Learning Topical Social Sensors

Authors
Affiliations

Abstract

Social media sources such as Twitter represent a massively distributed social sensor over a kaleidoscope of topics ranging from social and political events to entertainment and sports news. However, due to the overwhelming volume of content, it can be difficult to identify novel and significant content within a broad theme in a timely fashion. To this end, this paper proposes a scalable and practical method to automatically construct social sensors for generic topics. Specifically, given minimal supervised training content from a user, we learn to identify topical tweets from millions of features capturing content, user and social interactions on Twitter. On a corpus of over 800 million English Tweets collected from the Twitter streaming API during 2013 and 2014 and learning for 10 diverse themes ranging from social issues to celebrity deaths to the "Iran nuclear deal", we empirically show that our learned social sensor automatically generalizes to unseen future content with high ranking and precision scores. Furthermore, we provide an extensive analysis of features and feature types across different topics that reveals, for example, that (1) largely independent of topic, simple terms are the most informative feature followed by location features and that (2) the number of unique hashtags and tweets by a user correlates more with their informativeness than their follower or friend count. In summary, this work provides a novel, effective, and efficient way to learn topical social sensors requiring minimal user curation effort and offering strong generalization performance for identifying future topical content.

1 Introduction

Social media sites such as Twitter present a double-edged sword for users. On one hand these sources contain a vast amount of novel and topical content that challenge traditional news media sources in terms of their timeliness and diversity. Yet on the other hand they also contain a vast amount of spam and otherwise low-value content for most users' information needs where filtering out irrelevant content is extremely time-consuming. Hence, while it is widely acknowledged that social media sources can be used as topical content sensors (indeed, an entire European Union project was focused on related "Social Sensor" research[1].), automatically learning high-precision sensors (i.e., ranking and retrieval methods) for arbitrary topics that generalize to future unseen content remains an open question in the literature and comprises the key problem we seek to address in this paper.

In this work, we contribute a novel supervised method for training social sensors with minimal user curation by using a small seed set of hashtags as topical proxies for automatic supervised data labeling. Then we proceed to train supervised classification and ranking methods to learn topical content from a large feature space of source users and their locations, terms, hashtags, and mentions. On a corpus of over 800 million English Tweets collected from the Twitter streaming API during 2013 and 2014 and covering 10 diverse topics ranging from social issues to celebrity deaths to the "Iran nuclear deal", we empirically show that two simple and efficiently trainable methods — logistic regression and naive Bayes — generalize well to unseen future topical content (including content with no hashtags) in terms of their mean average precision (MAP) and Precision@n for a range of n. Furthermore, we show that terms and locations are among the most useful features — surprisingly more so than hashtags, even though hashtags were used to label the data. And perhaps even more surprisingly, the number of unique hashtags and tweets by a user correlates more with their informativeness than their follower or friend count.

In summary, this work fills a major gap in event detection and tracking from social media on identifying emerging topics from long-running themes with minimal user supervision. Our results suggest that these sensors generalize well to unseen future topical content and provide a novel paradigm for the extraction of high-value content from social media.

2 Related Work

The concept of social media as a sensor is prevalent in the literature and in this section we survey four related areas of active research: (1) trending topic detection, (2) tweet recommendation, (3) friend sensors, and (4) and specific event detection such as earthquake or influenza sensors. Despite the partial overlap and superficial similarities between this paper and related social sensor work, we argue that no prior work has learned targeted social sensors for arbitrary topics using supervised learning methods as done in this paper.

Trending Topic Detection represents one of the most popular types of social sensor and can be subdivided into many

[1]http://www.socialsensor.eu/

categories. The first general category of methods define trends as topically coherent content and focus on clustering across lexical, linguistic, temporal and/or spatial dimensions (Petrović, Osborne, and Lavrenko 2010; Ishikawa et al. 2012; Phuvipadawat and Murata 2010; Becker, Naaman, and Gravano 2011; O'Connor, Krieger, and Ahn 2010; Weng and Lee 2011). The second general category of methods define trends as temporally coherent patterns of terms or keywords and focus largely on detecting bursts of terms or phrases (Mathioudakis and Koudas 2010; Cui et al. 2012; Zhao et al. 2011; Nichols, Mahmud, and Drews 2012; Aiello et al. 2013). The third category of methods extends the previous categories by additionally exploiting network structure properties (Budak, Agrawal, and El Abbadi 2011). Despite this important and very active area of work that can be considered a type of social sensor, trending topic detection is intrinsically unsupervised and not intended to detect targeted topics. In contrast, the work in this paper is based on supervised learning of a specific topical social sensor derived from the topical set of hashtags provided by the user.

Tweet Recommendation represents an alternate use of social sensors and falls into two broad categories: personalized or content-oriented recommendation and retweet recommendation. For the first category, the objective of personalized recommendation is to observe a user's interests and behavior from their user profile, sharing or retweet preferences, and social relations to generate tweets the user may like (Yan, Lapata, and Li 2012; Chen et al. 2012). The objective of content-oriented recommendation is to use source content (e.g., a news article) to identify and recommend relevant tweets (e.g., to allow someone to track discussion of a news article) (Krestel et al. 2015). For the second category, there has been a variety of work on retweet prediction that leverages retweet history in combination with tweet-based, author-based, and social network features to predict whether a user will retweet a given tweet (Can, Oktay, and Manmatha 2013; Xu and Yang 2012; Petrovic, Osborne, and Lavrenko 2011). Despite that the fact all of these methods recommend tweets, they — and recommendation methods in general — are not focused on a specific topic but rather on predicting tweets that correlate with the preferences of a specific user or that are directly related to specific content. Rather the focus with learning topical social sensors is to learn to predict for a broad theme (independent of a user's profile) in a way that generalizes beyond existing labeled topical content to novel future topical content.

Specific Event Detection builds social sensors as we do in this work but focuses on highly specific events such as a disasters or epidemics. For the use case of earthquake detection, an SVM can be trained to detect earthquake events and coupled with a Kalman filter for localization (Sakaki, Okazaki, and Matsuo 2013). In another example use case to detect health epidemics such as influenza, researchers build purpose-specific classifiers targeted to this specific epidemic (Culotta 2010; Aramaki, Maskawa, and Morita 2011), e.g, by exploiting knowledge of users' proximity and friendship along with the contageous nature of influenza (Sadilek, Kautz, and Silenzio 2012). While these targeted event detectors have the potential of providing high precision event detection, they are highly specific to the target event and do not easily generalize to learn arbitrary event-based or topic-based social sensors as provided in this work.

Friend Sensors: The fourth and final class of social sensors are intended for early event detection (Kryvasheyeu et al. 2014; García-Herranz et al. 2012) by leveraging the concept of the "friendship paradox"(Feld 1991), to build user-centric social sensors. We note that our topical social sensors represent a *superset* of friend sensors since our work includes author features that the predictor may learn to use if this proves effective for prediction. However, as shown in our feature analysis, user-based features are among the least informative feature types for our topical social sensors suggesting that general social sensors benefit from a wide variety of features well beyond those of author features alone.

3 Learning Topical Social Sensors

TODO: Formal learning framework.

One critical bottleneck for learning targeted topical social sensors is to achieve sufficient supervised content labeling. With data requirements often in the thousands of labels to ensure effective learning and generalization over a large candidate feature space (as found in social media), manual labeling is simply too time-consuming for many users and crowdsourced labels are both costly and prone to misinterpretation of users' information needs. Fortuitously, hashtags have emerged in recent years as a pervasive topical proxy on social media sites — hashtags originated on IRC chat, were adopted later (and perhaps most famously) on Twitter, and now appear on other social media platforms such as Instagram, Tumblr, and Facebook. Hence as a simple enabling insight that serves as a catalyst for effective topical social sensor learning, we leverage a (small) set of user-curated topical hashtags to efficiently provide a large number of supervised topic labels for social media content.

TODO: Formal math for defining the set of positive and negative labeled tweets via the hashtag set H.

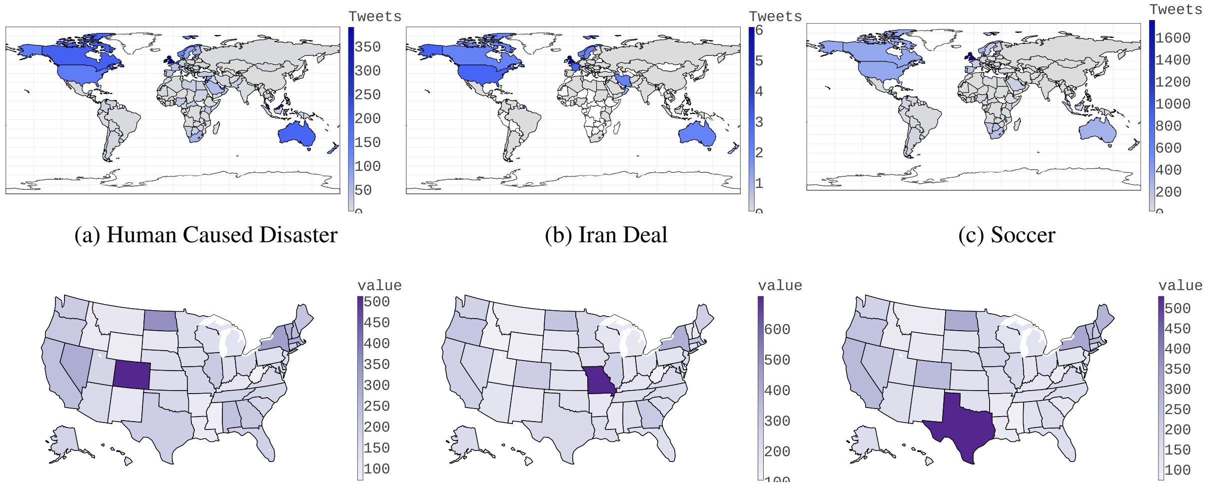
With the data labeling bottleneck resolved, we proceed to train supervised classification and ranking methods to learn topical content from a large feature space of source users and their locations, terms, hashtags, and mentions.

TODO: An simple enumeration of the training steps?

4 Data Description

Now we provide details of the Twitter testbed for topical social sensor learning that we evaluate in this paper. We crawled Twitter data using Twitter Streaming API for two years spanning 2013 and 2014 years. The total number of tweets collected is $829,026,458$. In the context of Twitter, we consider five feature types for each tweet. Each tweet has a *From* feature (i.e., the person who tweeted it), a possible *Location* (i.e., a string provided as meta-data), and a time stamp when it was posted. A tweet can also contain one or more of the following:

- *Hashtag*: a topical keyword specified using the # sign.
- *Mention*: a Twitter username reference using the @ sign.



(a) Human Caused Disaster (b) Iran Deal (c) Soccer

(d) Health Epidemics (e) Social Issues (f) Space

Figure 1: Distribution of tweets across International locations (top row) and U.S. locations (bottom row)

- *Term*: any non-hashtag and non-mention unigrams.

We provide more detailed statistics about each feature in Table 1. For example, authors (*From* users) have used a median value of 2 unique hashtags and a hashtag has been used on average by 10.08 unique users.

Fig. 1 shows per capita tweet frequency across different international and U.S. locations for different topics. While English speaking countries dominate English tweets, we see that the Middle East and Malaysia additionally stand out for the topic of Human Caused Disaster (MH370 incident), Iran and Europe for the "Iran deal", and soccer for many countries where it is popular. For U.S. states, we see that Colorado stands out for health epidemics (both whooping cough and pneumonic plague), Missouri stands out for social issues (#blacklivesmatter in St. Louis), and Texas stands out for space due to NASA's presence there.

5 Empirical Evaluation

With the 5 set of features defined as $From$, $Mention$, $Location$, $Term$, $Hashtag$, we proceed to define the methodology for retrieving ranked list of tweets for a given topic. The list of topics were defined to be a set of 10 various topics covering very specific e.g., $IranDeal$, and very broad e.g., $SocialIssues$ topics: $Tennis$, $Space$, $Soccer$, $IranDeal$, $HumanDisaster$ (HumanCausedDisaster), $CelebrityDeath$, $SocialIssues$, $NaturalDisaster$, $Epidemics$, and $LGBT$.

Our goal is to retrieve a ranked list of tweets T_i by employing machine learning methods using defined features. Fig **??** showed unique number of values for each feature. These values sum up to a total number of $538,365,507$ features. Also, as noted earlier, we are dealing with $829,026,458$ number of tweets. This shows the need for providing techniques to:

1. Annotate the tweets as topical and non-topical
2. Select a set of features for learning the model

#Unique Features

From	Hashtag	Mention	Location	Term
95,547,198	11,183,410	411,341,569	58,601	20,234,728

Feature Usage in #Tweets

Feature	Max	Avg	Median	Max entity
From	10,196	8.67	2	running_status
Hashtag	1,653,159	13.91	1	#retweet
Mention				
Location				
Term	241,896,559	492.37	1	rt
Feature Usage by #Users				
Hashtag	592,363	10.08	1	#retweet
Mention	26,293	5.44	1	dimensionist
Location	739,120	641.5	2	london
Term	1,799,385	-	1	rt
Feature Using #Hashtags				
From	18,167	2	0	daily_astrodata

Table 1: Feature Statistics of our $829,026,458$ tweet corpus.

In regards to annotating the tweets, first, a set of topical hashtags are manually curated for each topic. This set is annotated manually with 2 annotator individually and inner-annotator agreement was achieved by reviewing these sets by 2 more individuals. Each $Hashtag$ has a birthday which is defined as the first time it has been used in our dataset. The criteria of choosing hashtags included

- To be related to the topic
- To be preferably born during our time span i.e. not being used before e.g., #ebola
- The entire set of hashtag birth dates to cover the two years of our data

After choosing these sets, we tag a tweet as topical if it contains at least one topical hashtag.

In order to conduct our experiments, tweets are temporally divided over 2 years to provide train and test sets. Since our tweet labeling is through topical hashtags, this division is done in a way to retain enough hashtags for train, validation, and test timespan. To this purpose, hashtags are divided based on their birth date with 50 percent of hashtags being born at train timespan, 10 percent born at validation timespan, and the last 40 percent born at test timespan. Table 2 provides samples of hashtags, number of train hashtags, test hashtags, and topical tweets for each topic. We can see that some topics such as $HumanDisaster$ and $Soccer$ are more general topics and have higher number of topical tweets while some other ones such as $IranDeal$ is more specific, thus having less number of topical tweets.

Regarding feature selection, it is clear that it is not possible to learn a model with total number of $538,365,507$ features. Even if it was possible, we would have a problem for providing enough training samples, and our feature vectors would be extremely sparse considering 140 characters limitation of Twitter. Therefore, we performed a primary feature selection based on frequency of each feature. The feature selection process included:

- Cleaning $Term$ feature to remove stop-words
- Choosing a cut-off threshold of 159 for $From$, $Mention$, $Term$ features
- Choosing a cut-off threshold of 50 for $Location$ and $Hashtag$ features (These features had much lower number of unique values)

This results in roughly 1 million features, denoted as social features set SF where

$$\begin{aligned} SF_m \in \\ \{Hashtag_{m1}, From_{m2}, Mention_{m3}, Term_{m4}, Location_{m5}\} \\ m = \{1, ..., 1166582\}, m1 = \{1, .., 184702\}, \\ m2 = \{1, ..., 361789\}, m3 = \{1, .., 244478\}, \\ m4 = \{1, .., 317846\}, m5 = \{1, .., 57767\} \end{aligned} \quad (1)$$

Classification Algorithms

Now that we defined the primary steps for preparing the features and dataset, we can use them to build a ranking approach for topical tweet selection. Our method is based on classification/ranking approaches defined in the literature to weight features. These weights are further used to rank tweets for each topic. Here, we use the following classification approaches:

1. Logistic Regression
2. Naive Bayes
3. Rocchio (centroid)
4. RankSVM

To this purpose, we define the problem as assigning a weight W_i to tweet X as a measure of similarity to the given topics t_i. W_i is the sum of weights of features in the x_i:

$$W_i = \sum_k w_k \times f_k \quad (2)$$

where w_k is the weight of feature f_k and $f_k \in \{true, false\}$ represents whether each of the features in $SF_m, m = \{1, .., 1166582\}$ is present in tweet X or not. The weights w_k are learned by applying one of the classification algorithms.

For example, in case of Naive Bayes, the problem is defined as probability of tweet being in topic t_i given feature vector F_x for

$$P(t_i|X) = P(t_i|F_x) = \frac{P(t_i)P(F_x|t_i)}{P(F_x)} \quad (3)$$

In order to learn the models, we take the following steps for each topic:

1. All the positive tweets for the given topic, in addition to sub-sampled set of negative tweets are collected
2. A set of top N features are selected based on the Mutual Information values of features for the given topic. N is selected during hyper-parameter tuning phase from the set of $10E1, 10E2, 10E3, 10E4, 1166582$ values.
3. Train, validation, and train sets are further built based on the division process explained in 5 and selected set of top N features.
4. The model's hyper-parameter in addition to N parameter are tuned on the validation set
5. The model learns the weight vector based on the tuned hyper-parameters on the full train set and validation set

The Liblinear (Fan et al. 2008) package is used for implementing LR and $RankSVM$. The reason for deciding to tune the models on top N features based on Mutual Information, comes from our primary feature analysis on the dataset which showed the ability of Mutual Information measure to pick more correlated features for each topic. This is discussed in more details in section 6. The model hyper-parameters are tuned for LR and NB. The $Rocchio$ method is parameter free and the LibLinear (Fan et al. 2008) implementation of $RankSVM$ does not provide manual tuning of the model's hyper-parameter.

Analysis

After experimenting each mentioned model on our dataset, we provide the following metrics:

- MAP: Mean average precision for a set of topics is the mean of the average precision scores for each topic.
- P@K: Precision at K for $K \in \{10, 100, 1000\}$, the number of relevant results on the first K search results page

The model's hyper-parameters are tuned based on MAP scores, having MAP as our most important metrics. Table 3 provides these metrics for each topic. Logistic Regression method is the method that performs best on average. Generally, Naive Bayes performed comparable/better to Logistic Regression having second best average value of MAP. We

	Tennis	Space	Soccer	IranDeal	HumanDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT
#TrainHashtags	58	98	126	12	49	28	31	31	52	29
#TestHashtags	36	63	81	5	29	16	19	19	33	17
#TopicalTweets	55,053	239,719	860,389	8,762	408,304	163,890	230,058	230,058	210,217	282,527
Sample Hashtags	#usopenchampion	#asteroids	#worldcup	#irandeal	#gazaunderattack	#robinwilliams	#policebrutality	#policebrutality	#ebola	#loveislove
	#novakdjokovic	#astronauts	#lovesoccer	#iranfreedom	#childrenofsyria	#ripmandela	#michaelbrown	#michaelbrown	#virus	#gaypride
	#wimbledon	#satellite	#fifa	#irantalk	#iraqwar	#ripjoanrivers	#justice4all	#justice4all	#vaccine	#uniteblue
	#womenstennis	#spacecraft	#realmadrid	#rouhani	#bombthreat	#mandela	#freetheweed	#freetheweed	#chickenpox	#homo
	#tennisnews	#telescope	#beckham	#nuclearpower	#isis	#paulwalker	#newnjgunlaw	#newnjgunlaw	#theplague	#gaymarriage

Table 2: Test/Train Hashtag samples and statistics

also provide the top 5 tweets returned by Logistic Regression for each topic as anecdotal results in Table 4. In this table, the signs in the beginning of the tweet represent the following:

- ✗represents the tweets that are method has incorrectly ranked as highly topical
- ✓represents the tweets correctly ranked as highly topical
- ★represents the tweets that don't have any topical hashtags and therefore are not labeled as correctly ranked topical. However, looking at the tweets, we can see that they are in fact related to the topic

The fact that there are cases of tweets not being correctly labeled as topical, provides evidence that our method of labeling tweets has limitations and our MAP and P@K values are actually suffering from this problem. However, this shows the power of Logistic Regression method in generalizing from a small set of hashtags.

6 Feature Analysis

. In this section, we analyze the informativeness of each feature for learning topical tweets by looking at different characteristics for each feature in our dataset. For example, one characteristic of hashtags could be the number of the tweets that contain those hashtags. Does this have an effect on importance of the hashtag when it comes to learning topical tweets or not. In this sense, this section would bring insights to the following questions:

- **What are the best features for learning social sensors, do they differ by topic? (Why?)**
- **For each feature type, do any attributes correlate with importance?**

A famous method for measuring informativeness is Mutual Information which is a measure of amount of information one random variable contains about another random variable. In order to calculate amount of information that a feature $f_k \in \{from, hashtag, mention, term, location\}$ provides w.r.t $t_i \in \{NaturalDisaster, Epidemics, ...\}$, mutual information is defined as:

$$I(t_i, f_k) = \sum_{t_i \in \{true,false\}} \sum_{f_k \in \{true,false\}} p(f_k, t_i) \log \left(\frac{p(f_k, t_i)}{p(f_k)p(t_i)} \right) \quad (4)$$

Higher values for this metric indicates more informative features for the specified topic.

In order to answer the first question on what are the best features for learning social sensors, we provide mean of Mutual Information values for each feature across different topics in Table 3. The last column in this table shows average of mean Mutual Information for the feature. The following observations are from the analysis of Table 3:

- $Term$ feature is the most prevalent feature and in general, the more features you have, the better the chance that one is useful.
- $Location$ and $Hashtag$ feature provide second and third most informative features respectively.
- A few topics such as $IranDeal$ and tennis are less sensitive to selection of a specific features.
- Location feature provides more information regarding $HumanDisaster$, $LBGT$, and $Soccer$ topics.
- Sorting features based on their average mean values across different topics results in the following order: $Term$, $Location$, $Hashtag$, $Mention$, $From$

In general, this presents evidence on the need for learning the weights of features for each topic, because there is no specific selection of features that would separate various topics from each other.

Also, in order to show the power of Mutual Information criteria, we present the top 5 features for each topic in table 5. It can be observed how different locations, hashtags, or terms showed as the top features based on mutual information are actually in relation with the specific topic.

In order to answer the second question on whether any attributes correlate with importance for each feature, we provide two set of analysis. The first one, provides Mutual Information values of each feature across feature's attribute values shown by violin plots in figure 4. The attributes for each feature are:

- From: favorite count (the number of tweets the user has favorited), followers count (the number of users who follow the user), friends count (the number of users followed by the user), hashtag count (number of hashtags used by the user), tweet count (the number of tweets from the user)
- Hashtag: tweet count, user count (the number of users using the hashtag)
- Location: user count
- Mention: tweet count

		Tennis	Space	Soccer	IranDeal	HumanDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT	Mean
LR	**MAP**	**0.918**	0.870	0.827	0.811	0.761	0.719	0.498	**0.338**	**0.329**	**0.165**	**0.623±0.19**
NB	**MAP**	0.908	**0.897**	0.731	**0.824**	**0.785**	**0.748**	**0.623**	0.267	0.178	0.092	0.605±0.22
Rocchio	**MAP**	0.690	0.221	**0.899**	0.584	0.481	0.253	0.393	0.210	0.255	0.089	0.407±0.18
RankSVM	**MAP**	0.702	0.840	0.674	0.586	0.603	0.469	0.370	0.248	0.136	0.082	0.471±0.18
LR	**P@10**	**1.000**	0.000	0.200	0.700	**0.600**	0.000	0.100	0.200	0.300	0.500	0.360±0.24
NB	**P@10**	**1.000**	**0.900**	0.700	0.600	**0.600**	**0.700**	**1.000**	0.100	0.400	0.100	**0.610±0.23**
Rocchio	**P@10**	0.800	0.000	**1.000**	**0.900**	0.000	0.000	0.000	**0.500**	**0.500**	0.100	0.380±0.29
RankSVM	**P@10**	**1.000**	0.800	0.600	0.800	0.400	0.300	0.000	0.100	0.000	**0.200**	0.420±0.26
LR	**P@100**	0.950	0.580	0.650	0.870	0.620	0.490	0.640	**0.690**	**0.790**	**0.210**	**0.649±0.15**
NB	**P@100**	**0.980**	**0.850**	0.600	**0.880**	0.750	**0.860**	**0.730**	0.230	0.090	0.190	0.616±0.23
Rocchio	**P@100**	**0.980**	0.000	**1.000**	0.690	0.170	0.000	0.280	0.170	0.680	0.120	0.409±0.28
RankSVM	**P@100**	0.730	0.720	0.310	0.700	**0.880**	0.440	0.480	0.340	0.020	0.100	0.472±0.20
LR	**P@1000**	**0.963**	**0.954**	0.816	**0.218**	0.899	0.833	**0.215**	0.192	**0.343**	**0.071**	**0.550±0.26**
NB	**P@1000**	0.954	**0.954**	0.716	**0.218**	**0.904**	**0.881**	**0.215**	**0.195**	0.141	0.060	0.524±0.28
Rocchio	**P@1000**	0.604	0.000	**0.925**	**0.218**	0.359	0.000	**0.215**	0.167	0.144	0.065	0.270±0.21
RankSVM	**P@1000**	0.799	0.922	0.764	**0.218**	0.525	0.547	**0.215**	0.173	0.154	0.064	0.438±0.22

Table 3: Different learning methods results on topics with hyper-parameter tuning based on MAP

Tennis	**Space**
✓rt @espntennis: shock city. darcis drops rafa in straight sets. first time nadal loses in first rd of a. major in career. #espnwimbledon #w	✗rt @jaredleto: rt @30secondstomars: icymi: mars performing a cover of @rihanna's #stay on australia's @triplemmelb - video _ http://t.co/uq
✓@ESPNTennis: Shock city. Darcis drops Rafa in straight sets. First time Nadal loses in first rd of a. Major in career.	✗voting mars @30secondstomars @jaredleto @shannonleto @tomofromearth xobest group http://t.co/dlsozvjinf
✓@ESPNTennis: Djokovic ousts the last American man standing @Wimbledon, beating Reynolds 7-6 6-3 6-1 #ESPNWimbledon	✗rt @jaredleto_com: show everyone how much you are proud of @30secondstomars !#mtvhottest 30 seconds to mars http://t.co/byxnri4t67
✓Nadal's a legend. After 3 years; Definitely He's gonna be the best of all the time. Unbelievable performance. @RafaelNadal #USOpenFinal	✗rt @30secondstomars: missed the big news? mars touring with @linkinpark + special guests @afi this summer!_ http://t.co/3e5rm9pwrd
✓@calvy70 @ESPNTennis @Wimbledon I see, thanks for the info and enjoy #Wimbledon2014	✗rt @30secondstomars: to the right,to the left,we will fightto the death.go #intothewildonvyrt with mars, starting weekly, nov 30 _ htt
Soccer	**IranDeal**
✗rt @tomm_dogg: #thingstodobeforeearthends spend all my money.	✓rt @iran_policy: @vidalquadras:@isjcommittee has investigated 10 major subjects of irans controversial #nuclear program #irantalksvienna
★@mancityonlineco nice performance	✓rt @iran_policy: @vidalquadras:@isjcommittee has investigated 10 major subjects of irans controversial #nuclear program #irantalksvienna
★rt @indykaila: podolski: "let's see what happens in the winter. the fact is that i'm not happy with it, that's clear." @arsenal	✗rt @negarmortazavi: thank you @hassanrouhani for retweeting. let's hope for a day when no iranian fears returning to their homeland. http:/
★rt @indykaila: wenger: "i don't believe match-fixing is a problem in england." #afc	✗rt @iran_policy: iran: details of savage attack on political prisoners in evin prison http://t.co/xdzuakqdiv #iran #humanrights
✗@indykaila you never got back to me about tennis this week	✓rt @iran_policy: chairman ros-lehtinen speaking on us commitment 2 protect camp liberty residents. #iranhrviolations http://t.co/1g6dhx1znu
HumanDisaster	**CelebrityDeath**
✓rt @baselsyrian: there've been peaceful people in #homs not terrorists! #assad,enemy of #humanity destroyed it. #eyeonhoms #withsyria http:	★rt @sawubona_chris: today is my birthday & also the day my hero @nelsonmandela has died. lets never forget what he taught us. forgiveness i
✓what a helpless father, he can do nothing under #assad's siege!#speakup4syrianchildren http://t.co/vgle3byebw#syria #syriawarcrimes #un	★rt @nelsonmandela: death is something inevitable.when a man has done what he considers to be his duty to his people&his country,he can res
★exclusive: us formally requested #un investigation; russia pressured #assad to no avail;chain of evidence proof hard http://t.co/560t2rvdfw	★rt @nelsonmandela: la muerte es algo inevitable.cuando un hombre ha hecho lo que considera que es su deber para con su gente y su pas,pued
★#save_aleppo from #assadwarcrimes#save_aleppo from #civilians -targeted shelling of #assad regime#syria #aleppo http://t.co/k3dfxh0pxl	✗#jacques #kallis: a phenomenal cricketing giant of all time - #cricket #history #southafrica http://t.co/ms5pmwoag9
✓rt @canine_rights: why does the #un allow this to continue? rt@tintin1957 help raise awareness of the suffering in #syriawarcrimes http://t	✗@sudesh1304 south africa has the most beautiful babies....so diverse,so unique...so god!! lol #durban #southafrica
SocialIssues	**NaturalDisaster**
★the us doesn't actually borrow is the thing. i believe in a creationist theory of the us dollar @usanationdebt @nationaldebt	✗us execution in #oklahoma : not cruel and unusual? maybe just barbaric, inhumane and reminiscent of the dark ages!
★rt @2anow: according to @njsenatepres women's rights do not include this poor nj mother's right to defend herself http://t.co/xzbslnqkh6 #	✗#haiti #politics - the haiti-dominican crisis - i agree with how martelly is handling the situation: i totally... http://t.co/ro4pswsszs
★rt @2anow: confiscation ? how many carry permits are in the senate and assembly? give us ours or turn them in. @senatorlorettaw @lougreenw	★rt @soilhaiti: a new reforestation effort in #haiti. local compost, anyone? http://t.co/xpad0rqbjk @richardbranson @clintonfdn @virginunite
★rt @2anow: vote with your wallet against #guncontrolforest city enterprises does not support the #2a http://t.co/tpkok3berm#nj2as #tcot	✗mes cousins jamais ns hantent les nuits de duvalier #haiti #duvalier
★@2anow @momsdemand @jstines3 they dont have a plan for that , which is why they should never be allowed to take our guns	✓tony burgener of @swisssolidarity says you can't compare the disaster response in #haiti with the response to #haiyan in #philippines @iheid
Epidemics	**LGBT**
✓rt @who: fourteen of the susp. & conf. ebola cases in #conakry, #guinea, are health care workers, of which 11 died #askebola	★rt @jackmcoldcuts: @lunaticrex @fingersmalloy @toddkincannon @theanonliberal anthony kennedy just wrote opinion granting...
✗@who who can afford also been cover in government health insurance [with universal health coverage]	✗@toddkincannon your personal account, your interest. separate from your business.
✓#ebolaoutbreak this health crisis..unparalleled in modern times, @who dir. aylward - requires $1 billion to stem http://t.co/rjzqhydb3d	✗why would you report someone as spam if he is not spam? @illygirlbrea @toddkincannon
✗rt @medsin: @who are conducting a survey on the social determinants of health in medical teaching. fill the survey in at https://t.co/aj59x	✗rt @t3h_arch3r: @toddkincannon thanks for your tl having the female realbrother. between them is 600 lbs. 104 iq points. and a lot of hate.
✗augmentation vertigineuse de 57,4% en 1 an des actes islamophobes en france, dit le collectif contre l'islamophobie http://t.co/2qjhocegi5	✗@toddkincannon who us dick trickle.

Table 4: Top Tweets for each topic based on MAP tuned results

- Term: tweet count

As we can see in the violin plots, the general pattern is that the more number of tweets, users, or hashtags count a feature has, the higher the chance of becoming topical will be. This pattern exists on other attributes of $From$ feature, although a bit less clear than the tweets, users, or hashtags counts attributes. In addition, we further analyzed the density plots of favorite count, follower count, friends count, hashtag count attributes of $From$ feature shown in Fig 5. These plots represent a bi-modality in the distribution. Further analysis of data showed that the top mode belongs to users who have at least one topical tweet while bottom mode are users with no topical tweet.

7 Conclusions and Future Work

This work fills a major gap in event detection and tracking from social media on identifying emerging topics from long-running themes with minimal user supervision. Our results suggest that these sensors generalize well to unseen future topical content and provide a novel paradigm for the extraction of high-value content from social media. Future work should explore the following enhanced topical social sensor learning tasks: (1) optimizing rankings not only for topicality but also to minimize the lag-time of novel content identification, (2) optimizing queries for boolean retrieval oriented APIs such as Twitter, and (3) utilizing more social network structure to exploit a more expressive graph-based features.

References

[Aiello et al. 2013] Aiello, L. M.; Petkos, G.; Martín, C. J.; Corney, D.; Papadopoulos, S.; Skraba, R.; Göker, A.; Kompatsiaris, I.; and Jaimes, A. 2013. Sensing trending topics in twitter. *IEEE Transactions on Multimedia* 15(6):1268–1282.

[Aramaki, Maskawa, and Morita 2011] Aramaki, E.; Maskawa, S.; and Morita, M. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11.

[Becker, Naaman, and Gravano 2011] Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on twitter. In *Proceedings of the*

Topics/Top10	NaturalDisaster	Epidemics	IranDeal	SocialIssues	LBGT	HumanDisaster	CelebrityDeath	Space	Tennis	Soccer
From	earthquake_wo	changedecopine	mazandara	nsingerdebtpaid	eph4_15	ydumozyf	nmandelaquotes	daily_astrodata	tracktennisnews	losangelessrh
From	earthalerts	drdaveanddee	hhadi119	debtadvisoruk	mgdauber	syriatweeten	boiknox	freesolarleads	tennis_result	shoetale
From	seelites	joinmentornetwk	140iran	debt_protect	stevendickinson	tintin1957	jacanews	houston__jobs	i_roger_federer	sport__agent
From	globalfloodnews	followebola	setarehgan	negativeequityf	lileensvf1	sirajsol	ewnreporter	star_wars_gifts	tennislessonnow	books_you_want
From	gcmcdrought	localnursejobs	akhgarshabaneh	dolphin_ls	truckerbooman	rt3syria	paulretweet	lenautilus	kamranisbest	makeupbella
Hashtag	earthquake	health	iran	ferguson	tcot	syria	rip	science	wimbledon	lfc
Hashtag	haiyan	uniteblue	irantalks	mikebrown	p2	gaza	riprobinwilliams	starwars	usopen	worldcup
Hashtag	storm	ebola	rouhani	ericgarner	pjnet	isis	ripcorymonteith	houston	tennis	arsenal
Hashtag	tornado	healthcare	iranian	blacklivesmatter	uniteblue	israel	mandela	sun	nadal	worldcup2014
Hashtag	prayforthephilippines	depression	no2rouhani	fergusondecision	teaparty	mh370	nelsonmandela	sxsw	wimbledon2014	halamadrid
Location	philippines	usa	tehran	st.louis	usa	malaysia	southafrica	germany	london	liverpool
Location	ca	ncusa	u.s.a	mo	bordentown	palestine	johannesburg	roodepoort	uk	manchester
Location	india	garlandtx	nederland	usa	newjersey	syria	capetown	houston	india	london
Location	newdelhi	oh-sandiego	iran	dc	sweethomealabama!	israel	pretoria	austin	pakistan	nigeria
Location	newzealand	washington	globalcitizen	washington	aurora	london	durban	tx	islamabad	india
Mention	oxfamgb	foxtramedia	4freedominiran	deray	jjauthor	ifalasteen	nelsonmandela	bizarro_chile	wimbledon	lfc
Mention	weatherchannel	obi_obadike	iran_policy	natedrug	2anow	revolutionsyria	realpaulwalker	nasa	usopen	arsenal
Mention	redcross	who	hassanrouhani	antoniofrench	govchristie	drbasselabuward	robinwilliams	j_ksen	andy_murray	realmadriden
Mention	twcbreaking	obadike1	un	bipartisanism	a5h0ka	mogaza	rememberrobin	jaredleto	serenawilliams	ussoccer
Mention	abc7	c25kfree	statedept	theanonmessage	barackobama	palestinianism	tweetlikegiris	30secondstomars	espntennis	mcfc
Term	philippines	health	iran	police	obama	israel	robin	cnblue	murray	madrid
Term	donate	ebola	regime	protesters	gun	gaza	williams	movistar	tennis	goal
Term	typhoon	acrx	nuclear	officer	rights	israeli	nelson	enero	federer	cup
Term	affected	medical	iranian	protest	america	killed	mandela	imperdible	djokovic	manchester
Term	relief	virus	resistance	cops	gop	children	cory	greet	nadal	match

Table 5: Top 5 features for each topic based on Mutual Information

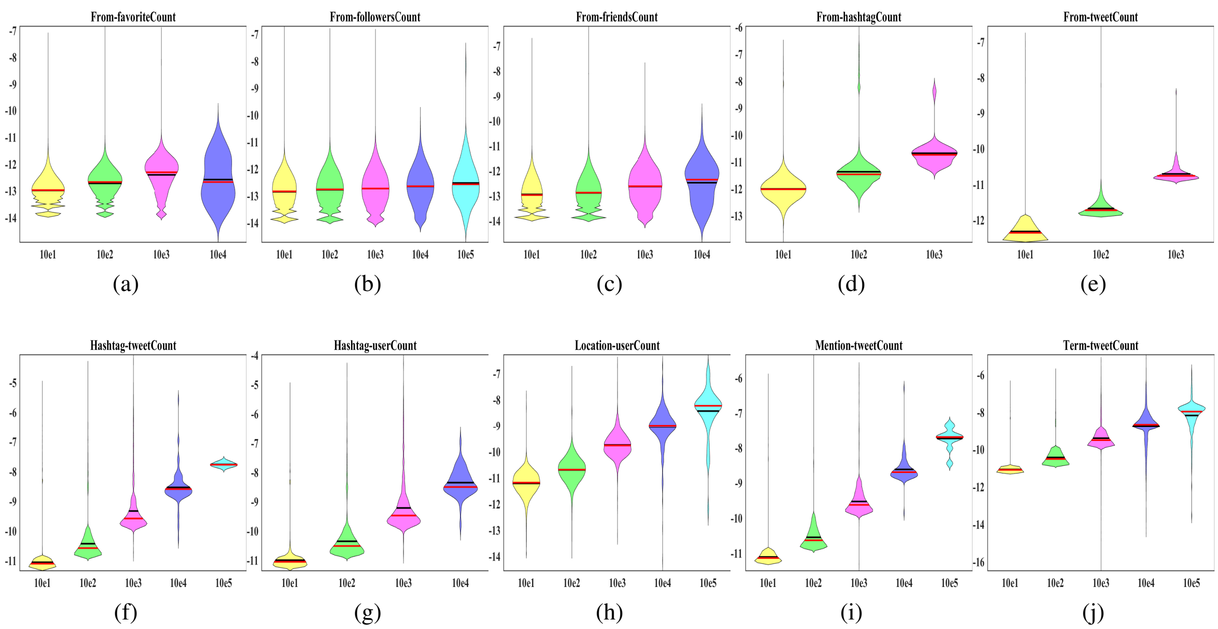


Figure 4: ViolinPlots for feature attributes counts vs. MI. Top row shows attributes {favoriteCount, followerCount, friendCount, hashtagCount, tweetCount} for $From$ feature. Bottom row shows attributes tweetCount and/or userCount for $Hashtag$, $Location$, $Mention$,and $Term$ features.

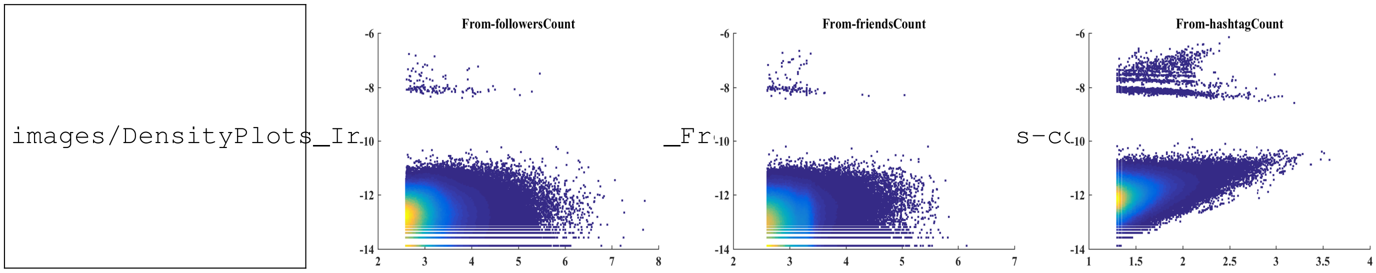
Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011.

[Budak, Agrawal, and El Abbadi 2011] Budak, C.; Agrawal, D.; and El Abbadi, A. 2011. Structural trend analysis for online social networks. *PVLDB* 4(10):646–656.

[Can, Oktay, and Manmatha 2013] Can, E. F.; Oktay, H.; and Manmatha, R. 2013. Predicting retweet count using visual cues. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, 1481–1484.

[Chen et al. 2012] Chen, K.; Chen, T.; Zheng, G.; Jin, O.; Yao, E.; and Yu, Y. 2012. Collaborative personalized tweet recommendation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, 661–670. New York, NY, USA: ACM.

[Cui et al. 2012] Cui, A.; Zhang, M.; Liu, Y.; Ma, S.; and



(a) (b) (c) (d)

Figure 5: DensityPlots for feature attributes counts vs. MI. (a-d) show attributes {favoriteCount, followerCount, friendCount, hashtagCount} for $From$ feature

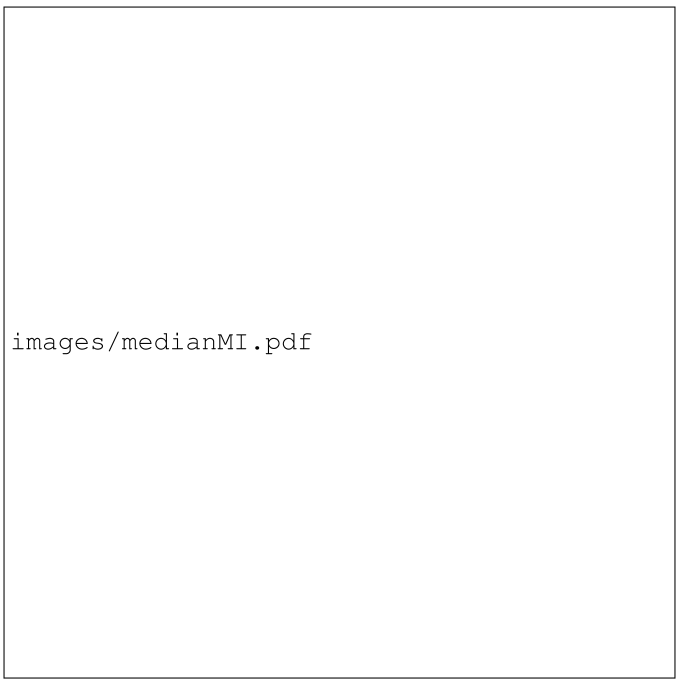


Figure 2: Median MI for different features vs. Topics, last two column show mean value and stderr across all topics

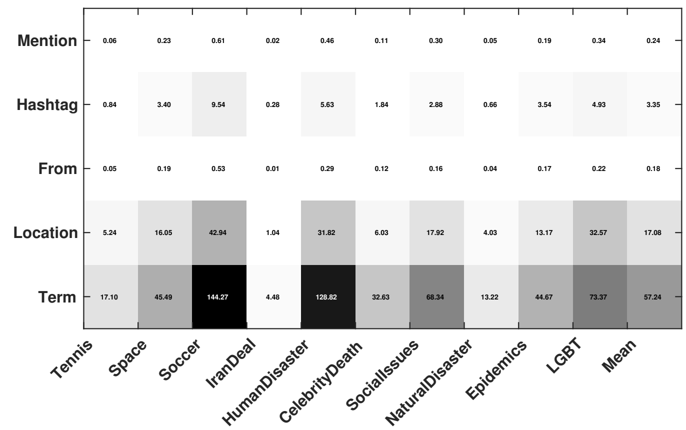


Figure 3: Average MI for different features vs. Topics, last two column show mean value and stderr across all topics

Zhang, K. 2012. Discover breaking events with popular hashtags in Twitter. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, 1794–1798.

[Culotta 2010] Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10.

[Fan et al. 2008] Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; and Lin, C. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

[Feld 1991] Feld, S. L. 1991. Why your friends have more friends than you do. *American Journal of Sociology* 1464–1477.

[García-Herranz et al. 2012] García-Herranz, M.; Egido, E. M.; Cebrián, M.; Christakis, N. A.; and Fowler, J. H. 2012. Using friends as sensors to detect global-scale contagious outbreaks. *PloS one* abs/1211.6512.

[Ishikawa et al. 2012] Ishikawa, S.; Arakawa, Y.; Tagashira, S.; and Fukuda, A. 2012. Hot topic detection in local areas using Twitter and wikipedia. In *ARCS Workshops (ARCS), 2012*, 1–5.

[Krestel et al. 2015] Krestel, R.; Werkmeister, T.; Wiradarma, T. P.; and Kasneci, G. 2015. Tweet-recommender: Finding relevant tweets for news articles. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, 53–54. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

[Kryvasheyeu et al. 2014] Kryvasheyeu, Y.; Chen, H.; Moro, E.; Hentenryck, P. V.; and Cebrián, M. 2014. Performance of social network sensors during hurricane sandy. *PLoS one* abs/1402.2482.

[Mathioudakis and Koudas 2010] Mathioudakis, M., and Koudas, N. 2010. Twittermonitor: trend detection over the Twitter stream. In *Proceedings of the ACM SIGMOD In-*

ternational Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, 1155–1158.

[Nichols, Mahmud, and Drews 2012] Nichols, J.; Mahmud, J.; and Drews, C. 2012. Summarizing sporting events using Twitter. In *17th International Conference on Intelligent User Interfaces, IUI '12, Lisbon, Portugal, February 14-17, 2012*, 189–198.

[O'Connor, Krieger, and Ahn 2010] O'Connor, B.; Krieger, M.; and Ahn, D. 2010. Tweetmotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*.

[Petrović, Osborne, and Lavrenko 2010] Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 181–189. Stroudsburg, PA, USA: Association for Computational Linguistics.

[Petrovic, Osborne, and Lavrenko 2011] Petrovic, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in Twitter. In *ICWSM*.

[Phuvipadawat and Murata 2010] Phuvipadawat, S., and Murata, T. 2010. Breaking news detection and tracking in Twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology - Workshops, Toronto, Canada, August 31 - September 3, 2010*, 120–123.

[Sadilek, Kautz, and Silenzio 2012] Sadilek, A.; Kautz, H. A.; and Silenzio, V. 2012. Modeling spread of disease from social interactions. In *Proceedings of the Sixth International Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012*.

[Sakaki, Okazaki, and Matsuo 2013] Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2013. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on* 25(4):919–931.

[Weng and Lee 2011] Weng, J., and Lee, B. 2011. Event detection in Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*.

[Xu and Yang 2012] Xu, Z., and Yang, Q. 2012. Analyzing user retweet behavior on Twitter. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August 2012*, 46–50.

[Yan, Lapata, and Li 2012] Yan, R.; Lapata, M.; and Li, X. 2012. Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, 516–525. Stroudsburg, PA, USA: Association for Computational Linguistics.

[Zhao et al. 2011] Zhao, S.; Zhong, L.; Wickramasuriya, J.; and Vasudevan, V. 2011. Human as real-time sensors of social and physical events: A case study of Twitter and sports games. *Technical Report TR0620-2011, Rice University and Motorola Mobility* abs/1106.4300.