Learning Topical Social Sensor

Authors
Affiliations

Abstract

Twitter represents a massively distributed social sensor of a rich underlying topic space that drives its content generation. Yet Twitter content is so diverse, decentralized, and dynamic in nature, that it is hard to automatically aggregate this topical content. To address this need, we provide a novel way of learning topical social sensors on Twitter that learn from a provided set of topical hashtags and generalize to identify topical tweets with previously unseen tags. These learning social sensors leverage a variety of user-based, hashtag-based, term-based, and location-based features for distinguishing topical from non-topical tweets; we further analyze these features to understand which features are most useful and why. We further assess general global topical trends and how our learning sensors are able to follow these trends by drawing from a rich variety of sources on the Twittersphere to enable a first generation of learning social sensors for Twitter.

Introduction

Twitter hosts lots of information, on average more than $2,200$ new tweets every second. This can get up to 3 to 4 times increase during large events such as tsunami. [1]

- Twitter is a vast sensor of content generated by latent phenomena (e.g., flu, political sentiment, elections, environment).
- Learning topical social sensors (politicians in NY, road conditions in Toronto) – very broad topics for which its hard to manually specify a useful query.
- But there is interesting topical content and wouldn't it be cool if we could learn a social sensor for a targeted topic?
- Key insight is that hashtags are topical and can be used to bootstrap a supervised learning system that as we will show generalizes well beyond the seed hashtags.
- Conclusion is a new way to build topical real-time feeds that are otherwise difficult to do with existing Twitter tools (???).

sectionLearning Topical Social Sensors

Start off with the questions that we want to answer in this section:

- How to evaluate, labeling (problem of no supervised labels for tweets, indirect via hashtags as topical surrogates, leads to question of hashtag curation)?

- Which classification algorithm is best / most robust for learning topical social sensors?

Dataset Statistics

We crawled Twitter data using Twitter Streaming API for two years spanning 2013 and 2014 years. This type of crawling provides us with a very sparse set of data, roughly 1% of all tweets [2]. The total number of tweets collected is $829,026,458$. In the context of Twitter, we consider a list of 5 features for each tweet. Each tweet has a $From$, the person who tweeted it, and a $Time$ which is the date information of the tweet. It can also contain

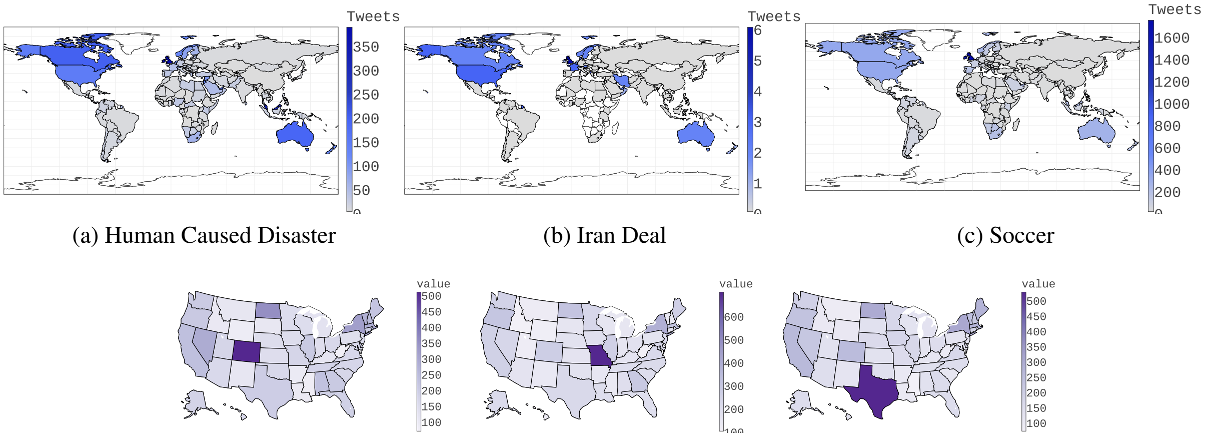
- $Hashtag(s)$, keywords specified using # sign
- $Mention(s)$, another Twitter username being mentioned using @ sign
- $Term(s)$, uni-grams which we extract from the 140 characters of the tweet. These uni-grams are later cleaned to remove $Term$s with no meaning (total number of $Term$s before cleaning was $20,234,729$)

Table 1 provides more detail statistics about each feature. For each feature, we reported the count of the feature in our dataset, in addition to maximum, average, median counts of each feature across the tweets. Lower part of the table provides these counts across user dimension meaning that for example a hashtag has been used in average by 10.08 users. Last part of the table shows the statistics for the hashtag usage of our users e.g., users have used 2 hashtags in average.

Figure **??** shows details of number of tweets per month and figure **??** shows the power law plots of tweet counts and hashtag counts for users. We chose 10 topics for our experiments. Tweets are temporally divided over 2 years to provide train and test sets. Table 2 provides samples of training hashtags and number of train hashtags, test hashtags, topical tweets for each topic. Some topics such as $HumanCausedDisaster$ and $Soccor$ are more general

[1] https://blog.twitter.com/2011/the-engineering-behind-twitter-s-new-search-experience

[2] http://allthingsd.com/20101110/twitter-firehose-too-intense-take-a-sip-from-the-garden-hose-or-sample-the-spritzer—



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

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

Figure 1: Choropleths of top: International map and bottom: U.S. map

	Tweets				
Feature	**Max**	**Avg**	**Median**	**Max entity**	**Count**
From	10,196	8.67	2	running_status	95,547,198
Hashtag	1,653,159	13.91	1	#retweet	11,183,410
Mention					411,341,569
Location					58,601
Term	2024529	7,450.58	323	taking	20,234,729
	Users				
From					
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	6,616.65	305	taking	
	Hashtags				
From	18,167	2	0	daily_astrodata	

Table 1: Feature Statistics

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.

Figure **??** shows distribution of tweets across different location in U.S. and international locations overall and for each topic(?).

Experimental Methodology

How we curated hashtags: need to make up good story here. Inner-annotator agreement of 3/4.

Train/validation/test split date selection – temporally .5,.1,.4

Feature selection: threshold per feature 159 and 50 (just explain rationale for lower hashtag and location thresholds).

Formal notation, how do we train/test and tune hyperparameters for a generic classifier.

Classification Algorithms

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

All above over 1,000,000 features, *same* training data for all algorithms.

Not breaking down by feature type yet – that's for the feature analysis section.

Analysis

- Table of rows:alg, cols: MAP, P@k (k in 10,100,1000) with stderrs over all topics

- Could do a bar graph (below) each for MAP, P@100 with topics as major columns and algs as neighboring bars

- Anecdotal results for each topic – point out deficiency in our labels (a good thing, we generalized well from small hashtag set), manual evaluation of relevance for top-100 for best algorithm?

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:

Topics/Top10	NaturalDisaster	Epidemics	IranDeal	SocialIssues	LBGT	HumanCausedDisaster	CelebrityDeath	Space	Tennis	Soccer
#TrainHashtags	31	52	12	31	29	49	28	98	58	126
#TestHashtags	18	33	5	19	17	29	16	63	36	81
#TotalTopicalTweets	42,987	210,217	8,762	230,058	282,527	408,304	163,890	239,719	55,053	860,389
Sample Train Hashtags	#earthquake	#ebola	#irandeal	#policebrutality	#loveislove	#gazaunderattack	#robinwilliams	#asteroids	#usopenchampion	#worldcup
	#storm	#virus	#iranfreedom	#michaelbrown	#gaypride	#childrenofsyria	#ripmandela	#astronauts	#novakdjokovic	#lovesoccer
	#tsunami	#vaccine	#irantalk	#justice4all	#uniteblue	#iraqwar	#ripjoanrivers	#satellite	#wimbledon	#fifa
	#abfloods	#chickenpox	#rouhani	#freetheweed	#homo	#bombthreat	#mandela	#spacecraft	#womenstennis	#realmadrid
	#hurricanekatrina	#theplague	#nuclearpower	#newnjgunlaw	#gaymarriage	#isis	#paulwalker	#telescope	#tennisnews	#beckham

Table 2: Test/Train Hashtag samples and statistics

Method	Metric	Tennis	Space	Soccer	IranDeal	HumanCausedDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT	Mean±Std
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

$$I(t_i, f_k) = \sum_{t_i \in \{true, false\}} \sum_{f_k \in \{true, false\}} p(f_k, t_i) \log \left(\frac{p(f_k, t_i)}{p(f_k) p(t_i)} \right) \tag{1}$$

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

First, we provide mutual information values for each feature across different topics shown by boxplots in figure **??**, and average values of mutual information for each feature vs different topics shown in table 2. The last column in table 2 shows average mutual information for the feature with the standard error range provided. We make a few observations from the analysis of Table 2:

- Term features provide more information for all of the topics on average which shows the importance of uni-grams when it comes to selection of topical tweet.
- From and mention features are the least informative features for all of the topics.
- 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 HumanCausedDisaster, LBGT, and Soccer topics.
- Sorting features based on their average mean value across different topics results in the following order:

1. Term
2. Location
3. Hashtag
4. Mention
5. From

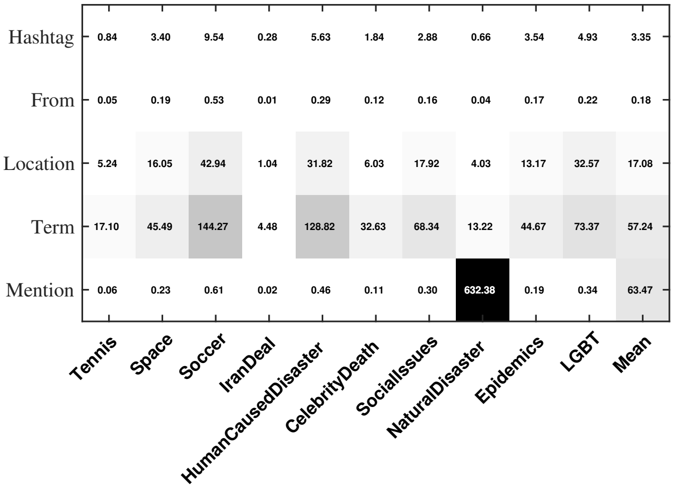


Figure 2: Average MI for different features vs. Topics, last two column show mean value and stderr across all topics

It is important to note that due to very large amount of Term features, they were cleaned based on their frequency (having at least frequency value of 100).

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
✓rt @espntennis: shock city. darcis drops rafa in straight sets. first time nadal loses in first rd of a. major in career. #espnwimbledon #w	✗voting mars @30secondstomars @jaredleto @shannonleto @tomofromearth xobest group http://t.co/dlsozvjinf
✓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_com: show everyone how much you are proud of @30secondstomars !#mtvhottest 30 seconds to mars http://t.co/byxnri4t67
✓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 @30secondstomars: missed the big news? mars touring with @linkinpark + special guests @afi this summer!_ http://t.co/3e5rm9pwrd
✓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 @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
HumanCausedDisaster	**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 &amp; 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&amp;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. &amp; 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 legal protection to cupcake kiplers
✗@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

For each feature type, do any attributes correlate with importance? In order to give a better sense of what features are better for each topic, we provided 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.

scatterplots of feature MI – the absolute last thing we do (density plots?!!) **which plots below, and for which topics? Could pick out most useful features for topics in part (a)(i) and just show selected scatter plots below for these feature types. from, mention MIs vs. followers, favorites, friends, hashtags, tweets hashtag MI vs. #tweets, #users location MI vs. #users term MI vs. #tweets

Related Works

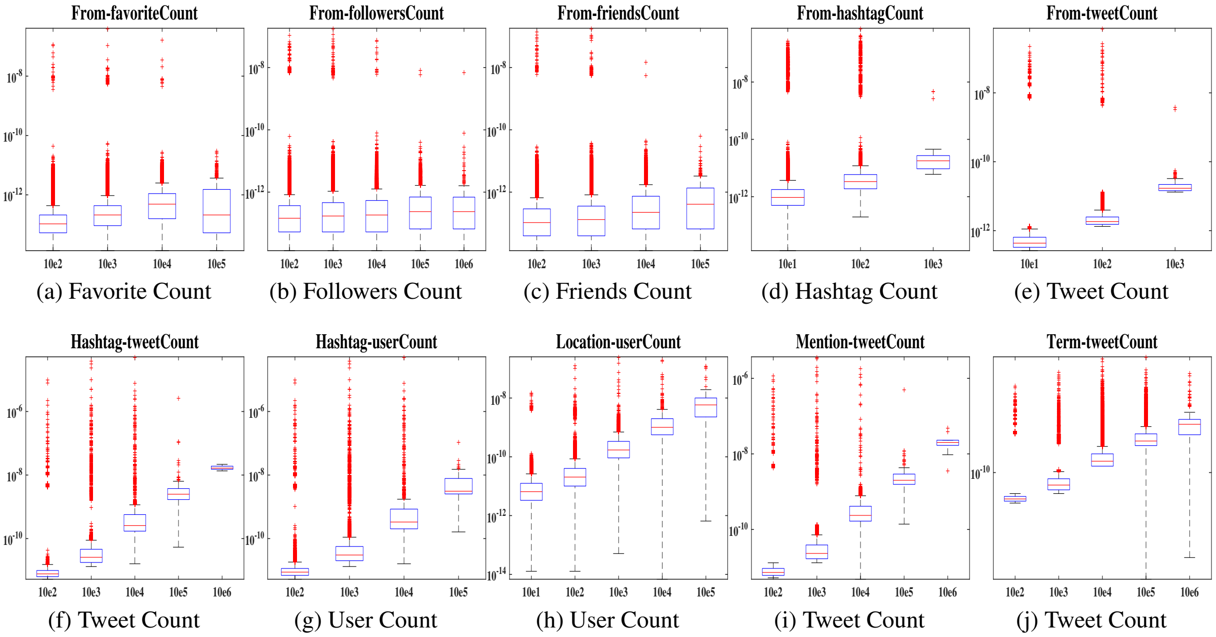
This section provides existing research on the use of social media as a sensor for topic detection on social media. Here, we focus on related research on both events and topics detection from social media since events are special type of topics and can be considered as a topic.

Historically, event detection has been studied extensively in text mining, NLP, and IR to find events from conventional media sources such as news streams (Yang, Pierce, and Carbonell 1998). With the growth of social media sites such as Facebook, Twitter and other microblogs, social media sites have become known as powerful communication tools for sharing and exchanging information about such events. To see how different works address topic detection on social media, we focus on the two highly studied types of topic detections: Trending Topic Detection, Event Detection. First group of works focus on trending topic detection methods. Majority of works on detecting trending topics use bursts as the indicator of events, where a burst is defined as a sudden change in posting rates of some keywords, hashtags, etc. These can further be divided into multiple categories based on how they use bursts to extract the event. First category, clustering-based methods, focus on the hypothesis that trends are topical and topics are defined by collection of relevant content, hence trends can be detected by clustering content. (**?**) proposed a graphical model to discover latent events clustered in the spatial, temporal and lexical dimensions. (**?**) focus on the task of multi-label classification of tweets into living aspects such as eating.(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). Second category, term-based methods focus on the hypothesis that topics can be detected by focusing on temporal patterns of terms/keywords independent of contents of documents (Mathioudakis and Koudas 2010; Cui et al. 2012; Zhao et al. 2011; Nichols, Mahmud, and Drews 2012). Third category, query-based methods, focus on the hypothesis that trending topics can be detected by measuring user-defined criteria (Albakour, Macdonald, and Ounis 2013; Sakaki et al. 2012). Last work, network Structure-based method, focused on the hypothesis that trending topics can be detected by studying the network structure of users (Budak, Agrawal, and El Abbadi 2011).

However, trending topics detection methods are not targeted. Our method differs from trending topic detection methods in the sense that we are focusing on a set of topics that can not necessarily be detected using burst.

Second groups of works focus on detection of a specific targeted topic such as disaster or epidemic. Regarding disaster, predictive studies (Kryvasheyeu et al. 2014), studied the network of users and focused on choosing the best groups of users in order to achieve lead-times i.e. faster detection of disastrous event (following the concept of "friendship paradox"[3]). (Sakaki, Okazaki, and Matsuo 2013) used SVM classifier for detecting earthquakes and employed location estimation method such as Kalman Filtering for localizing it. Sakaki, Okazaki, and Matsuo extracted statistical features e.g., the number and position of words in a tweet, keyword features and word context features. These studies investigated the real-time nature of Twitter and provided promising results. However another set of works focused on descriptive studies on disaster by discussing the

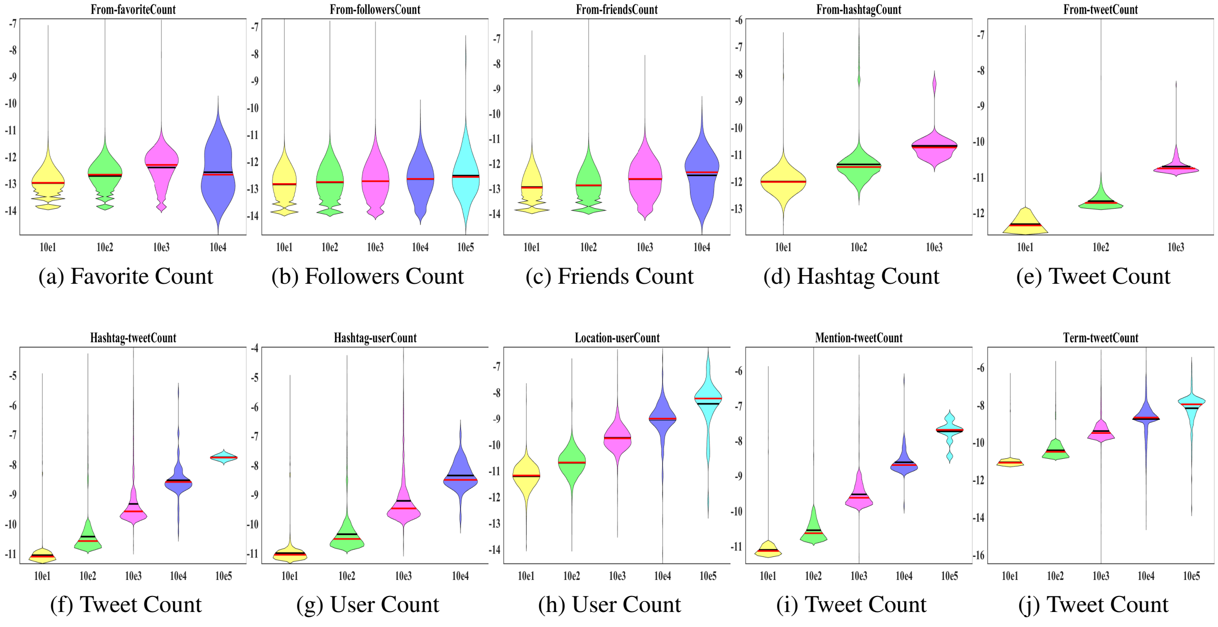
[3] On average, most people have fewer friends than their friends have



(a) Favorite Count (b) Followers Count (c) Friends Count (d) Hashtag Count (e) Tweet Count

(f) Tweet Count (g) User Count (h) User Count (i) Tweet Count (j) Tweet Count

Figure 3: Box Plots 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.



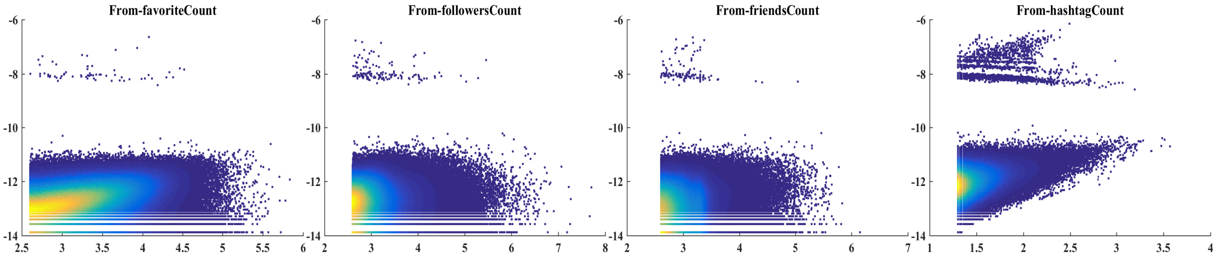
(a) Favorite Count (b) Followers Count (c) Friends Count (d) Hashtag Count (e) Tweet Count

(f) Tweet Count (g) User Count (h) User Count (i) Tweet Count (j) Tweet Count

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.

Topics/Top10	NaturalDisaster	Epidemics	IranDeal	SocialIssues	LBGT	HumanCausedDisaster	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



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

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

behavior of Twitter users during crisis (Vieweg et al. 2010; Cheong and Cheong 2011; Starbird and Palen 2010) and do not address exploiting detection of crisis events. They investigated the use of social media during crisis in order to identify information propagation properties, social behavior of users e.g. retweeting behavior, information contributing to situational awareness, and active players in communicating information. However, this behavioral information could be exploited in development of sensors.

Regarding health epidemic detection, researchers used content-based method and/or structure-based methods. Content-based methods, (Culotta 2010) and (Aramaki, Maskawa, and Morita 2011) both tried to identify influenza-related tweets and find correlations of these tweets to CDC statistics. Both works extracted bag-of-words as features. As for methodology, the former used single and multiple linear regression showing that multiple linear regression works better, while the latter employed SVM. Results showed high correlation of their estimation of influenza in early stages with values from U.S CDC and Japan's Infection Disease Surveillance Center. Structure-based method, (García-Herranz et al. 2012) use the friendship paradox concept (Feld 1991) for early detection of contagious outbreaks. They provided a method for choosing sensor groups from friends of random sets of users to find more central individuals in order to enforce early detection. They claim that this sensor group represents more central individuals and individuals at the center of a network are likely to receive a contagion sooner than randomly-chosen members of the population (because central individuals are a smaller number of steps away from the average individual in the network). As a result, (García-Herranz et al. 2012) argued that this selection process of sensor groups helps in early detection of outbreaks.

On the other hand, hybrid methods (Sadilek, Kautz, and Silenzio 2012) ,exploited both content of tweets and structural information of users network. They employed a semi-supervised approach to learn a SVM classifier using n-grams as features in order to detect ill individuals. Then, they estimated physical interaction between healthy and sick people based on co-location and friendship. This enabled them to study the effect of these two factors of social activity (co-location for contact network and friendship for social ties) on public health.

These method focus on finding a specific topic, thus using a very primitive method for curating the data e.g., querying keyword "earthquake". In addition, there is no discussion on how can these methods be generalized for other topics.

In a more similar settings, (Krestel et al. 2015) compared four different methods of language model, topic model, logistic regression and boosting to evaluate recommended tweets for a given news article on a dataset of 55k news articles and 121k tweets. (Yan, Lapata, and Li 2012; Chen et al. 2012) focus on tweet recommendation based on user preferences and tweet popularity, thus focusing on user's own tweet history, retweet history and social relations between users as features.

Conclusions

Acknowledgments

Copyright

References

[Albakour, Macdonald, and Ounis 2013] Albakour, M.-D.; Macdonald, C.; and Ounis, I. 2013. Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13.

[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 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.

[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.

[Cheong and Cheong 2011] Cheong, F., and Cheong, C. 2011. Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods. In *Pacific Asia Conference on Information Systems, PACIS 2011: Quality Research in Pacific Asia, Brisbane, Queensland, Australia, 7-11 July 2011*, 46.

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

[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 International 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.

[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 et al. 2012] Sakaki, T.; Matsuo, Y.; Yanagihara, T.; Chandrasiri, N.; and Nawa, K. 2012. Real-time event extraction for driving information from social sensors. In *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012 IEEE International Conference on*, 221–226.

[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.

[Starbird and Palen 2010] Starbird, K., and Palen, L. 2010. *Pass it on?: Retweeting in mass emergency*. International Community on Information Systems for Crisis Response and Management.

[Vieweg et al. 2010] Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010*, 1079–1088.

[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*.

[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.

[Yang, Pierce, and Carbonell 1998] Yang, Y.; Pierce, T.; and Carbonell, J. G. 1998. A study of retrospective and on-line event detection. In *SIGIR '98: Proceedings of the 21st Annual International (ACM) (SIGIR) Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, 28–36.

[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.