgeous outside | so I'm pretty sure | I have no excuse not | to get this | long run in.
I got a new kitten | and he has blue eyes and | stripes and | I need a good name | but nothing | that's normal

Canonicalization: Generally speaking, there are many alternative ways to describe or report on a personal experience when writing a social media message, leading to the need to identify and canonicalize phrases with substantially the same meaning. To do so, we cluster phrases based on their distributional similarity. Specifically, for each phrase, we build a distribution of co-occurring (single-word) tokens. We use agglomerative hierarchical clustering to group together all phrases that are within a distance threshold d of each other, where the distance between two phrases is measured as the cosine similarity between their token distributions. (We use $d = 0.75$ in our experiments). Table 3 shows example phrase canonicalizations.

4.4 Precedent and Subsequent Events

There are multiple methods to identify the distinguishing precedent and subsequent events when comparing timelines containing an event E^+ to those containing an event E^-. In this paper, we report our experiences with two methods: a simple correlational analysis, and a correlational analysis with semantic scoping. These two techniques make different assumptions and are appropriate for different purposes.

Correlational Analysis: Our first technique looks at simple correlations between a target event and the events that precede or follow it. Our goal in this analysis is to find events that are more correlated with occurring before or after E^+ (but not both before and after) than occuring before or after E^-. As shown graphically in Figure 3, our goal corresponds to finding events that are more likely to occur in one quadrant (say, E^+ for $t > 0$) than in its immediately neighboring quadrants (E^- for $t > 0$ and E^+ for $t < 0$) [2].

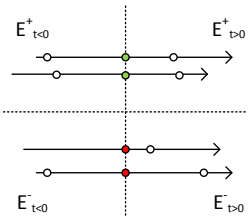

Figure 3: Quadrants of our two sets of timelines

More formally, we begin by defining the pair-wise comparison of likelihoods of an event occurrence between a target quadrant q and a neighboring quadrant u. Let $N_q(e)$ be the number of occurrences of an event e in a given quadrant, $|N_q|$ be the total number of events in a quadrant, and $\hat{p}_q(e) = \frac{N_q(e)}{|N_q|}$.

Our score, $S_{q,u}(e)$, is the relative likelihood of an event occurrence in q as compared to u. We calculate this as:

$$S_{q,u}(e) = \frac{\tilde{p}_{q,u}(e)}{\hat{p}_u(e)} \quad (1)$$

where $\tilde{p}_{q,u}(e)$ is the Laplace-smoothed probability:

$$\tilde{p}_{q,u}(e) = \frac{N_q(e) + \hat{p}_u(e)m}{|N_q| + m} \quad (2)$$

Smoothing the likelihood of $\hat{p}_q(e)$ toward the neighboring quadrant has the effect of requiring greater evidence of a difference in likelihood to appear significant. In our experiments, we set $m = 10^4$. For an event to be considered important, we require $S_{q,u}(e) \gg 1$ for both neighboring quadrants. For example, when considering an event in the quadrant $E^+_{t>0}$, we will calculate the score for both $u = E^+_{t<0}$ and $u = E^-_{t>0}$. The final reported score is the minimum of the two.

Correlational analysis has the advantage of being straightforward and requiring no inputs beyond the definitions of E^- and E^+. Because it is not a causal analysis, however, we expect its results to be better suited for tasks such as predictions which do not require a causal interpretation. Furthermore, correlational analysis may find relationships that are difficult to easily explain or interpret, and thus may not be appropriate for end-user facing applications.

Correlational analysis with semantic scoping: Our semantic correlation is the same as the correlational analysis above, with the added restriction that we only consider those events that are believed to be semantically closely related to our domain of interest. Let us define E' to be a set of events known to be in our domain then we will consider only e_i that co-occurs at least once with E' in our corpus.

Semantic correlation makes an assumption that if an event e_i is related to our target events E^+ and E^-, then at least one person would have clearly mentioned e_i in the recognizable context of our target domain. Our expectation is

[1]The MSR Phrase Breaker Service is available for demonstration and programmatic access at http://weblm.research.microsoft.com/PhraseBreakerDemo.aspx

[2]Recall that all of timelines were aligned such that the events E^+ and E^- occur at time $t = 0$

that the ranked events e_i will be more robust to noise and confounds. Furthermore, we expect that any events found to be correlated is more likely to be easily interpretable by humans, due to the enforced domain proximity. The cost, however, is that we essentially extend our query model to require a specification of the domain of interest.

While the outcomes of actions can vary based on context, our analyses are context-independent. Extending them to incorporate individual demographics, past actions, location, seasonality, social and other contextual information is important future work.

5. CASE STUDIES

In this section, we present two case studies extracting various forms of action-outcome relationships from social media data. First, we demonstrate an example of subsequent event analysis. We evaluate the quality of analysis results and measure the quality reduction when experiential message filtering, phrase clustering, or semantically scoped correlation are removed. Secondly, we demonstrate an example of precedent event analysis, where we measure the increase in likelihood of goal achievement given the occurrence of a precedent event. We ground our first case study in identifying the consequences of pet adoption, and the second in achieving the goal of running a marathon.

5.1 Data

While we are designing our architecture to process a full, unfiltered archive of social media data, our first small-scale implementation demonstrates and evaluate the feasibility of the techniques through archive subsets. For our first case study, we create an archive subset of the timelines of English-language Twitter users who mentioned getting a dog, cat, puppy or kitten during the period of August 1-15, 2013. This procedure identified 6232 Twitter users who had mentioned adopting a pet. We then collected the entire Twitter timelines for these users from the period of August 1-September 15, 2013, encompassing a total of 4.6M tweets.

For our second case study, we create an archive subset of the timelines of English-language Twitter users during the period of March 1-31, 2014 who mentioned running or training for a marathon. We then collected 2 month timelines for each of these users, from February 1-March 31, 2014. In total, this data set consists of 40,591 users and 21M tweets, with retweets removed. In addition, we used a random sample of 260M tweets to provide background statistics.

5.2 Subsequent Events and Choice Exploration

In our first case study, we wish to test the basic components of our analysis pipeline to better understand the quality implications of each analysis stage: Namely, how important are the subtasks of identifying experiential tweets and canonicalizing phrases with similar meaning? How much perceived benefit is there to restricting precedent and subsequent events to those with a semantic correlation to the target domain?

To do this, we ground our study in the specific task of automatically generating a "pros and cons" list to aid people deciding whether or not to adopt a kitten or cat. A "pros and cons" list is a simple decision making aid for clearly evaluating the benefits (pros) and disadvantages (cons) of taking some action (in this case, adopting a pet). Writing a pros and cons list is often recommended to individuals facing a significant decision to ensure that all potential consequences are considered and evaluated.

In this case study, we apply our analysis techniques to automatically extract the subsequent events that follow declarations of pet adoption in social media timelines. More formally, our query consists of an E^+ that consists of a boolean OR search for the following phrases: {"got a *pet*", "got a new *pet*"} where *pet* is either "cat" or "kitten". The set of E^+ timelines consist of all messages written by users who wrote a tweet matching E^+. In this query, our E^- is the null event, capturing all timelines—essentially a background model of user timelines. The semantic scoping of our correlational analysis consists of limiting our analysis to those events that co-occurred at least once with the main topic words "cat" or "kitten".

Table 4 shows the top entries of the pros/cons list generated by our system. We split outcome events into pros and cons by looking at the aggregate affect valence of all mentions of these outcomes across all of our E^+ set of timelines. Events with a valence > 0.6 are added to the pros column, and < 0.4 are added to the cons column. Events are ranked by their relative likelihood of occurrence, as compared to their occurrence in E^- timelines.

To evaluate the importance of each of the analysis stages, we regenerate our pros/cons list while disabling aspects of our pipeline, one at a time. First, we disable experiential tweet classification, and keep all tweets for analysis. Second, we disable phrase clustering and treat all distinct phrases independently. Third, we switch to correlation analysis, instead of semantic correlation.

To evaluate the quality impact of disabling each of these aspects of the system, we post the items of each of the 4 generated pros/cons lists for evaluation by crowdsourced workers. For each item, we display to workers the event title, and 3 messages mentioning the event (Table 4 only shows 1 message due to space limitations). We then ask workers to label, on a scale of 0 to 4 whether or not each item and messages are useful and relevant to deciding whether or not to adopt a cat. We use these labels to calculate a discounted cumulative gain (DCG) score for the entire set of results: $DCG_p = r_1 + \sum_{i=2}^{p} r_i/log(i)$, where r_i is the label at rank position i, and DCG_p is the accumulated score at rank position p.

The results provide interesting insights into the role that each stage of the pipeline plays. Our complete pipeline achieves the highest DCG score, of 20.7 summed across both the pros and cons list. Disabled-Experiential filtering is the 2nd best variation with a DCG score of 19.5. The results are very similar to our complete pipeline, though there are ranking differences and several results related to cat videos. Our pipeline without clustering is the third best variation, achieving a DCG of 16.0 after discounting duplicate items. Significant semantic duplication of results is the biggest drawback to not clustering phrases. Finally, our fourth variation of regular correlation achieves the worst performance, with a DCG of just 0.38. Most of the items found by this variation are not clearly related to cats or kittens at all. While this may be due to the relatively small data sizes, it is a striking result nonetheless, and emphasizes the importance of perceived topical relevance and the important need for an end-user to understand why correlations exist in results.

	Pros					Cons			
	Event	**Example message**	**PosNeg**	**RL**		**Event**	**Example messages**	**PosNeg**	**RL**
1	cat named	We just got a cat and named it Versace	0.70	9.3x	1	ran upstairs	But I ran upstairs and fell and now my head hurts	0.20	9.5x
2	I've got a cat	I've got a kitten asleep on my lap, and my heart has softened.	0.67	7.3x	2	damn kitten	Had practically no sleep because the damn kitten kept going nuts and runniy round my room	0.22	6.2x
3	Love my new kitten	I love my new kitten	0.88	7.2x	3	cat is literally	My cat is literally the devil	0.31	5.9x
4	named my cat	I named my cat tapenga if that's how you spell it	0.63	6.1x	4	cat just ate	My cat just ate something off the floor I don't know what it was gross	0.24	5.8x
5	love the fact that	Love the fact that our kitten Marley has a massive "M" on his forehead	0.64	5.3x	5	cat just jumped	My cat just jumped on me and scratched me	0.21	5.7x

Table 4: Top positive and negative events observed to occur after new cat ownership. PosNeg is the mood valence (1=good,0=bad). RL is the relative likelihood of the event occurring, compared to timelines where a pet adoption did not occur within our observation period.

5.3 Precedent Events and Goal Achievement

In a second analysis, we consider the effect of selected precedent actions on a specific, declared goal. In particular, we choose to look at the relative importance of various marathon training actions on the eventual outcome of a marathon race.

5.3.1 Marathon Event Identification

In the first case study we exclusively analyzed events explicitly mentioned in social media messages in an open-domain way, only requiring the user to specify four phrases and two keywords. Our second case study demonstrates our pipeline's ability to incorporate higher-level events, namely, marathon participation inferred from information mentioned across multiple social media messages. We infer the date of a marathon for individuals who have been tweeting about their training, but do not explicitly tweet about their race on the day of their run. Secondly, we report on experiments learning correlations between marathon training actions and declarations of personal record achievement.

We use official marathon result data from www.marathonguide.com to label a small set of 558 Twitter user timelines with the specific dates on which they ran a marathon by matching on the person's name and mentioned race. From these labeled timelines, we train a classifier to detect marathon dates. The features for the classifier included tokens used in tweets during a 3-day sliding window before and after the official marathon date, and tokens used in tweets that used temporal expressions to reference a date within 3-days of the official marathon date. Using these features, we built a hierarchical classifier by first estimating the likelihood that any given day was a "immediately-before-marathon" day or an "immediately-after marathon day". Then, we learned a logistic regression classifier over these estimates to find the most likely actual marathon date. Our final classifier is able to identify the true marathon date for 83% of a held out set of 42 test users within an average of 1.3 days of the actual day. The remaining 17% are not assigned to any marathon day. We applied this classifier to our entire data set and identified 1436 individuals with identifiable marathon dates during the month of March 2014.

Once we have inferred a marathon date for a user, we insert an artificial <`inferred marathon event`> symbol into the user's timeline. Without this additional inference step, we could certainly rely on explicitly mentioned marathon phrases, such as "ran a marathon today". However, implicit event identification enables us to further recognize individuals who have, for example, mentioned their excitement before a marathon and their soreness and exhaustion afterwards.

5.3.2 Measures of Marathon Success

While there are certainly several ways that individuals might determine the success of their own marathon, we use a simple definition here: whether the individual declared that they achieved a personal record (PR) after running the marathon. Our query E^+ is a boolean AND search for the phrases "PR" and <`inferred marathon event`>, where the latter is the event identifier output by our marathon date inference described above. E^- is a boolean AND search for <`inferred marathon`> and NOT "PR". Against this, we measure the correlation between a person tweeting about taking a specific training action (whether they chose to "taper", trained with "long runs", ate carbs before the race) and reporting that they achieved a personal record. Table 5 shows the results. Overall, we found that reporting the action of going for long runs and tapering (reducing exercise before the marathon) were most correlated with later reporting a personal record. Reporting eating carbohydrates (carbs) before the marathon had a minor effect as well.

Figure 4 shows the temporal dynamics of these precedent actions. Such a visualization could be useful for understanding when people take actions. For example, we see that people eat carbs the day or night before their race; go on long runs weekly for many weeks before the race; and taper their exercise 7-10 days before their race.

6. DISCUSSION

There are, of course, several challenges that our presentation above has so far elided. For example, relying on experiential social media data to learn outcomes can introduce bias due

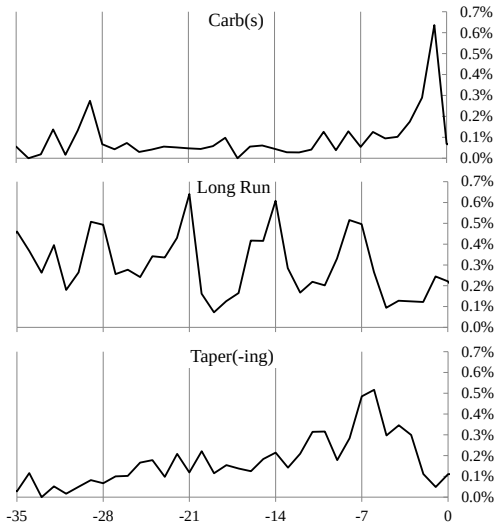

Figure 4: Temporal dynamics of carbs, long run and taper mentions. The y-axis is the percentage of tweets on a given day containing the phrase(s).

Action	Increase in PR likelihood
Carbs	+9%
Long run	+27%
Taper	+45%

Table 5: Actions reported by marathon runners on Twitter and the relative increase in reporting a personal record.

to population and self-reporting biases [30, 12, 22]. Significantly, the absence of an event in our social media timeline does not necessarily mean that an event did not occur. Understanding the implications of previous empirical studies for our inference processes, as well as the implications for how such biases circumscribe our ability to learn parts of the semantic space of relationships is important future work.

In our pipeline, we currently ignore much of the semantics of the language people use, in favor of a simplistic approach of treating all phrases in experiential tweets as candidate events in a person's timeline. Considering additional semantics and even interpreting people's own statements of causal inference, is a potentially rich area for future exploration.

An important challenge is that a true action-outcome model is essentially a model of causal relationships. There is a rich literature on the inference of causal relationships from purely observational data [43, 34] though there is debate about the reliability of causal inference in the absence of randomized, active intervention [39]. Luckily, at least for some initial applications of these models, inference of the true causal relationships seems likely unnecessary and simpler analyses such as temporal prediction and propensity scored relationships may be sufficient for the extracted results to be useful.

An area left largely unexplored in this paper is the question of how information about actions and their outcomes can best be used to aid people, and the implications of these application patterns for the action-outcome extraction pipeline. For example, many decisions involve comparing multiple choices, rather than the two-sided choice implied by the query E^+, E^- in our pipeline. Our pipeline will have to be adapted to such scenarios—perhaps through all-pairs comparisons, or multiple comparisons to a single base case.

Perhaps a more immediate consideration is whether or not the results of a particular algorithm are appropriate for a particular application or user interaction paradigm. We saw in our first case study that regular correlational analysis, when not scoped to a semantic domain, generated results that were not interpretable and marked as irrelevant by our labelers. It is quite possible that such correlations would have worked well if an application called for predictive power. But in the context of an end-user interface, the human interpretability of results is paramount. Better understanding of how to ensure results are interpretable, through correct presentation, supporting information and scoping as necessary, is an important area for further study.

Closely related to this issue is that of *actionability*. If we are to recommend actions, as we might be tempted to do based on the precedent analysis in our second case study, we must ensure that the actions we are recommending are feasible. For example, the event most predictive of a successful marathon outcome might be the simple declaration that the author "loves running!". However, recommending to a user that they should "love running" to ensure success, while perhaps insightful, is not necessarily actionable.

7. CONCLUSIONS

As computing devices continue to become more embedded in our everyday lives, they are mediating an increasing number of our interactions with the world around us. From helping people search for the best product to buy, to recommending a restaurant we are likely to enjoy, computing services enable users to evaluate options and take action with "one click". While such services model many facets of the options they present, they do not model the higher-level implications and trade-offs inherent in deciding to take one action instead of another. For example, a restaurant recommender service will not know that suggesting a carb-heavy Italian restaurant the evening before a person is going to run a marathon might improve their race outcomes. Today, people reason about these trade-offs based on their own past experiences and learnings, combined with their own "gut instinct". People with a relevant knowledge may do well; but many others do not. By aggregating the combined experiences of hundreds of millions of people into a knowledge base of ac-tions and their consequences, we believe that our computing devices may provide significant assistance to augment our own decision-making abilities.

In this paper, we focused on the question of feasibility: Can relatively straightforward techniques identify action-outcome relationships from social media data? As demonstrated in our initial results, even a relatively small scale of social media data — weeks as opposed to the years of data available — allows us to discover rich action-outcome relationships. As future work, we are continuing to develop more sophisticated techniques, as well as evaluate with broader workloads and applications.

8. REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In *Intl Workshop on Cyber-Physical Networking Systems (CPNS)*. IEEE, 2011.

[2] S. Asur and B. A. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2010.

[3] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11, 2011.

[4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[5] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*. IEEE, 2010.

[6] G. M. Chen. Tweet this: A uses and gratifications perspective on how active twitter use gratifies a need to connect with others. *Computers in Human Behavior*, 27(2):755–762, 2011.

[7] H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[8] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*, 2012.

[9] M. De Choudhury, S. Counts, and E. Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *CHI*. ACM, 2013.

[10] M. De Choudhury, M. Gamon, and S. Counts. Happy, nervous or surprised? classification of human affective states in social media. In *ICWSM*, 2012.

[11] B. De Longueville, R. S. Smith, and G. Luraschi. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Intl. Workshop on Location Based Social Networks*. ACM, 2009.

[12] F. Diaz, M. Gamon, J. Hofman, E. Kiciman, and D. Rothschild. Online and social media data as a flawed continuous panel survey. Working Paper http://research.microsoft.com/flawedsurvey.

[13] J. DiMicco, D. R. Millen, W. Geyer, C. Dugan, B. Brownholtz, and M. Muller. Motivations for social networking at work. In *CSCW*. ACM, 2008.

[14] N. B. Ellison et al. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[15] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. What can search predict. In *WWW*, 2010.

[16] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[17] P. M. Gollwitzer and P. Sheeran. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology*, 38:69–119, 2006.

[18] S. Guo, M.-W. Chang, and E. Kiciman. To link or not to link? a study on end-to-end tweet entity linking. In *HLT-NAACL*, 2013.

[19] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Workshop on Web mining and social network analysis*. ACM, 2007.

[20] Y. Jin, E. Kıcıman, K. Wang, and R. Loynd. Entity linking at the tail: sparse signals, unknown entities, and phrase models. In *WSDM*. ACM, 2014.

[21] A. N. Joinson. Looking at, looking up or keeping up with people?: motives and use of facebook. In *CHI*. ACM, 2008.

[22] E. Kıcıman. Omg, i have to tweet that! a study of factors that influence tweet rates. In *ICWSM*, 2012.

[23] N. Kokkalis, T. Köhn, J. Huebner, M. Lee, F. Schulze, and S. R. Klemmer. Taskgenies: Automatically providing action plans helps people complete tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(5):27, 2013.

[24] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[25] Y.-L. Kuo, J. Hsu, and F. Shih. Contextual commonsense knowledge acquisition from social content by crowd-sourcing explanations. In *Proceedings of the Fourth AAAI Workshop on Human Computation*, pages 18–24, 2012.

[26] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*. ACM, 2010.

[27] C. Lampe, N. Ellison, and C. Steinfield. A face (book) in the crowd: Social searching vs. social browsing. In *CSCW*. ACM, 2006.

[28] E. Law and H. Zhang. Towards large-scale collaborative planning: Answering high-level search queries using human computation. In *AAAI*, 2011.

[29] I. Mani and G. Wilson. Robust temporal processing of news. In *ACL*. Association for Computational Linguistics, 2000.

[30] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

[31] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8), 2013.

[32] M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *CSCW*. ACM, 2010.

[33] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 2011.

[34] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[35] A. Prestwich, M. Perugini, and R. Hurling. Can implementation intentions and text messages promote brisk walking? a randomized trial. *Health Psychology*, 29(1):40, 2010.

[36] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[37] M. Richardson. Learning about the world through long-term query logs. *ACM Transactions on the Web (TWEB)*, 2(4):21, 2008.

[38] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

[39] J. M. Robins and L. Wasserman. On the impossibility of inferring causation from association without background knowledge. *Computation, causation, and discovery*, pages 305–321, 1999.

[40] A. Sadilek, H. A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*, 2012.

[41] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[42] T. Spiliotopoulos and I. Oakley. Understanding motivations for facebook use: Usage metrics, network structure, and privacy. In *CHI*. ACM, 2013.

[43] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.

[44] R. W. White, N. P. Tatonetti, N. H. Shah, R. B. Altman, and E. Horvitz. Web-scale pharmacovigilance: listening to signals from the crowd. *Journal of the American Medical Informatics Association*, 2013.

[45] E. Yom-Tov and E. Gabrilovich. 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.