A Longitudinal Study of Topic Classification on Twitter

Zahra Iman[†], Scott Sanner[‡], Mohamed Reda Bouadjenek[*], Lexing Xie[§]
[†]Oregon State University, Corvallis, OR, USA
[‡]University of Toronto, Toronto, ON, Canada
[*]University of Melbourne, Melbourne, VIC, Australia
[§]Australian National University and Data61, Canberra, ACT, Australia
zahra.iman87@gmail.com, ssanner@mie.utoronto.ca, reda.bouadjenek@unimelb.edu.au, lexing.xie@anu.edu.au

Abstract

Twitter represents a massively distributed information source over a kaleidoscope of topics ranging from social and political events to entertainment and sports news. While recent work has suggested that variations on standard classifiers can be effectively trained as topical filters (Lin, Snow, and Morgan 2011; Yang et al. 2014; Magdy and Elsayed 2014), there remain many open questions about the efficacy of such classification-based filtering approaches. For example, over a year or more after training, how well do such classifiers generalize to future novel topical content, and are such results stable across a range of topics? Furthermore, what features and feature classes are most critical for long-term classifier performance? To answer these questions, we collected a corpus of over 800 million English Tweets via the Twitter streaming API during 2013 and 2014 and learned topic classifiers for 10 diverse themes ranging from social issues to celebrity deaths to the "Iran nuclear deal". The results of this long-term study of topic classifier performance provide a number of important insights, among them that (1) such classifiers can indeed generalize to novel topical content with high precision over a year or more after training and (2) simple terms and locations are the most informative feature classes (despite training on classes labeled via hashtags).

Learning Topical Social Sensors

Our objective is to evaluate binary classifiers that can label a previously unseen tweet as topical (or not). Following the approach of (Lin, Snow, and Morgan 2011), for a topic t, we leverage a (small) set of user-curated topical hashtags H^t to efficiently provide a large number of supervised topic labels for training. As standard for machine learning methods, we divide our training data into train and validation sets — the latter for hyperparameter tuning to control overfitting and ensure generalization to unseen data. As a critical insight for topical generalization where we view correct classification of tweets with *previously unseen topical hashtags* as a proxy for topical generalization, we *do not* simply split our data temporally into train, validation, and test sets and label both with *all* hashtags in H^t. *Instead*, we split H^t into three disjoint sets H^t_{train}, H^t_{val}, and H^t_{test} according to two time stamps $t^{\text{val}}_{\text{split}}$ and $t^{\text{test}}_{\text{split}}$ for topic t and the first usage time stamp $h_{\text{time}*}$ of each hashtag $h \in H^t$. In short, all hashtags $h \in H^t$ with $h_{\text{time}*} < t^{\text{val}}_{\text{split}}$ are used to generate positive labels in the training data, those with $h_{\text{time}*} \geq t^{\text{test}}_{\text{split}}$ are used for positive labels in the test data and the remainder are used for positive labels in the validation data.

The key point to observe is that we not only partition the train, validation, and test data temporally, but we also divide the hashtag class labels temporally and label each data partition with an entirely disjoint set of topical hashtags. The purpose behind this training and validation data split and labeling is to ensure that learning hyperparameters are tuned so as to prevent overfitting and maximize generalization to unseen topical content (i.e., new hashtags). We remark that a classifier that simply memorizes training hashtags will fail to correctly classify the validation data except in cases where a tweet contains both a training and validation hashtag.

#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	Most frequent
From	10,196	8.67	2	running_status
Hashtag	1,653,159	13.91	1	#retweet
Mention	6,291	1.26	1	tweet_all_time
Location	10,848,224	9,562.34	130	london
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	6,616.65	1	rt

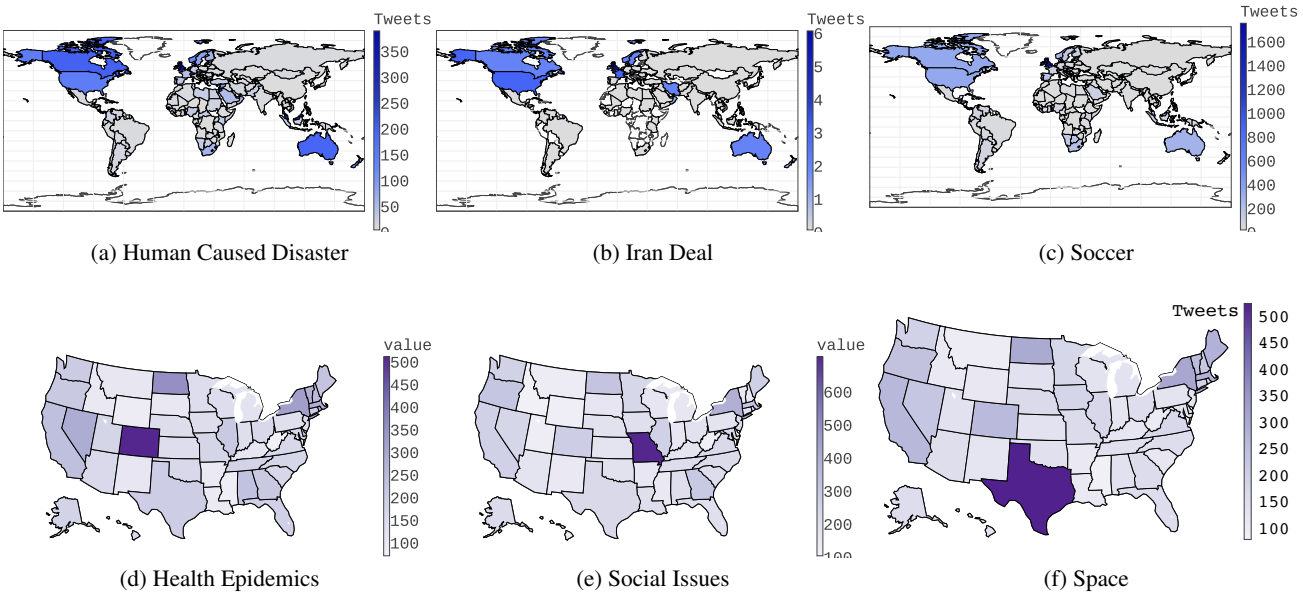
Feature Using #Hashtags

From	18,167	2	0	daily_astrodata
Location	2,440,969	1,837.79	21	uk

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

Data Description

We crawled Twitter data using the Twitter Streaming API for two years spanning 2013 and 2014. We collected more than 2.5 TB of compressed data, which contains a total of



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

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

Figure 1: Tweets per 1 Million capita tweet frequency across different international and U.S. locations for different topics.

$829,026,458$ English tweets. 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* (i.e., a topical keyword specified using the # sign), *Mention* (i.e., a Twitter username reference using the @ sign), *Term* (i.e., any non-hashtag and non-mention unigrams). We provide detailed feature statistics in Table 1.

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, U.S., and Europe for nuclear negotiations the "Iran deal", and soccer for some (English-speaking) countries where it is popular. For U.S. states, we see that Colorado stands out for health epidemics (whooping cough and pneumonic plague occurred in the data collection period), Missouri stands out for social issues (#blacklivesmatter in St. Louis), and Texas stands out for space due to NASA's presence there.

Empirical Evaluation

With the formal definition of learning topical classifiers provided in Sec. and the overview of our data in Sec. , we proceed to outline our experimental methodology on our Twitter corpus. We manually curated a broad thematic range of 10 topics shown in the top row of Table 2 by annotating hashtag sets H^t for each topic t. We used 4 independent annotators to query the Twitter search API to identify candidate hashtags for each topic, requiring an inner-annotator agreement of 3 annotators to permit a hashtag to be assigned to a topic set. Per topic, hashtags were split into train and test sets according to their first usage time stamp roughly according to a 3/5 to 2/5 proportion (the test interval spanned between 9-14 months). The training hashtag set was further temporally subdivided into train and validation hashtag sets according to a 5/6 to 1/6 proportion. We show a variety of statistics and five sample hashtags per topic in Table 2. Here we can see that different topics had varying prevalence in the data with *Soccer* being the most tweeted topic and *IranDeal* being the least tweeted according to our curated hashtags.

As noted in Sec. , positively occurring features may include *From*, *Mention*, *Location*, *Term*, and *Hashtag* features. Because we have a total of $538,365,507$ unique features in our Twitter corpus, it is critical to pare this down to a size that is robust to overfitting and amenable for efficient learning. To this end, we thresholded all features according to the frequencies listed in Table 3. The rationale in our thresholding was initially that all features should have the same frequency cutoff in order to achieve roughtly 1 million features. However, in initial experimentation, we found that a high threshold pruned a large number of informative terms and locations. To this end, we lowered the threshold for terms and locations noting that even at these adjusted thresholds, we still have more authors than terms. We also removed common English stopwords which further reduced the unique term count. Overall, we end up with $1,166,582$ candidate features (*CF*) for learning topical classifiers.

Supervised Learning Algorithms

With our labeled training and validation datasets defined in Sec. and our candidate feature set *CF* defined previously, we proceed to apply different probabilistic classification and ranking algorithms for learning topical classifiers as defined in Sec. . In this paper, we experiment with the following four classifiers or rankers:

	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	#earthquake	#ebola	#loveislove
	#novakdjokovic	#astronauts	#lovesoccer	#iranfreedom	#childrenofsyria	#ripmandela	#michaelbrown	#storm	#virus	#gaypride
	#wimbledon	#satellite	#fifa	#irantalk	#iraqwar	#ripjoanrivers	#justice4all	#tsunami	#vaccine	#uniteblue
	#womenstennis	#spacecraft	#realmadrid	#rouhani	#bombthreat	#mandela	#freetheweed	#abfloods	#chickenpox	#homo
	#tennisnews	#telescope	#beckham	#nuclearpower	#isis	#paulwalker	#newnjgunlaw	#hurricanekatrina	#theplague	#gaymarriage

Table 2: Test/Train Hashtag samples and statistics.

	Threshold	#Unique Values
From	159	361,789
Hashtag	159	184,702
Mention	159	244,478
Location	50	57,767
Term	50	317,846
Features (CF)	-	1,166,582

Table 3: Cutoff threshold and corresponding number of unique values of candidate features *CF* for learning.

1. **Logistic Regression** using LibLinear (Fan et al. 2008)
2. **Bernoulli Naïve Bayes** (McCallum and Nigam 1998)
3. **Rocchio** (Manning, Raghavan, and Schütze 2008) (a centroid-based classifier)
4. **RankSVM** (Lee and Lin 2014)

As noted in Sec , tuning of hyperparameters on validation data is critical. In our experiments, we tune as follows:

- *Logistic Regression and RankSVM*: L_2 regularization constant C for both methods is tuned for $C \in \{10^{-12}, 10^{-11}, ..., 10^{11}, 10^{12}\}$.
- *Naïve Bayes*: Dirichlet prior α is tuned for $\alpha \in \{10^{-20}, 10^{-15}, 10^{-8}, 10^{-3}, 10^{-1}, 1\}$.
- *All Classfiers*: The number of top features M selected based on their Mutual Information is tuned for $M \in \{10^2, 10^3, 10^4, 10^5, 1166582 \text{ (all features) }\}$.

We remark that many algorithms such as Naive Bayes and Rocchio performed better with feature selection and hence we used feature selection for all algorithms (nb., it is possible to select all features). We tune the hyperparameters via exhaustive grid search and select the configuration with the highest Average Precision (AP) (Manning, Raghavan, and Schütze 2008) ranking metric discussed more below.

Performance Analysis

While our training data is provided as supervised class labels, we remark that topical classifiers are targeted towards individual users who will naturally be inclined to *examine only the highest ranked tweets*. Hence we believe ranking metrics represent the best performance measures for the intended use case of this work. While RankSVM naturally produces a ranking, all classifiers are score-based, which also allows them to provide a natural ranking of the test data that we evaluate via the following ranking metrics:

- **AP:** Average precision over the ranked list; the mean over all topics provides mean AP (mAP).

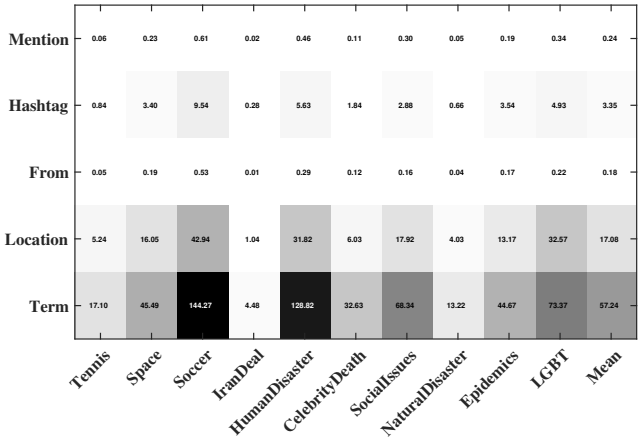


Figure 2: Matrix of mean Mutual Information values (divided by $1E + 10$ for different feature types vs. topics. The last column is the average of mean values across all topics.)

- **P@k:** Precision at k for $k \in \{10, 100, 1000\}$.

While P@10 may be a more standard retrieval metric for tasks such as ad-hoc web search, we remark that the short length of tweets relative to web documents makes it more plausible to look at a much larger number of tweets, hence the reason for also evaluating P@100 and P@1000.

Table 4 evaluates these metrics for each topic. *Logistic Regression* is the best performing method on average except for $P@10$. We conjecture the reason is that *Naïve Bayes* tends to select fewer features for training, which allows it to achieve higher precision over the top-10 at the expense of lower $P@100$ and $P@1000$. These results suggest that in general both *Logistic Regression* and *Naïve Bayes* make for effective topical learners and generalize to new unseen topics up to a year after training. Also notable is that trained classifiers outperform RankSVM on the ranking task thus justifying the use of trained topic classifiers for ranking.

Feature Analysis

We now analyze the informativeness of our defined features in Sec and the effect of their attributes on learning targeted topical classifiers. We use Mutual Information (MI) (Manning, Raghavan, and Schütze 2008) as our primary metric for feature evaluation, where higher values for this metric indicate more informative features for the given topic.

We provide the mean Mutual Information values for each feature across different topics in Fig. 2. The last column in Fig. 2 shows the average of the mean Mutual Information

		Tennis	Space	Soccer	IranDeal	HumanDisaster	CelebrityDeath	SocialIssues	NaturalDisaster	Epidemics	LGBT	Mean
LR	**AP**	**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	**AP**	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	**AP**	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	**AP**	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 4: Performance of algorithms across metrics (best in bold) and topics with mean performance over all topics at right.

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: The top 5 features for each feature type and topic based on Mutual Information.

for each feature type. We observe the following:

- The *Term* and *Location* features are the most informative features on average (despite class labels being *Hashtags*).
- The Location feature provides the highest MI regarding the topics of *HumanDisaster*, *LBGT*, and *Soccer* indicating that the content in these topics is heavily localized.
- Looking at the overall average values, the order of informativeness of feature types appears to be the following: *Term*, *Location*, *Hashtag*, *Mention*, *From*.

As anecdotal evidence to support these conclusions and provide additional insights regarding the informativeness of each feature type, we refer to Table 5, which displays the top five feature instances for each feature type and topic. Among many remarkable insights in this table, one key aspect we note is that the *terms appear to be the most generic* (and hence most generalizable) features, providing strong intuition as to why these features are informative over the two year time span of our data. The top *locations are also highly relevant to most topics* indicating the overall importance of these tweet features for identifying topical tweets.

Acknowledgments

We thank Dan Nguyen for his help with data processing.

References

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.

Lee, C.-P., and Lin, C.-J. 2014. Large-scale linear RankSVM. *Neural Computing* 26(4):781–817.

Lin, J.; Snow, R.; and Morgan, W. 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *KDD*, 422–429.

Magdy, W., and Elsayed, T. 2014. Adaptive method for following dynamic topics on twitter. In *ICWSM*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge Univ. Press.

McCallum, A., and Nigam, K. 1998. A comparison of event models for naive bayes text classification. In *In AAAI-98 Workshop On Learning For Text Categorization*, 41–48.

Yang, S.-H.; Kolcz, A.; Schlaikjer, A.; and Gupta, P. 2014. Large-scale high-precision topic modeling on twitter. In *KDD*, 1907–1916.