Extracting Diurnal Patterns of Real World Activity from Social Media

Nir Grinberg
Rutgers University
nirg@cs.rutgers.edu

Mor Naaman
Rutgers University
mor@rutgers.edu

Blake Shaw
Foursquare, Inc.
blake@foursquare.com

Gilad Lotan
SocialFlow, Inc.
gilad@socialflow.com

Abstract

In this study, we develop methods to identify verbal expressions in social media streams that refer to real-world activities. Using aggregate daily patterns of Foursquare checkins, our methods extract similar patterns from Twitter, extending the amount of available content while preserving high relevance. We devise and test several methods to extract such content, using time-series and semantic similarity. Evaluating on key activity categories available from Foursquare (coffee, food, shopping and nightlife), we show that our extraction methods are able to capture equivalent patterns in Twitter. By examining rudimentary categories of activity such as nightlife, food or shopping we peek at the fundamental rhythm of human behavior and observe when it is disrupted. We use data compiled during the abnormal conditions in New York City throughout Hurricane Sandy to examine the outcome of our methods.

1 Introduction

The huge volumes of social media activity readily expose and reflect people's attitudes, attention, interests, and activities. In particular, data from the social media services Foursquare and Twitter often reflects activities of users at real-world venues including restaurants, coffee shops, shopping venues and more. For example, Figure 1 shows the weekly average of activity in four broad categories as captured on Foursquare. Modeling diurnal activities from social media provides unique opportunities to understand and draw insights about social activities in urban (and other) settings, to reason about differences between cities and other locations, and to provide alerts and indications when patterns are disrupted. Such models have the potential to transform the study of urban communities, with implications for diverse social challenges such as public health, emergency response, community safety, transportation and resource planning.

Previously, researchers studied temporal patterns of online activity, real-world activities and language use in disconnect. While it is clear that real-life experiences shape our online presence and vice versa, the cross-section between the two worlds received little attention by the research community. Foursquare checkins are an ideal source of information about people's physical location and a strong indicator of their actions. Using large scale dataset of Twitter posts we address the question of identifying verbal expressions that lead to or follow on actions in the real-world, with high volume and semantic coherence.

The challenges in studying diurnal patterns range from finding textual phrases that reflect real-world activities in a noisy, constantly changing environment, to modeling temporal differences and abnormal conditions. First, the lack of proper grammar, use of abbreviations and slang on social media pose a serious challenge to traditional nature language processing (NLP) tools (Wang et al. 2013; Ritter et al. 2011; Ramage, Dumais, and Liebling 2010). The fact that there are many number of ways to refer, in text, to a certain real-world activity only exacerbate the problem. Therefore, robust methods are needed to effectively find footprints of real-world activity in language. Secondly, these methods need to handle the sparsity and variability of content volume as well as the noise inherent to social platforms. Further, once temporal patterns are extracted and represented in time-series form, we need to identify intervals of time where differences are statistically significance with fine-grain time-resolution.

Our goal in this work is to develop and test a set of robust methods to extract diurnal patterns of real-world activities from social media, and reason about differences between patterns (e.g. in different times or locations). To this end, we explore several approaches based on statistical measures of correlation and mutual information and evaluate them using Twitter and Foursquare data. With diurnal patterns at hand, we model the differences across platforms with fine-grain time resolution. We study deviations from normal behavior during Hurricane Sandy as means for validating our methods and demonstrating their effectiveness.

Our contributions are therefore:

1. Methods for extracting information from Twitter that corresponds with real-world activities such as eating, shopping, and others (sections 4, 5).
2. A method of modeling differences of activity time-series that can help reason about behavioral patterns and intended use of social media platforms (Section 6.1).
3. A study of how these techniques can help expose irregular activities in social media data, in a case study of Hurricane

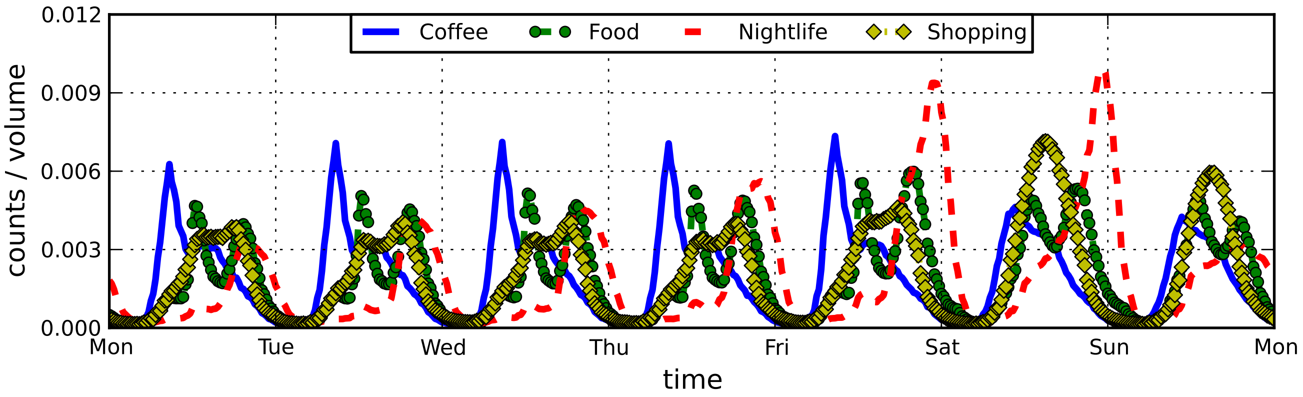

Figure 1: Weekly patterns of four broad categories of real-world activities as reflected in Foursquare checkins.

Sandy (Section 6.2).

We begin with a description of background research and survey related work.

2 Background and Related Work

The widespread proliferation of mobile devices and social networking services offers new opportunities for studying human behavior in urban settings. Cuff et al. defined Urban Sensing as the analysis of data ”generated through decentralized collection, shared freely and amenable to distributed sense-making not only for the pursuit of science but also advocacy, art, play, and politics” (Cuff, Hansen, and Kang 2008). Previous work examined, for example, cellphone use to understand spatio-temporal human dynamics (Candia et al. 2008), geo-coded image posting rate to learn about points of interest (Girardin et al. 2008; Pereira et al. 2011) and activity on Foursquare to model city dynamic structure and functional use (Cranshaw et al. 2012; Kling and Pozdnoukhov 2012).

In this work we use Foursquare checkins as proxy for actions in the real-world, and find Twitter posts that reference such actions. The Foursquare mobile app allows users, who are physically present in a place, to share that information with their friends. Twitter micro-blogging service offers its users the ability to share short messages (limited to 140 characters) with their followers in almost real-time. Twitter (500 million active users) and Foursquare (30 million registered users) are both very popular and have fine-grained time resolution, which makes them ideal for studying diurnal patterns of human behavior at scale. Considering the streams of Foursquare and Twitter together allows one to utilize Foursquare’s emphasis on physical location with the high volume of activity on Twitter. While the two social networks may not be representative of the general population, combining platforms can potentially reduce such bias.

Previous work examined diurnal, seasonal and other other temporal patterns of social media activity. Golder et al. (2007) found consistent weekly and seasonal patterns of social interaction among college students on Facebook. Later, Golder and Macy (2011) drew a connection between sentiment on Twitter posts to cycles of sleep and seasonality. Naaman et al. (2012) studied the variations of keyword use on Twitter diurnal patterns and assessed their robustness across geographical locations. Leskovec et al. developed a framework for tracking variants of short textual phrases over time (Leskovec, Backstrom, and Kleinberg 2009) and found prototypical temporal patterns in the spread of news stories (Yang and Leskovec 2011).

In the domain of online search, several studies considered the temporal aspect of search engine queries. Chein and Immorlica (2005) showed semantic similarity between search queries with no lexical overlap (e.g. Halloween and pumpkin). Shimshoni et al. (2009) and Choi and Varian (2012) looked at trends and seasonality in Google searches across categories and estimated their predictability. Shokouhi and Radinsky (2012) improved query auto-completion by considering temporal patterns of search engine queries. Our work differs from traditional search because it is focused on matching specific temporal patterns across platforms, maximizing the total number of obtained results and requiring high quality throughout the results set (not just the top N).

Numerous algorithms applied probabilistic approaches to semantic expansion and clustering, but none considered recurring language use following a specific temporal pattern. Along the lines of semantic similarity, PMI-IR (Turney 2001) used PMI scores based on search engine results to assess similarity of two words. SOC-PMI (Islam and Inkpen 2006) improved semantic similarity by taking into account co-occurrence in the context of words. Pantel and Pennacchiotti (2006) discovered new forms of semantic relations by starting with a small seed list and expanding it.

Several Probabilistic Topic Models were proposed for thematic analysis of Twitter messages (Zhao et al. 2011; Hu et al. 2012; Ramage et al. 2009). However, all models relied on grouping messages per user, specific event time or hashtag. In addition, the results of Topic Models are not always interpretable to humans. Mimno et al. (2011) mitigated the issue of interpretability by introducing coherence measure similar to PMI in regulating topics. Our methods build on the good semantic properties of PMI, extend it and optimize the similarity of resulting patterns to real-world diurnal patterns.

3 Extracting Real World Activities

We are interested in identifying Twitter posts that correspond to physical world activities, as reflected in Foursquare checkins. In other words, we are looking for Twitter status updates ("tweets") that reference actions people take in the real-world, like eating, meeting friends, shopping, sleeping and more. By identifying real-world activities on Twitter we gain access to greater volume of activity and improve the robustness of patterns, even in geographic locations where Foursquare is not yet fully adopted. However, tweets can reflect a wide array of topics, activities, thoughts and responses to anything at all – in the physical world or not. Furthermore, even if we limit our vocabulary to the order of 10^4 and look for keywords indicative of real-world activity, finding the optimal set is a search over $2^{(10^4)}$ options. Therefore, we resort to approximation methods and evaluate them in Section 5.

Ideally, we want our extracted patterns to have 3 properties:

1. High specificity: we want our extracted terms (and by extension our extracted tweets) to be as specific as possible to the category at hand.
2. Similarity to Foursquare diurnal patterns, without overfitting: references to real-world activities should resemble the patterns of taking these actions in the world. However, we do expect to see differences or lag between platforms. For example, talking about coffee on Twitter can occur at any time of the day while checking-in on Foursquare is much more likely during business hours of coffee shops.
3. High volume: high volume of temporally consistent patterns such that even in smaller cities the patterns are stable.

We devise a set of methods to identify keywords and the resulting tweet patterns according to these criteria, presented in Section 4. In Section 5, we validate the methods according to the criteria mentioned above. First, we define few concepts that will be used throughout the paper.

3.1 Preliminaries

We formalize the definition of content items, users, venues, geographic locations and aggregated measures on Twitter and Foursquare. **Users** and **Venues** are marked $u \in U$ and $v \in V$, which can be minimally modeled by a user ID and venue ID, respectively. Similarly, we denote **Geographic locations** by $g \in G$ and consider only non-overlapping geographic areas. **Content items** are marked $m \in M$ for tweets and $n \in N$ for check-ins. m can be minimally modeled by a tuple (u_m, c_m, t_m, g_m) containing the identity of the user posting the message $u_m \in U$, the content of the message c_m, the posting time t_m and the location of the user at the time of posting g_m. n can be minimally modeled by a tuple (u_n, v_n, t_n) containing the identity of the user checking-in $u_n \in U$, the check-in venue v_n and the posting time t_n. The check-in location g_n can be inferred from the venue location. For simplicity, we associate each check-in with the venue primary category, $cat_n \in \{coffee, food, nightlife, shopping\}$.

We now formalize other aggregate concepts, i.e., the content's diurnal patterns. We use the following definitions:

- $X_K^t(p, g)$ (in short, X_K^t) is the mean number of tweets in geographical location g during the time period p that contain any of the terms in the non-empty set of keywords K.
- $X_{cat}^{4sq}(p, g)$ (in short, X_{cat}^{4sq}) is the mean number of Foursquare (4sq) check-ins in geographical location g during time period p with primary category cat.

In this work we use 20 minutes-long bins, therefore $p \in P_{day} = \{1..72 = 3 \times 24\}$ for diurnal time-series or $p \in P_{week} = \{1..504 = 3 \times 24 \times 7\}$ for weekly time-series. In addition, we usually refer to aggregated time-series over all geographical locations, therefore omitting g. Time-series from multiple locations were first adjusted to local time and then aggregated.

We borrow two concepts from Information Theory: Mutual Information and Pointwise Mutual Information. Let $W_K \in \{0, 1\}^M = \Omega$ be a binary, random variable of dimension M with non-zero entries for tweets containing any of the keywords in the non-empty set K. We will use the shortened notation w_1 and w_2 to designate realizations of $W_{\{k_1\}}$ and $W_{\{k_2\}}$ with keyword k_1 and k_2, respectively. Mutual Information (MI) is then defined as:

$$MI(W_1, W_2) = \sum_{w_1, w_2 \in \Omega} P(w_1, w_2) \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{1}$$

In practice, it is often impractical to compute MI for high-dimensional vectors and point estimates are used instead. Pointwise Mutual Information (PMI) is defined as follows:

$$PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{2}$$

In Natural Language Processing (NLP) it is common to estimate these distributions using Maximum Likelihood Estimates (MLE). Using the MLE the joint and marginal distributions in Equation 2 are replaced with the observed frequency of co-occurrence and occurrences of terms, respectively. One benefit of using PMI instead of a simple co-occurrence model is that co-occurring terms are discounted by their individual frequency. For example, simple co-occurrence model would rank high the word "get" given the word "coffee" since the bigram "get coffee" is very frequent. However, PMI would relatively decrease the score of "get" since it is very frequent on its own and occurs with many terms other than "coffee".

4 Methods

This section describe different approaches for selecting keywords indicative of real-world activities. We start by describing a method solely based on similarity to Foursquare diurnal patterns. Then, we turn to develop keyword expansion methods that take a small set of keywords and possibly the Foursquare patterns and find additional semantically relevant terms.

4.1 Temporal Correlation

Given the Foursquare data X_{cat}^{4sq} (i.e. the target time-series), a simple approach for finding related terms in Twitter is to score individual terms by their similarity to the target time-series. Mathematically, Temporal Correlation (TempCorr) uses the measure of correlation between $X_{\{k\}}^{t}$ and the target time-series:

$$TempCorr(k, cat) = Corr(X_{\{k\}}^{t}, X_{cat}^{4sq}) \tag{3}$$

The choice of computing weekly correlation was made to reduce the noise of one-time abnormal occurrences, as well as to preserve within-week variations (e.g. weekend to weekday differences). Therefore, in TempCorr terms are ranked based on the level of correlation to the target time-series. Of course, high temporal correlation does not guarantee semantic relevance. For instance, a word like ”morning” may score high in the category of coffee merely based on its occurrence at similar times as coffee terms.

4.2 Aggregated Pointwise Mutual Information

Aggregated Pointwise Mutual Information (APMI) captures the context shared between two or more terms, controlling for the popularity of individual terms. Given κ, a set of keywords describing the category cat, we use APMI to retrieve semantically relevant keywords as follows:

$$APMI(w|\kappa) = \sum_{w' \in \kappa} log \frac{P(w, w')}{P(w)P(w')} \tag{4}$$

The range of this score is $(-\infty, \infty)$ where higher values correspond to more relevant terms to the initial seed list. For seed list of size 1, APMI simply reduces to PMI, but adding terms to κ surfaces keywords that co-occur frequently with the seed list. For instance, given keywords relevant to coffee activity (e.g. ”coffee”, ”starbucks” and few others) APMI was able to find keywords such as ”#frappuccino”, ”iced” and ”#barista”.

4.3 Context Pointwise Mutual Information

The Context Pointwise Mutual Information (ContextPMI) method is focused at finding terms with similar context, discounting generally popular terms. We define *Context* as the set of top n terms according to PMI. Intuitively, ContextPMI score is the number of shared terms between a word and a reference context, i.e. higher scores corresponds to words that have more terms in common. Mathematically, ContextPMI takes a seed list of keywords κ and the number of top context terms to consider n and expands the list:

$$ContextPMI(w|\kappa, n) = |C(w, n) \cap C_{ref}(\kappa, n)| \tag{5}$$

Where $C(w, n)$ is the context of the term w obtained using the top n terms by PMI, $C_{ref}(\kappa, n) = \bigcup_{w' \in \kappa} C(w', n)$ and κ is the initial seed list as before. The score is an integer in the range $[0, min(|C(w, n)|, |C_{ref}(\kappa, n)|)]$. For example, given the same seed list for the category of ”coffee”, ContextPMI generated ”#cappuccino”, ”macchiatos”, ”#sbux” (shortened for Starbucks).

ContextPMI may be sensitive to small subsets of seed list terms that have very similar context, but share very little with the concept represented by the rest of the seed list. To mitigate that, we introduced Bootstrap Aggregation (Bagging) into ContextPMI. Breiman (1996) showed that taking m samples of the original dataset of size n with replacement, yields a dataset where each of the m samples it expected to have $1 - 1/e \approx 63.2\%$ original items and the rest are duplicates. We sampled terms from the original seed list 100 times with replacement and modified the reference context in Equation 5 to be a multiset. We also set a threshold for minimal ContextPMI score at 500 - each term has to share at least an average of 5 terms with the seed list context (sampled 100 times).

4.4 Hybrid Approach

In order to extract robust patterns that resemble real-world activities and have good semantic properties we devise a hybrid approach. Our hybrid method takes Foursquare time-series X_{cat}^{4sq} and an initial seed list κ and assigns score to term w as follows:

$$Hybrid(w|\kappa, n, X_{cat}^{4sq}) = \begin{cases} ContextPMI(w|\kappa, n) & \text{if (7) holds} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$Corr(X_{\kappa \cup \{w\}}^{t}, X_{cat}^{4sq}) > Corr(X_{\kappa}^{t}, X_{cat}^{4sq}) \tag{7}$$

That is, the Hybrid score is simply the ContextPMI score for terms that bring the correlation of the combined Twitter time-series closer to the target time-series X_{cat}^{4sq} and 0 otherwise.

In the next section we evaluate the performance of our four extraction methods in terms of quality of extracted keywords, the similarity of the induced diurnal patterns to the Foursquare patterns, and the robustness of the patterns.

5 Results

In this section we describe the datasets used in our experiments, report on their results and evaluate the extracted patterns according to the criteria defined in Section 3. We quantify the specificity of extracted patterns (property #1) by manually labeling the top 200 terms of each method. We use correlation between Twitter and Foursquare time-series to evaluate the similarity of the extracted patterns (property #2). Finally, we inspect the volume of extracted patterns, examine its quality and assess its robustness (property #3).

5.1 Experimental Setup

Our Foursquare dataset consisted of all checkins from 2011 and 2012 (except December 2012) aggregated in 20 minutes bins by category and urban area. Foursquare categories included the broad categories of Coffee Shops, Restaurants, Bars and Shopping venues as well as the their finer subcategories. From Twitter we had only about 10 weeks of data (August 27th, 2012 to November 5th, 2012) aggregated into time bins in similar fashion. Based on the total volume available from individual urban areas, we restricted our analysis to seven of the biggest urban areas in the United States, including the New York metropolitan area, the San Francisco bay area and others.

To properly consider diurnal patterns in local time, we had to infer the user location with high accuracy. We used the method described in (Naaman et al. 2012), which uses the user hometown location to identify tweet-level location. We evaluated this method independently by verifying that the tweet location falls inside the metro area we associated with the user. Since the user is expected to move about without changing their profile location, we expect a solid but not a perfect match between the two location fields. Indeed, we found the match had precision and recall of around 75%.

We used more than 300 million tweets from a single day of the Twitter Firehose to calculate the PMI score for any pair of words, hashtags and Twitter handles. Since our approach is purely statistical, we included tweets in all languages. For practical reasons, we retained for each term its top 100 terms according to PMI. Throughout the evaluation, we excluded terms that appeared in less than 200 tweets. We identified additional domain-specific stopwords, such as ”url”, by taking the top terms according to APMI of a random sample of 10,000 keywords. For methods that require initial seed list we manually created and used the same initial seed for all method. The seed list per category consisted of 10 keywords and their hashtag form. Our manually created seed lists can be seen in Appendix A and the top terms extracted by each method are provided in Appendix B.

5.2 Content Relevance and Volume

Figure 2 shows the quality of extracted keywords for each method and category of activity. Based on the top 200 manually labeled keywords, the four sub-plots compare the percent of relevant keywords in the top 10, 20, ..., 200 terms. For example, in the category of coffee (shown on the top right), out of the top 50 terms identified by the Hybrid method only about 85% were labeled as relevant for the category. Notice that ContextPMI and the Hybrid method may generate fewer than 200 terms because of the score threshold. Significant differences are evident both between categories and among extraction methods. ContextPMI and the Hybrid method generally achieve better accuracy and their deterioration in quality is slower compared with APMI and TempCorr. As expected, the worst method in terms of semantic relevance is the TempCorr method, which ignores semantics altogether.

Figure 3 is similar to Figure 2, but compare the percent of relevant tweets with the volume of newly discovered content. Let’s denote Σ_K as the number tweets that contain at least one of the terms in keyword set K. $K(n)$ is the set of top n keywords as ranked by either ContextPMI or the Hybrid method and $K_{true}(n)$ contains only the keywords manually labeled as relevant. κ is the set of seed list keywords as before. On the Y-axis we present $\Sigma_{K_{true}}/\Sigma_K$, the precision in identified tweets. The X-axis is $\Sigma_{K\cup\kappa}/\Sigma_K$, the ratio of total available content to the volume induced by the seed list. Each point in the figure corresponds to 10 addition keywords, that is n’s in 10, 20, ..., 200. We excluded APMI and TempCorr from the figure because they generated huge numbers of volume ratio with very low precision, which masked the more stable performance of ContextPMI and the Hybrid method.

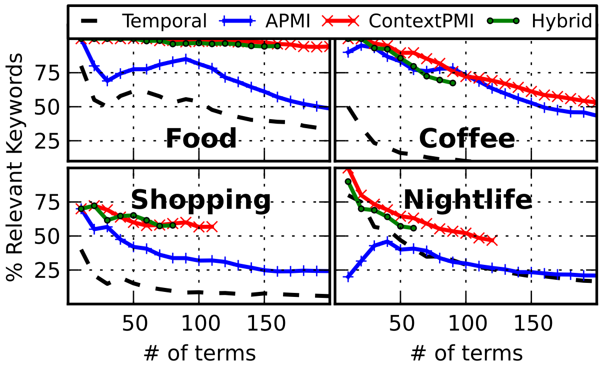

Figure 2: Percent of Relevant Keywords by each extraction method and category of activity. See Appendix B for the top terms extracted by each method.

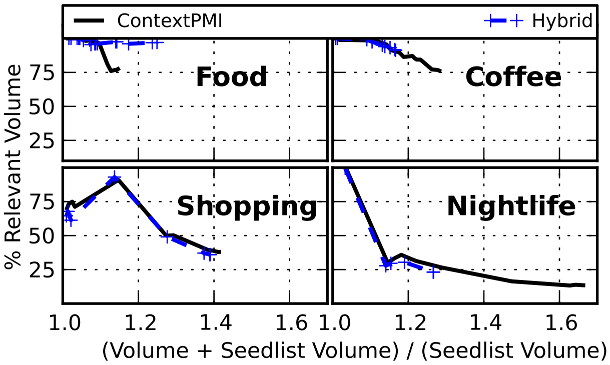

Figure 3: Quality vs. quantity in identified tweets. Each point corresponds to the inclusion of 10 additional keywords.

Figure 3 demonstrate the impact keywords have on the induced volume and its accuracy. For example, in the category of Shopping, the first terms obtained by both extraction methods provide 1.15 (+15%) more tweets than the seed list where 90% of those tweets are relevant to the Shopping category. In categories like Nightlife and Shopping, selection of irrelevant terms with high-frequency caused considerable decrease in volume precision. In ”Food” and ”Coffee” precision remains more than 75% while attaining more than 25% new content. The Hybrid method is on par with ContextPMI in most categories, except for the category of ”Food” where it retained more than 90% relevance. Also, the hybrid method selects fewer terms and stops before the quality deteriorates any further. These findings suggest that the criteria in the Hybrid method (Equation 7) improves both temporal similarity and semantic relevance.

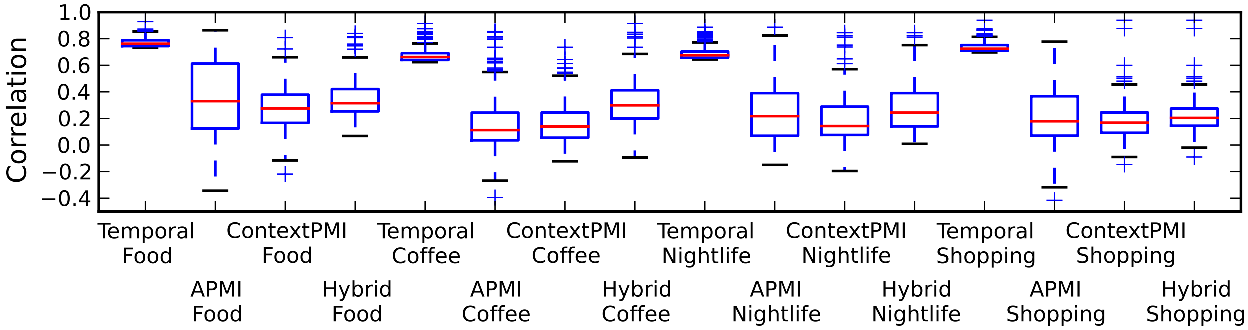

Figure 4: Comparing four extraction methods by correlation of individual terms to real-world patterns over four categories of activity.

5.3 Temporal Similarity

In order to investigate the temporal aspect of our extraction methods we first construct their weekly, binned and time-zone adjusted time-series $X^t_{\{k\}}$. The Box plots in Figure 4 summarize the distribution of correlation values of individual terms with Foursquare weekly patterns. Correlation was calculated as $TempCorr(X^t_{\{k\}}, X^{4sq}_{cat})$ for every term k in the top 300 results of each extraction method and category. For example, "Food" terms extracted by APMI have median correlation value of 0.35, first quartile at 0.1, third quartile at 0.6 and no outliers below -0.35 or above 0.85. Of course, TempCorr attains much better correlation scores since it is directly designed to optimize this aspect. With the exception of "Coffee", APMI scores have wider variation than the rest of the methods. These findings suggest that while APMI selects terms that appear together they may not correspond with real-world patterns. ContextPMI and the Hybrid method are comparable in terms correlation, with slightly higher median for the Hybrid method.

Surprisingly, the three expansion methods that have relatively weak temporal correlation for individual terms, have very different temporal correlation in their aggregated form as shown in Figure 5. The figure characterizes the change in temporal correlation with Foursquare weekly patterns as a we incorporate more keywords into the Twitter patterns. For example, in the category of "Coffee" (shown on the top right) the Hybrid method obtained correlation value of 0.82 after including its top 50 terms. Out of the three keyword-based expansion methods, the Hybrid method achieves the best similarity to Foursquare patterns. In the category of "Food" (shown on the top left) the Hybrid method even performed better than the greedy approach of TempCorr.

5.4 Robustness

We now turn to assess the robustness of our extraction methods as a function of available volume in geographic locations. We harness the widely used Signal-to-noise ratio (S/N) for the task, averaging over all time bins:

$$S/N(K, g) = \frac{1}{|P_{week}|} \sum_p \frac{X^t_K(p, g)}{\sigma^t_K(p, g)} \tag{8}$$

where K is a set of keywords, g geographic location, $X^t_K(p, g)$ the mean number of tweets at time bin p and $\sigma^t_K(p, g)$ is its standard deviation. Higher S/N ratio in our case mean patterns with more volume and less variation over time.

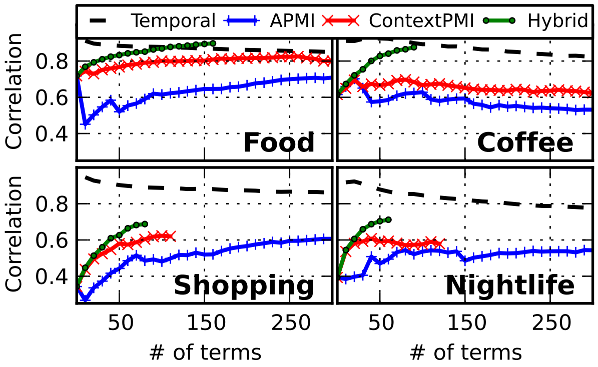

Figure 5: Comparing aggregated patterns correlation with real-world patterns.

Figure 6 shows how robustness depend on amount of available content. The Y-axis depicts the S/N as in Equation 8, computed using the top 200 terms for each extraction method and averaged across categories. The X-axis represents the total number of tweets available from different geographical locations (in log scale). For example, we had slightly more than 16 million tweets from Boston, MA and using only the seed list terms (black dotted line) the four categories averaged S/N is 2.65. The figure demonstrates that methods based on multiple keyword selection are relatively robust - the smallest city is more than 4 orders of magnitude smaller in volume than the biggest city, but its S/N only decreases by 0.8, keeping within the same order of magnitude. TempCorr and the APMI method represent two optimums in S/N space: one for the most similar terms to Foursquare activity and the other for seed list, by choosing the most co-occurring terms with it. Therefore, it is not surprising that in smaller cities TempCorr and APMI are more robust. With more data however, as in the case of New York City, the gap vanishes. Nevertheless, The advantage TempCorr and APMI have in terms of robustness does come at the expense of their specificity to real-world actions as we have seen in Figure 2.

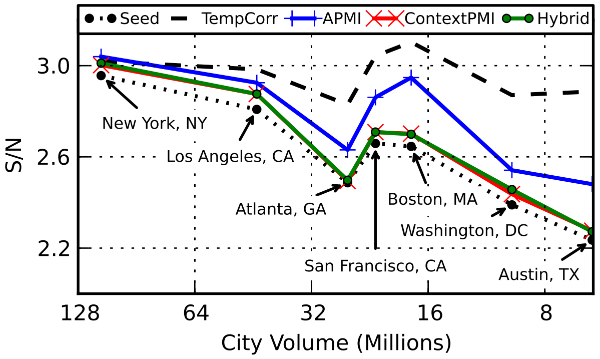

Figure 6: Robustness of extraction methods: Signal-to-noise ratio (S/N) as function of volume (in log scale).

To summarize, our results show that the Hybrid approach is comparable to ContextPMI in terms of quality, volume and robustness, but attains better temporal similarity to real-world patterns.

6 Modeling Time-Series Differences

In this section we use the previously described methods to extract categorical patterns from Twitter and compare them with real-world activity as observed on Foursquare. We model differences between two diurnal time-series using Gaussian Processes to highlight times when the differences are of statistical significance, without assuming particular causal structure. Finally, we examine deviations from normality during the recent Hurricane Sandy as a case study for understanding abnormalities during extreme conditions and verify that our methods capture real-world events.

6.1 Diurnal Differences Using Gaussian Processes

Gaussian Processes (GP) are Bayesian time-series modeling technique that assume a Gaussian likelihood on the space of functions relating one time point to the other. The prior for GP is usually specified implicitly as a Kernel function, allowing for infinite basis functions to be induced. As a Bayesian model, GP provide estimates and confidence intervals around them. For our purposes, the power of the GP model lays in its ability to infer from data the number of points contributing to prediction, as well as to provide arbitrary fine time-resolution in interpolation between points.

We use the same datasets described in Section 5.1, but aggregate activity on a period of a day. We construct the time-series of differences by subtracting Twitter daily activity from its Foursquare equivalent. The resulting time-series is modeled using GP with the widely used Squared-Exponential kernel and a noise term:

$$K(t_i, t_j) = \sigma_{SE}^2 \exp\left[-\left(\frac{t_i - t_j}{2l}\right)^2\right] + \sigma_{noise}^2 \delta_{t_i,t_j} \quad (9)$$

Where t_i and t_j are two time-stamps, σ_{SE}^2 and σ_{noise}^2 are variance terms for the squared-exponential and noise component, respectively. The parameter l is considered the typical length-scale as it determines the effective time scale of data points contributing for prediction. Lastly, δ_{t_i,t_j} is the Kronecker delta which is equal to 1 if $t_i = t_j$ and 0 otherwise.

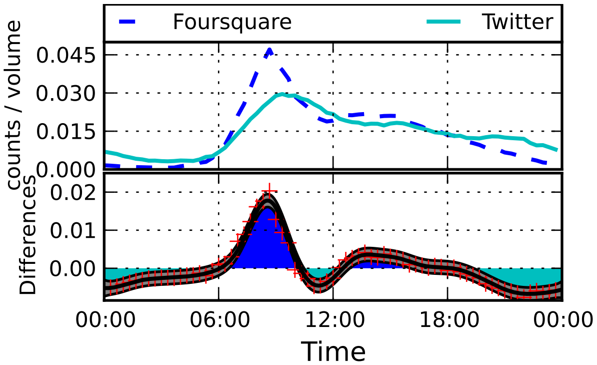

Figure 7: Coffee diurnal patterns (top) and differences (bottom) of weekday activity.

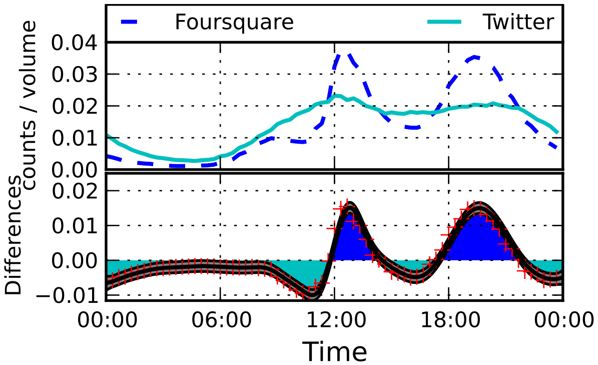

Figure 8: Food diurnal patterns (top) and differences (bottom) of weekday activity.

Figures 7-10 show for each category the diurnal patterns of activity on Twitter and Foursquare and their differences. Twitter patterns were extracted independently of the Foursquare signal by using the top 200 terms according to ContextPMI (using the Hybrid method would introduce a bias). We plot weekday patterns for ”Coffee” and ”Food” (Figures 7 and 8) and weekend patterns for ”Shopping” and ”Nightlife” (Figures 9 and 10). The top part of these figures depicts the diurnal patterns X_K^t and X_{cat}^{4sq}, normalized by their total volume. The bottom part of Figures 7-10, contains both the time-series of differences and the prediction of the GP model. Red crosses designate Twitter activity subtracted from Foursquare activity, solid line correspond the GP prediction with 95% confidence interval around it. We shaded areas above and below the zero axis when the differences time-series did not include the zero value within its 95% confidence interval. Shaded area above (below) the zero axis correspond to times where Foursquare activity is significantly greater (smaller) than the activity on Twitter. For example, in the category of Food (Figure 8) there are two peaks in activity: one at noon and the other at 7pm. At both peaks of ”Food” there is more activity on Foursquare than Twitter (positive difference), which is statistically significantly greater than zero (shaded).

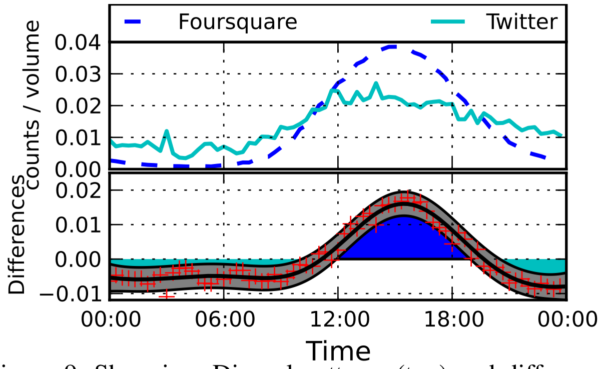

Figure 9: Shopping: Diurnal patterns (top) and differences (bottom) of weekend activity.

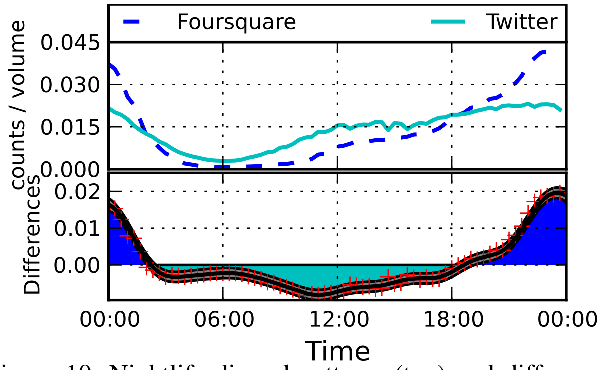

Figure 10: Nightlife diurnal patterns (top) and differences (bottom) of weekend activity.

The diurnal patterns in figures 7-10 overlap considerably. The similarity of Foursquare and Twitter across all categories suggest that ContextPMI is indeed capturing real-world activities on Twitter. Moreover, we observe from these figures that the differences between verbal and locational expressions are very much category specific. At the level of correlation (not causation!) we can say that people in general first have their coffee (mostly at 8:30am) and only afterwards talk about coffee on Twitter. Similarly, shopping takes place before it is discussed online. The opposite phenomena appears in Nightlife where people discuss activities earlier than checking-in to nightlife venues on Foursquare. Food activity on Twitter usually precedes or follows the peak of activity on Foursquare during meal times.

6.2 Case Study: Hurricane Sandy

When rare events at the scale of Hurricane Sandy happen, we expect them to leave an unquestionable mark on Social Media activity. Using the models and methods described so far, we would like to measure the effect of such vast and prominent event on Twitter and Foursquare, compare and contrast the deviations on both platforms.

To analyze deviations from ”normality”, one must make predictions on unfolded time-series. While we leave for future work the task of modeling unfolded time-series using Gaussian Processes, we take a simpler approach in this section. We assume that data-points in each time bin are normally distributed and standardize them to obtain z-scores. We assume that the mean and variance are those of the weekly patterns observed over the 10 weeks in our dataset. Z-score of value -3 or 3 correspond to three standard deviation from the mean; about 99% of values fall in that range.

For the analysis of deviations we included additional Foursquare category of grocery shopping, that is checkins at supermarkets, grocery stores and others. We applied the Hybrid approach described in Section 4.4 and extracted the equivalent Twitter patterns for grocery shopping.

Figure 11 shows z-scores for three of the categorical irregularities occurring in New York Metropolitan area during hurricane Sandy. The z-scores outline deviations from normal weekly behavior over three periods: before the storm, in preparation for it and during the hurricane. Prior to the storm, activity is relatively normal with the exception of iMac release on 10/25. The big spikes in divergent activity in the two days right before the storm correspond with emergency preparations and the spike in nightlife activity follows the ”celebrations” pattern afterwards [1]. In the category of Grocery shopping (top panel) the deviations on Foursqaure and Twitter overlap closely, while on Nightlife the Twitter activity lags after Foursquare. On October 29 and 30 shops were mostly closed in NYC and we observe fewer checkins than usual, but interestingly more tweets about shopping. This finding suggests that opposing patterns of deviations may indicate of severe distress or abnormality, with the two platforms corroborating an alert.

7 Conclusions and Future Work

We developed methods for extracting references to real-world activities from text-based social media (namely, Twitter posts) and evaluated the volume, quality and robustness of the resultant data. The extracted patterns allow us to observe categorical human behavior at unprecedented scale and reason about differences and similarities between platforms. Our methods can potentially generalize to different domains and topics.

Our analysis showed that diurnal patterns in Foursquare (largely, a source of physical-world actions) closely resemble the patterns on Twitter (textual references to actions). The differences between platforms are category specific, with Twitter activity generally more spread around the peak of Foursquare activity, precede or follow it. The fundamental reasons for these differences are have yet to be explored: are the discrepancies due to the difference in nature of activities, different intent of platform use or the limitation of our methods.

Beyond the regular pattern, the deviations in the case study of Hurricane Sandy clearly separate normal and abnormal times. In some cases the deviations on both platforms closely overlap, while in others some time lag (or even opposite trend) is evident. Moreover, during the height of the storm Foursquare activity diminishes significantly, while Twitter activity is on the rise. These findings have immediate implications for event detection systems, both in com-

[1] the spike in shopping on the night of the 31th is due to Halloween pop-up stores

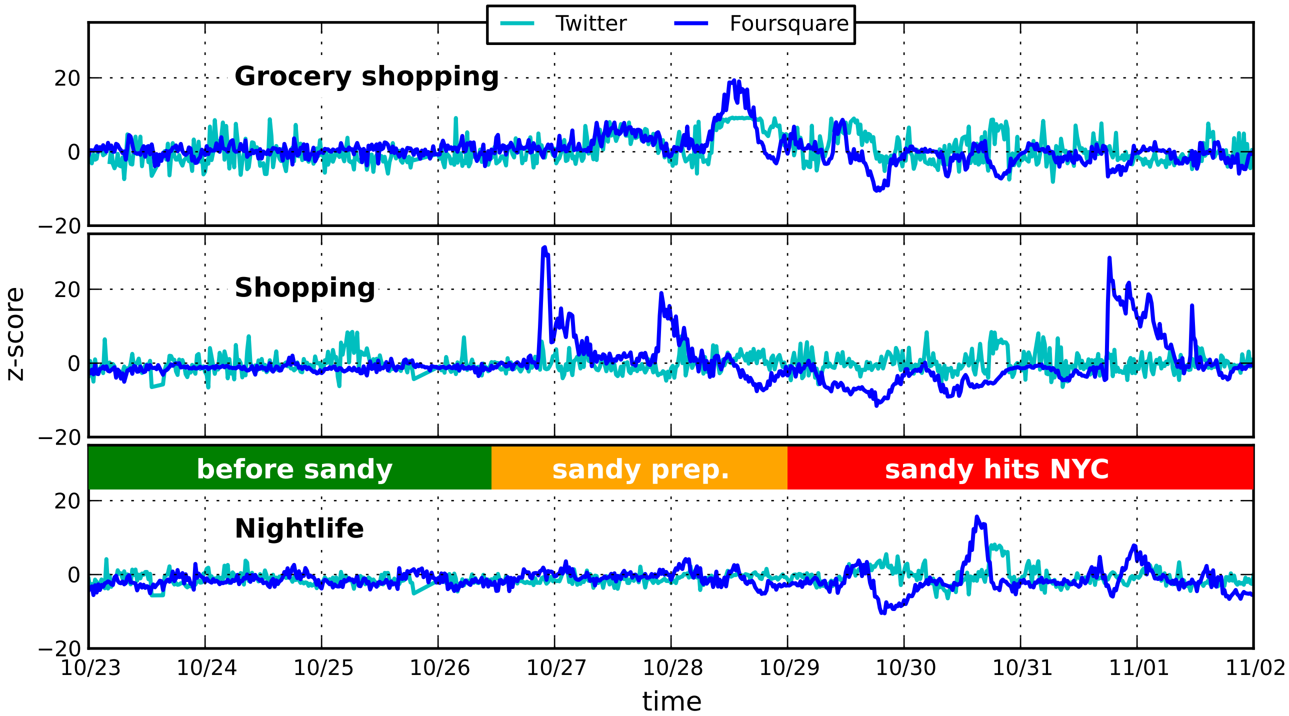

Figure 11: Pattern disruption on Fourquare and Twitter during Hurricane Sandy.

bining multiple sources of information and in using them to improving overall accuracy.

Our methods are based on token-level analysis of the Twitter posts text. Future work could model the language in a post more carefully, for instance by generalizing our methods to use Bigrams, Trigrams, etc. and employ NLP tools like part of speech tagging to perform deeper analysis of language constructs. For example, examining the tense used in tweets before and after lunch time could shed light on user's intent. In respect to time-series analysis, future work could examine the ”unfolded”, non-periodic signals and use multiple indicators for better prediction. Our approach focused on diurnal pattern and thus considered only predefined periods of in-phase signals. To that end, a natural extension could further explore the frequency space, allowing for some phase shift between time-series and drift over time.

References

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–140.

Candia, J.; González, M. C.; Wang, P.; Schoenharl, T.; Madey, G.; and Barabási, A.-L. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41(22):224015.

Chien, S., and Immorlica, N. 2005. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th international conference on World Wide Web (WWW'05)*, 2–11.

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

Cranshaw, J.; Schwartz, R.; Hong, J.; and Sadeh, N. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*.

Cuff, D.; Hansen, M.; and Kang, J. 2008. Urban sensing: out of the woods. *Communications of the ACM* 51(3):24–33.

Girardin, F.; Calabrese, F.; Fiore, F. D.; Ratti, C.; and Blat, J. 2008. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing* 7(4):36–43.

Golder, S., and Macy, M. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–81.

Golder, S. A.; Wilkinson, D. M.; and Huberman, B. A. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Communities and Technologies 2007*. Springer. 41–66.

Hu, Y.; John, A.; Wang, F.; and Kambhampati, S. 2012. Et-lda: Joint topic modeling for aligning events and their twitter feedback. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*.

Islam, A., and Inkpen, D. 2006. Second order co-occurrence pmi for determining the semantic similarity of words. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'06)*, 1033–1038.

Kling, F., and Pozdnoukhov, A. 2012. When a city tells a story: urban topic analysis. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL'12)*, 482–485.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Memetracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, 497–506.

Mimno, D.; Wallach, H.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, 262–272.

Naaman, M.; Zhang, A. X.; Brody, S.; and Lotan, G. 2012. On the study of diurnal urban routines on twitter. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*.

Pantel, P., and Pennacchiotti, M. 2006. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING/ACL'06)*, 113–120.

Pereira, F.; Vaccari, A.; Giardin, F.; Chiu, C.; and Ratti, C. 2011. 19 crowdsensing in the web: Analyzing the citizen experience in the urban space. *From Social Butterfly to Engaged Citizen: Urban Informatics, Social Media, Ubiquitous Computing, and Mobile Technology to Support Citizen Engagement* 353.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, 248–256.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, volume 5, 130–137.

Ritter, A.; Clark, S.; Mausam; and Etzioni, O. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, 1524–1534.

Shimshoni, Y.; Efron, N.; and Matias, Y. 2009. On the Predictability of Search Trends. Technical report, Google. Available at `http://research.google.com/archive/google_trends_predictability.pdf`.

Shokouhi, M., and Radinsky, K. 2012. Time-sensitive query auto-completion. In *Proceedings of the 35th ACM International Conference on Research and Development in Information Retrieval (SIGIR'12)*, 601–610.

Turney, P. D. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01)*, 491–502.

Wang, C.-K.; Hsu, B.-J. P.; Chang, M.-W.; and Kiciman, E. 2013. Simple and knowledge-intensive generative model for named entity recognition. Technical report, Microsoft Research.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*, 177–186.

Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; and Li, X. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*. Springer. 338–349.

A Keyword seed lists

Coffee	coffee, #coffee, starbucks, #starbucks, espresso, #espresso, lovecoffee, #lovecoffee, caffeineaddict, #caffeineaddict, venti, #venti, mugs, #mugs, latte, #latte, caf, #caf, coffeebean, #coffeebean.
Shopping	shopping, mall, outlet, decorating, bookstore, drugstore, florist, #shopsmall, #mall, shopsmall, #decorating, furnishing, #fleamarket, #artsandcrafts, #outlet, #bookstore, fleamarket, antiqueshop, artsncrafts, bridalshop, clothingstore, toystore, thriftstore.
Food	food, eat, dinner, lunch, breakfast, brunch, snack, #yummy, #hungry, supper, #dinner, #breakfast, #lunch, #starving, #brunch, #instafood, #eat, #snack, #foodstagram, #supper, ieat.
Nightlife	drunk, #drunk, drunkkk, #drunkkk, drinks, #drinks, drinkin, pub, #pub, bar, #bar, bartender, instabeer, #instabeer, partying, nightclub, #nightclub, dancebar, #dancebar, #club, #dance.

B Top 5 extracted terms

	TempCorr	**APMI**	**ContextPMI**	**Hybrid**
Coffee	coffee	#frappuccino	#cappuccino	#cappuccino
	today	#mocha	#mocha	#mocha
	dunkin	#latte	#barista	#barista
	latte	starbucks	#starbucks	#starbucks
	today's	iced	#frappuccino	#caramel
Shopping	mall	#sbsperk	#icing	#icing
	shopping	sbsperk	#thrifting	#thrifting
	nap	#crafts	#livingroom	#livingroom
	panera	fabrics	#cakes	#cakes
	#yelp	#shopper	#crafting	#crafting
Food	#yelp	#foodoftheday	#foodgasm	#foodgasm
	sushi	healthy	#foodie	#foodie
	grill	#foodstamping	#eggs	#pasta
	restaurant	#foodpic	#pasta	#delicious
	casa	#foodism	#delicious	#sausage
Nightlife	hookah	60a	#shocktop	#shocktop
	faded	shits	#beerporn	#beerporn
	drunk	nights	#yuengling	#yuengling
	henny	gettin	#beer	#beer
	sleepover	grill	#lushlife	#lushlife