A Longitudinal Study of Topic Classification on Twitter

Zahra Iman[1], Scott Sanner[2], Mohamed Reda Bouadjenek[2], Lexing Xie[3], and Daniel Xiaoliang Shi[2]

[1]**Oregon State University, Corvallis, OR, USA**
[2]**The University of Toronto, Toronto, ON, Canada**
[3]**Australian National University and Data61, Canberra, ACT, Australia**

ABSTRACT

Twitter represents a massively distributed information source over topics ranging from social and political events to entertainment and sports news. While recent work has suggested this content can be narrowed down to the personalized interests of individual users by training standard classifiers as topical filters, 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? In addition, how robust is a topic classifier over the time horizon, e.g., can a model trained in 2010 be used for making predictions in 2019? How to allow proper generalization over the time horizon and avoid overfitting that may occur by learning inappropriate feature weights related to specific events? Furthermore, what features, feature classes, and feature attributes 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: (i) such classifiers can indeed generalize to novel topical content with high precision over a year or more after training, (ii) simple terms and hashtags are the most informative feature classes, (iii) removing tweets containing training hashtags from the validation set allows a better generalization, (iv) the performance of classifiers drops overtime, and (v) 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 long-term study of topic classifiers on Twitter that further justifies classification-based topical filtering approaches while providing detailed insight into the feature properties most critical for topic classifier performance.

INTRODUCTION

With the emergence of the social Web in the mid-2000s, the Web has evolved from a static Web, where users were only able to consume information, to a Web where users are also able to interact and produce information. This evolution which is commonly known as Social Web has introduced new freedoms for the user in his relation with the Web by facilitating his interactions with other users who have similar tastes or share similar resources. These interactions result in a massive quantity of data that has to be leveraged for developing various Data-driven Decision-making applications.

In this context, 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 chatter and otherwise low-value content for most users' information needs where filtering out irrelevant content is extremely time-consuming. Previous work Lin et al. (2011); Yang et al. (2014); Magdy and Elsayed (2014) has noted this need for topic-based filtering and has adopted a range of variations on supervised classification techniques to build effective topic filters.

While these previous approaches have augmented their respective topical classifiers with extensions ranging from semi-supervised training to multiple stages of classification-based filtering to online tracking of foreground and background language model evolution, we seek to analyze the lowest common denominator of all of these methods, namely the performance of the underlying (vanilla) supervised

classification paradigm. Our fundamental research questions in this paper are hence focused on a longitudinal study of the performance of such supervised topic classifiers. 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? In addition, how robust is a topic classifier over the time horizon, e.g., can a model trained in 2010 be used for making predictions in 2019? How to allow proper generalization over the time horizon and avoid overfitting that may occur by learning inappropriate feature weights related to specific events? Furthermore, what features, feature classes, and feature attributes 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". We leverage ideas from Lin et al. (2011) for curating hashtags to define our 10 training topics and label tweets for supervised training; however, we also curate disjoint hashtag sets for validation and test to tune hyperparameters and test true generalization performance of the topic filters to future novel content.

The main outcome of this work can be summarized as follows:

- 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 average precision (AP) and Precision@n (for a range of n) evaluated over long time-spans of typically one year or more.

- Also, we demonstrate that the performance of classifiers tends to drop over time – roughly 35% drop in average precision after 350 days of training, which is a significant decrease. We impute this to the fact that over long periods of time, features that are predictive during the training period may prove ephemeral and fail to generalize to prediction at future times.

- To address the problem above, we show that one can remove tweets containing training hashtags from the validation set such that to allow a better generalization and avoid overfitting. Indeed, although our approach here is simple, it allows to get roughly 11% improvement for average precision, thus avoiding overfitting to some extend.

- 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.

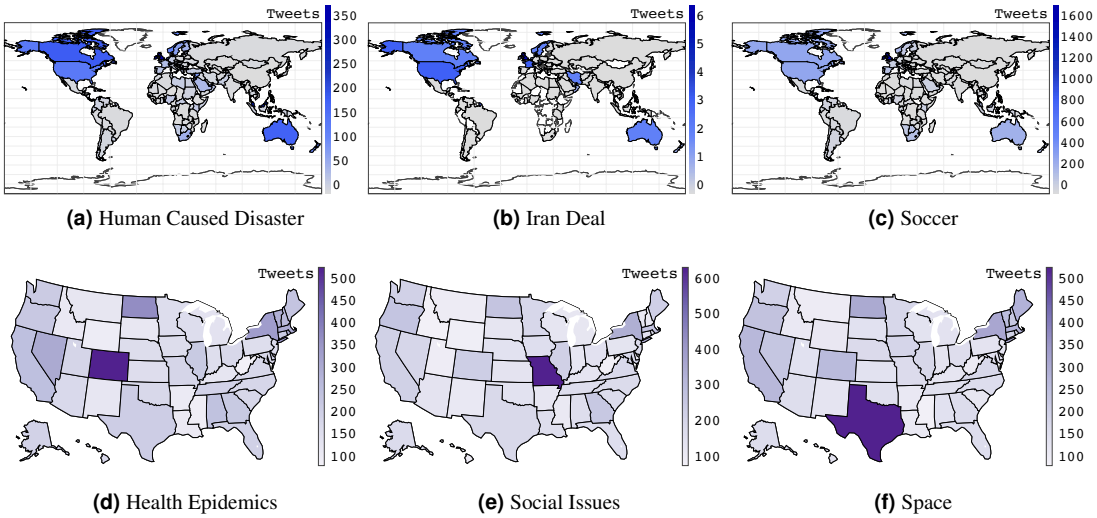
In summary, this work[1] provides a longitudinal study of Twitter topic classifiers that further justifies supervised approaches used in existing work while providing detailed insight into feature properties critical for their performance. The rest of this paper is organized as follows: we first describe the notation we use in this paper and provide a formal definition of the problem we address. Then, we provide a description of the dataset we used for the analysis carried out in this paper, followed by a description of the general methodology for learning topic classifiers over time. The next section gives a thorough description of the results obtained and the conclusions drawn. Finally, we review the literature before concluding and describing the future work in the last section.

NOTATION AND PROBLEM DEFINITION

Our objective in this paper is to carry out a longitudinal study of topic classifiers for Twitter. For each Twitter topic, we seek to build a binary classifier that can label a previously unseen tweet as topical (or not). To achieve this, we train and evaluate the classifier on a set of topically labeled historical tweets as described below.

Formally, given an arbitrary tweet d (a document in text classification parlance) and a set of topics $T = \{t_1, \ldots, t_K\}$, we wish to train a scoring function $f^t : D \to \mathbb{R}$ for each topic $t \in T$ over a subset of labeled training tweets from $D = \{d_1, \ldots, d_N\}$. We assume that each tweet $d_i \in D$ is represented by a vector of m features $d_i = [d_i^1, \ldots, d_i^M]$ with $d_i^M \in \{0,1\}$ to indicate that the feature M is associated with d_i (1) or not (0), and its associated label $t(d_i) \in \{0,1\}$ to indicate whether the tweet d_i is topical (1) or not

[1] This is an extended and revised version of a preliminary conference report that was presented in Iman et al. (2017).



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

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

Figure 1. Per capita tweet frequency across different international and U.S. locations for different topics. The legend provides the number of tweets per 1 Million capita.

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

#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

(0). As in any standard classification task, we wish to learn the mapping function $f^t(d_i)$ so that it can be used to predict the label of a new unseen tweet d_i with a high accuracy.

DATA DESCRIPTION

We begin with details of the Twitter testbed for topical classifier learning that we evaluate in this paper. We crawled Twitter data using Twitter Streaming API for two years spanning 2013 and 2014 years. We collected more than 2.5 TB of compressed data, which contains a total number of 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*: a topical keyword specified using the # sign.
- *Mention*: a Twitter username reference using the @ sign.

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

We provide more detailed statistics about each feature in Table 1. For example, there are over 11 million unique hashtags, the most frequent unique hashtag occurred in over 1.6 million tweets, a hashtag has been used on average by 10.08 unique users, and authors (*From* users) have used a median value of 2 tweets.

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 (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.

METHODOLOGY

In this section, we describe the formal framework we use for our longitudinal study of topical classifiers. Hence, we first describe how we propose to label the data using a set of hand-curated user hashtags. Then, we proceed to describe the way we propose to split the dataset into train, validation and test sets, which is critical for such a longitudinal study of topical classifiers. Finally, we provide a brief description of several classification algorithms we use in our analysis.

Dataset labelling

A critical bottleneck for learning targeted topical social classifiers 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, while 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 Internet Relay Chat (IRC), were adopted later (and perhaps most famously) on Twitter, and now appear on other social media platforms such as Instagram, Tumblr, and Facebook. Following the approach of Lin et al. (2011), for each topic $t \in T$, we leverage a (small) set of user hand-curated topical hashtags H^t to efficiently label a large number of supervised topic labels for social media content.

Specifically, 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 $t \in T$. 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.

Dataset splitting

We now provide a procedure for labeling data with H^t for training, validation and test. Following this, we proceed to train supervised classification and ranking methods to learn topical content from a large feature space – this feature space includes terms, hashtags, mentions, authors and their locations. In the following, we describe three key points related to the temporal splitting of the dataset, sampling negative examples, and hyper-parameter tuning.

Temporally split for train, validation and test using H^t: As standard for machine learning methods, we divide our training data into train, validation, and test sets — the validation set is used 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 and test sets and label both with *all* hashtags in H^t. Rather, we split each H^t into three disjoint sets H^t_{train}, H^t_{val}, and H^t_{test} according to two time stamps $t^{\text{train}}_{\text{split}}$ and $t^{\text{val}}_{\text{split}}$ for topic t and the first usage time stamp $h_{\text{time}*}$ of each hashtag $h \in H^t$. In short, all hashtags $h \in H^t$ first used before $t^{\text{train}}_{\text{split}}$ are used to generate positive labels in the training data, all hashtags $h \in H^t$ first used after $t^{\text{train}}_{\text{split}}$ and before $t^{\text{val}}_{\text{split}}$ are used to generate positive labels in the validation data, and the remaining hashtags are used to generate positive labels in the test data.

Table 2. Train/Validation/Test Hashtag samples and statistics.

	Tennis	Space	Soccer	Iran Deal	Human Disaster	Celebrity Death	Social Issues	Natural Disaster	Epidemics	LGBT
#TrainHashtags	62	112	144	12	57	33	37	61	55	30
#ValHashtags	14	32	42	2	8	4	5	4	17	9
#TestHashtags	14	17	21	3	12	7	8	17	13	5
#+TrainTweets	21,716	5,333	14,006	6,077	153,612	155,121	27,423	46,432	14,177	1,344
#-TrainTweets	191,905	46,587	123,073	54,045	1,363,260	1,376,872	244,106	411,609	125,092	11,915
#+ValTweets	884	2,281	4,073	1,261	53,340	23,710	3,088	843	4,348	50
#-ValTweets	7,860	20,368	36,341	11,363	473,791	210,484	27,598	7,456	39,042	443
#+TestTweets	1,510	5,908	11,503	368	34,055	7,334	14,566	5,240	3,105	692
#-TestTweets	13,746	53,348	103,496	3,256	305,662	65,615	130,118	47,208	27,828	6,325
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

To achieve this effect formally, we define the following:

$$H^t_{\text{train}} = \{h | h \in H^t \wedge h_{\text{time}*} < t^{\text{train}}_{\text{split}}\},$$
$$H^t_{\text{val}} = \{h | h \in H^t \wedge h_{\text{time}*} \geq t^{\text{train}}_{\text{split}} \wedge h_{\text{time}*} < t^{\text{val}}_{\text{split}}\},$$
$$H^t_{\text{test}} = \{h | h \in H^t \wedge h_{\text{time}*} \geq t^{\text{val}}_{\text{split}}\}.$$

Once we have split our hashtags into training and validation sets according to $t^{\text{train}}_{\text{split}}$ and $t^{\text{val}}_{\text{split}}$, we next proceed to temporally split our training documents D into a training set D^t_{train}, a validation set D^t_{val}, and a test set D^t_{test} for topic t based on the posting time stamp $d_{i,\text{time}*}$ of each tweet d_i as follows:

$$D^t_{\text{train}} = \{d_i | d_i \in D \wedge d_{i,\text{time}*} < t^{\text{train}}_{\text{split}}\},$$
$$D^t_{\text{val}} = \{d_i | d_i \in D \wedge d_{i,\text{time}*} \geq t^{\text{train}}_{\text{split}} \wedge d_{i,\text{time}*} < t^{\text{val}}_{\text{split}}\},$$
$$D^t_{\text{test}} = \{d_i | d_i \in D \wedge d_{i,\text{time}*} \geq t^{\text{train}}_{\text{val}}\}.$$

Then, to label the train, validation, and test data sets D^t_{train}, D^t_{val} and D^t_{test}, we use the respective hashtag sets H^t_{train}, H^t_{val}, H^t_{test} for generating the topic label for a particular tweet $t(d_i) \in \{0,1\}$ as follows:

$$t(d_i) = \begin{cases} 1 \text{ if } \exists\, h \in d_i \wedge h \in H^t \\ 0 \text{ otherwise} \end{cases}.$$

The critical insight here is that we do not only divide the train, validation, and test temporally, but we also divide the hashtag labels temporally and label the validation and test data with an entirely disjoint set of topical labels from the training data. The purpose behind this training, validation and test data split and labeling is to ensure that learning hyper-parameters 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.

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 train 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.

Sampling negative examples: Topic classification is very often an unbalanced classification task, since usually, there are much more negative examples than positive examples. Indeed, the large number of users on twitter, their diversity, their wide range interests, and the short lifetime of topics discussed on a daily basis make that diverse topics are discussed on a short time period, thus having only a small set of positive examples for each topic. For example, the open directory project (ODP also known as DMOZ), which provides a hierarchically classification structure of roughly 5M Web ressources over 1M topics[2].

[2]http://www.geniac.net/odp/

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

	Threshold	#Unique Values
From	235	206,084
Hashtag	65	201,204
Mention	230	200,051
Location	160	205,884
Term	200	204,712
Features (CF)	-	1,017,935

Therefore, given the huge amount of negative examples in a real classification scenario (e.g., roughly 800 million negative examples against 500k positive examples for the human disaster topic), to efficiently tune each classifier, we have chosen to sample negative examples such that positive examples represent 10% of the dataset and the negative examples represent 90% of the dataset. This rule is valid for the training, validation and test sets of each topic.

Training and hyper-parameter tuning: Once D^t_{train} and D^t_{val} have been constructed, we proceed to train our scoring function f^t on D^t_{train} and select hyperparameters to optimize Average Precision (AP) Manning et al. (2008) (a ranking metric) on D^t_{val}. Once the optimal f^t is found for D^t_{val}, we return it as our final learned topical scoring function f^t for topic t. Because $f^t(d_i) \in \mathbb{R}$ is a scoring function, it can be used to rank.

With train, validation, and testing data defined along with the training methodology, it remains now to extract relevant features, described next.

Empirical Evaluation

Positively occurring features D_i^+ in our d_i 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 amenable for efficient learning and robust to overfitting. 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 and our candidate feature set *CF* now defined, we proceed to apply different probabilistic classification and ranking algorithms to generate a score function f^t for learning topical classifiers as defined in the Methodology section. In this paper, we experiment with the following four state-of-the-art supervised classification and ranking methods:

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

As outlined in the Methodology section, tuning of hyperparameters on a validation dataset is critical. In our experiments, we tune the following hyperparameters:

Table 4. Performance of topical classifier learning algorithms across metrics and topics with the mean performance over all topics shown in the right column. The best performance per metric is shown in bold.

		Tennis	Space	Soccer	Iran Deal	Human Disaster	Celebrity Death	Social Issues	Natural Disaster	Epidemics	LGBT	Mean
LR	**AP**	0.9590	0.6452	0.5036	0.9807	0.6952	0.9293	0.5698	0.9428	0.4005	0.1559	0.6782±0.1724
NB	**AP**	0.5859	0.8471	0.3059	0.9584	0.4224	0.4658	0.5030	0.3518	0.4050	0.1689	0.5014±0.1494
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
RF	**AP**	0.9344	0.9314	0.5509	0.9757	0.6658	0.9571	0.8213	0.8306	0.5154	0.2633	0.7445±0.14764
KNN	**AP**	0.9550	0.7751	0.4739	0.9752	0.598	0.542	0.5078	0.9599	0.5317	0.1774	0.6496± 0.1618
LR	**P@10**	1.0	0.2	0.3	1.0	0.5	0.8	0.2	1.0	0.5	0.6	0.61±0.2012
NB	**P@10**	0.1	0.8	0.0	0.9	0.7	0.1	0.0	0.3	0.1	0.0	0.3±0.2225
RankSVM	**P@10**	**1.0**	0.8	0.6	0.8	0.4	0.3	0.0	0.1	0.0	0.2	0.42±0.26
RF	**P@10**	1.0	0.5	0.5	1.0	0.9	1.0	1.0	1.0	0.7	0.5	0.81 ±0.1444
KNN	**P@10**	1.0	0.0	1.0	1.0	0.7	0.9	0.0	0.9	0.3	0.4	0.62± 0.2543
LR	**P@100**	0.98	0.65	0.44	0.99	0.74	0.94	0.59	0.98	0.45	0.2	0.696±0.1721
NB	**P@100**	0.56	0.95	0.0	0.98	0.39	0.36	0.16	0.37	0.48	0.1	0.435±0.2033
RankSVM	**P@100**	0.73	0.72	0.31	0.70	**0.88**	0.44	0.48	0.34	0.02	0.100	0.472±0.20
RF	**P@100**	0.98	0.94	0.43	0.98	0.62	0.97	0.81	0.9	0.61	0.29	0.753 ± 0.1555
KNN	**P@100**	1.0	0.59	0.34	1.0	0.72	0.54	0.39	0.96	0.54	0.24	0.632 ±0.1731
LR	**P@1000**	0.653	0.703	0.545	0.299	0.666	0.884	0.574	0.919	0.267	0.076	0.5586±0.1682
NB	**P@1000**	0.551	0.667	0.29	0.333	0.338	0.542	0.655	0.287	0.319	0.169	0.4151±0.1073
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
RF	**P@1000**	0.728	0.464	0.576	0.331	0.463	0.914	0.789	0.728	0.397	0.159	0.5549 ± 0.1450
KNN	**P@1000**	0.571	0.821	0.53	0.329	0.476	0.84	0.49	0.929	0.234	0.083	0.5303±0.1696

- *Logistic Regression*: L_2 regularization constant C 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 $M \in \{10^2, 10^3, 10^4, 10^5\}$.
- How about RankSVM parameters?

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 (where it is possible to select all features). Hyperparameter tuning is done via exhaustive grid search using the Average Precision (AP) Manning et al. (2008) ranking metric to select the best scoring function f^t on the validation data. Once found, f^t can be applied to any tweet d_i to provide a score $f^t(d_i)$ used to *rank* tweets in the test data. Code to process the raw Twitter data and to train and evaluate these classifiers as described above is provided on github.[3]

RESULTS AND DISCUSSION

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).
- **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

[3] https://github.com/SocialSensorProject/socialsensor

Table 5. Top tweets for each topic from *Logistic Regression* method results, marked with ✗as irrelevant, ✓as relevant and labeled as topical, and ★as relevant but labeled as non-topical (a false negative).

Tennis	**Space**
✓rt @espntennis: shock city. darcis drops rafa in straight sets. first time nadal loses in first rd of a. major...	✗rt @jaredleto: rt @30secondstomars: icymi: mars performing a cover of @rihanna's #stay on australia's @trip...
✓@ESPNTennis: Shock city. Darcis drops Rafa in straight sets. First time Nadal loses in first rd of a...	✗voting mars @30secondstomars @jaredleto @shannonleto @tomofromearth xobest group http://t.co/dls...
✓@ESPNTennis: Djokovic ousts the last American man standing @Wimbledon, beating Reynolds 7-6...	✗rt @jaredleto_com: show everyone how much you are proud of @30secondstomars !#mtvhottest 30 seconds to...
✓Nadal's a legend. After 3 years; Definitely He's gonna be the best of all the time. Unbelievable perf...	✗rt @30secondstomars: missed the big news? mars touring with @linkinpark + special guests @afi this summer...
✓@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...
Soccer	**IranDeal**
✗rt @tomm_dogg: #thingstodobeforeearthends spend all my money.	✓rt @iran_policy: @vidalquadras:@isjcommittee has investigated 10 major subjects of iranÕs controversial #nuc...
★@mancityonlineco nice performance	✓rt @iran_policy: @vidalquadras:@isjcommittee has investigated 10 major subjects of iranÕs controversial #nuc...
★rt @indykaila: podolski: "let's see what happens in the winter. the fact is that i'm not happy with it, th...	✗rt @negarmortazavi: thank you @hassanrouhani for retweeting. let's hope for a day when no iranian fears retur...
★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...
✗@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. #iranhr...
HumanDisaster	**CelebrityDeath**
✓rt @baselsyrian: there've been peaceful people in #homs not terrorists! #assad,enemy of #humanity...	★rt @sawubona_chris: today is my birthday & also the day my hero @nelsonmandela has died. lets never...
✓what a helpless father, he can do nothing under #assad's siege!#speakup4syrianchildren http://t.co/vg...	★rt @nelsonmandela: N̈death is something inevitable.when a man has done what he considers to be his duty to...
★exclusive: us formally requested #un investigation; russia pressured #assad to no avail;chain of evidence...	★rt @nelsonmandela: la muerte es algo inevitable.cuando un hombre ha hecho lo que considera que es su...
★#save_aleppo from #assadwarcrimes#save_aleppo from #civilians -targeted shelling of #assad regime...	✗#jacques #kallis: a phenomenal cricketing giant of all time - #cricket #history #southafrica http://t.co/ms5p...
✓rt @canine_rights: why does the #un allow this to continue? rt@tintin1957 help raise awareness of the...	✗@sudesh1304 south africa has the most beautiful babies....so diverse,so unique...so god!! lol #durban #southa...
SocialIssues	**NaturalDisaster**
★the us doesn't actually borrow is the thing. i believe in a creationist theory of the us dollar @usanationdebt...	✗us execution in #oklahoma : not cruel and unusual? maybe just barbaric, inhumane and reminiscent of the...
★rt @2anow: according to @njsenatepres women's rights do not include this poor nj mother's right to defend...	✗#haiti #politics - the haiti-dominican crisis - i agree with how martelly is handling the situation: i totally... http...
★rt @2anow: confiscation ? how many carry permits are in the senate and assembly? give us ours or turn ...	★rt @soilhaiti: a new reforestation effort in #haiti. local compost, anyone? http://t.co/xpad0rqbjk @richardbran...
★rt @2anow: vote with your wallet against #guncontrolforest city enterprises does not support the #2a http...	✗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 allow...	✓tony burgener of @swisssolidarity says you can't compare the disaster response in #haiti with the response to...
Epidemics	**LGBT**
✓rt @who: fourteen of the susp. & conf. ebola cases in #conakry, #guinea, are health care workers, of...	★rt @jackmcoldcuts: @lunaticrex @fingersmalloy @toddkincannon @theanonliberal anthony kennedy just wro...
✗@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,6 @who dir. aylward - requires $1 billion ...	✗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...	✗rt @t3h_arch3r: @toddkincannon thanks for your tl having the female realbrother. between them is 600 lbs....
✗augmentation vertigineuse de 57,4% en 1 an des actes islamophobes en france, dit le collectif contre l'is...	✗@toddkincannon who us dick trickle.

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 with *Naïve Bayes* useful for its efficiency compared to its overall performance. *Notably, trained classifiers outperform RankSVM on the ranking task thus justifying the use of trained topic classifiers for ranking.*

To provide more insight into the general performance of our learning topical classifier framework, we provide the top five tweets for each topic according to *Logistic Regression* in Table 5. We've annotated tweets with symbols as follows:

- ✓: the tweet was labeled topical by our test hashtag set.
- ★: the tweet was determined to be topical through manual evaluation even though it did not contain a hashtag in our curated hashtag set (*this corresponds to a false negative due to non-exhaustive labeling of the data*).
- ✗: the tweet was not topical.

In general, we remark that our topical classifier based on logistic regression performs even better than the quantitative results in Table 4 would indicate: many of the highly ranked tweets are false negatives — *they are actually relevant*. Furthermore, even though we use hashtags to label our training, validation, and testing data, our topical classifier has highly (and correctly) ranked topical tweets that *do not contain hashtags*, indicating strong generalization properties from a relatively small set of curated topical hashtags.

Feature Analysis

In this section, we analyze the informativeness of feature sets defined in the Data Description section and the effect of their attributes on learning targeted topical classifiers. To this end, our goal in this section is to answer the following questions:

- What are the best features for learning classifiers and do they differ by topic?
- For each feature type, do any attributes correlate with importance?

To answer these questions, we use Mutual Information (MI) Manning et al. (2008) as our primary metric for feature evaluation. Mutual Information is a general method for measuring the amount of information one random variable contains about another random variable and is used to select predictive features

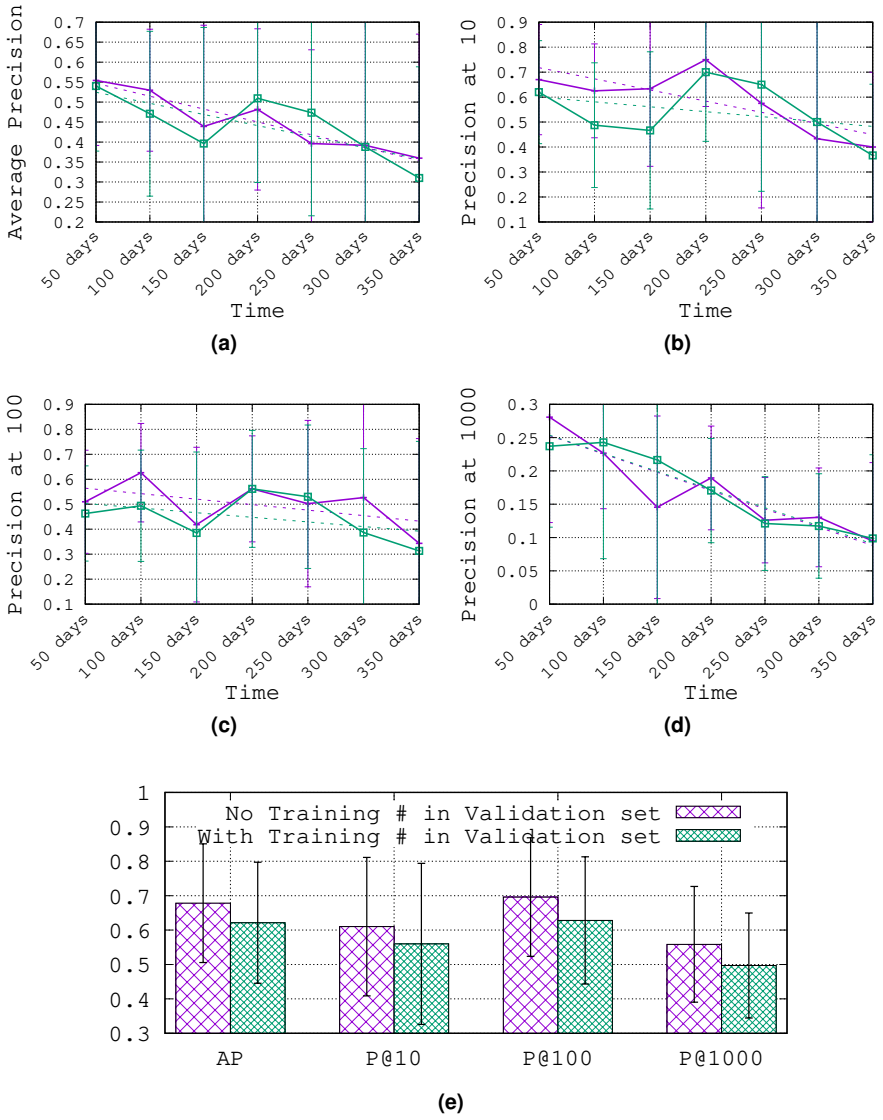


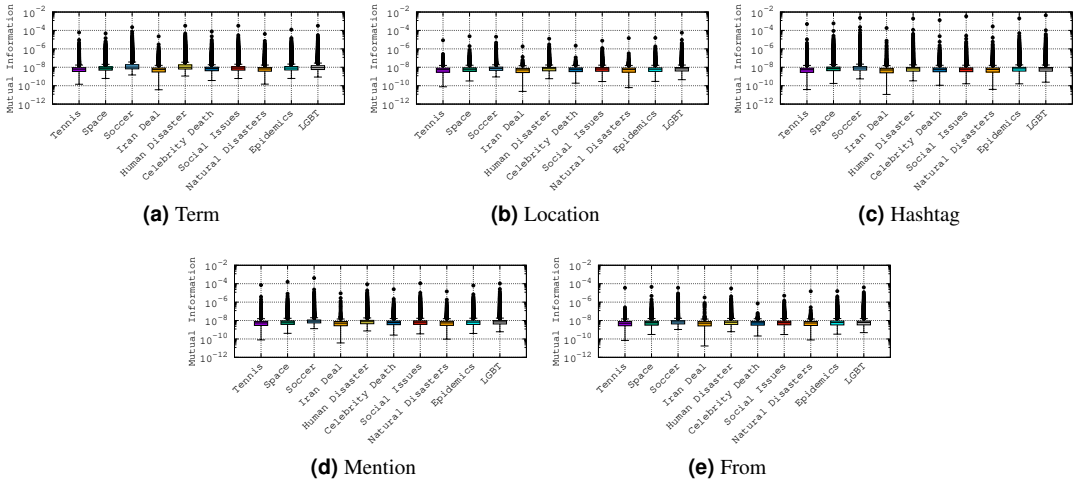
Figure 2. Performance obtained with validation sets with/without tweets containing training hashtags.

Table 6. The top 5 features for each feature type and topic based on Mutual Information.

Topics/Top10	Natural Disaster	Epidemics	Iran Deal	Social Issues	LBGT	Human Disaster	Celebrity Death	Space	Tennis	Soccer
From	from_japan	changedecopine	mazandara	debtadvisoruk	stevendickinson	witfp	boiknox	daily_astrodata	tracktennisnews	makeuphella
From	everyearthquake	stylishoz	freeiran9292	nsingerdebtpaid	mgdauber	ydumozyf	jacanews	freesolarleads	novakdjokovic_i	sport_agent
From	quakestoday	drdaveanddee	hhadi119	negativeequityf	lileensvf1	syriatweeten	ewnreporter	sciencewatchout	i_roger_federer	yasmingoode
From	equakea	soliant_schools	balouchn2	iris_messenger	kevinwhipp	rk70534	rowwsupporter	houston_jobs	andymurrayfans1	sportsroadhouse
From	davewinfields	msgubot	jeffandsimon	dolphin_ls	petermabraham	gosyrianews	flykiidchris	lenautilus	rafaelnadal_fan	losangelesrh
Hashtag	#earthquake	#health	#iran	#ferguson	#tcot	#syria	#rip	#science	#wimbledon	#worldcup
Hashtag	#haiyan	#uniteblue	#irantalks	#mikebrown	#pjnet	#gaza	#ripcorymonteith	#sun	#tennis	#lfc
Hashtag	#storm	#ebola	#iranian	#ericgarner	#p2	#israel	#riprobinwilliams	#houston	#usopen	#football
Hashtag	#PrayForThePhilippines	#healthcare	#rouhani	#blacklivesmatter	#uniteblue	#gazaunderattack	#rippaulwalker	#starwars	#nadal	#worldcup2014
Hashtag	#tornado	#fitness	#irantalksvienna	#icantbreathe	#teaparty	#isis	#robinwilliams	#scifi	#wimbledon2014	#sports
Location	With everyone	USA	France	St Louis MO	USA	Syria	South Africa	Houston TX	Worldwide	Liverpool
Location	Earth	Francophone	Tehran Iran	Washington DC	Bordentown New Jersey	Palestine	Pandaquotescom	Germany	London	Manchester
Location	Philippines	United States	Inside of Iran	St Louis	Global Markets	Syrian Arab Republic	Johannesburg South Africa	Houston	The Midlands	London
Location	Don't follow me am i a bot	Gainesville FL USA	Iran	Virginia US	The blue regime of Maryland	Israel	Johannesburg	Rimouski	London UK	Anfield
Location	Global planet earth	Boulder Colorado	Washington DC	Saint Louis MO	Lancaster county PA	Washington DC	Cape Town	In a galaxy far far ebay	Wimbledon	Bangil East Java Indonesia
Mention	@oxfamgb	@foxtramedia	@ap	@natedrug	@jjauthor	@ifalasteen	@nelsonmandela	@nasa	@wimbledon	@lfc
Mention	@gabriele_corno	@obi_obadike	@afp	@deray	@2anow	@drbasselabuward	@realpaulwalker	@philae2014	@usopen	@fifaworldcup
Mention	@weatherchannel	@who	@iran_policy	@antoniofrench	@gop	@revolutionsyria	@ddlovato	@maximaxoo	@atpworldtour	@ussoccer
Mention	@twcbreaking	@kayla_itsines	@4freedominiran	@bipartisanism	@pjnet_blog	@unicef	@robinwilliams	@esa_rosetta	@andy_murray	@mcfc
Mention	@redcross	@canproveit	@orgiac	@theanonmessage	@espuelasvox	@free_media_hub	@historicalpics	@astro_reid	@wta	@realmadriden
Term	typhoon	health	nuclear	police	obama	israeli	robin	space	tennis	liverpool
Term	philippines	ebola	regime	protesters	gun	israel	williams	solar	murray	cup
Term	magnitude	outbreak	iran	officer	america	gaza	walker	moon	djokovic	supporting
Term	storm	virus	iranian	cops	obamacare	palestinian	cory	houston	federer	match
Term	usgs	acrx	mullahs	protest	gop	killed	paul	star	nadal	goal

	Tennis	Space	Soccer	Iran Deal	Human Disaster	Celebrity Death	Social Issues	Natural Disaster	Epidemics	LGBT	Mean
Mention	0.53	1.33	4.68	0.22	3.81	0.52	2.21	0.31	1.4	4.38	1.94
Hashtag	2	9.35	25.66	1.16	17.53	4.49	7.53	2.08	10.29	18.99	9.91
From	0.1	0.55	1.6	0.07	0.91	0.07	0.4	0.11	0.83	2.29	0.69
Location	0.1	0.4	1.01	0.03	0.77	0.11	0.38	0.12	0.46	1.5	0.49
Term	1.51	3.05	13.51	0.58	12.81	2.78	5.71	1.51	4.21	8.11	5.38

Figure 3. Matrix of mean Mutual Information values for different feature types vs. topics. The last column as average of mean values across all topics. All values should be multiplied by 10^{-8}.)



(a) Term **(b)** Location **(c)** Hashtag

(d) Mention **(e)** From

Figure 4. Box plots of Mutual Information values (y-axis) per feature type across topics (x-axis labels).

in machine learning. To calculate the amount of information that each feature $j \in \{From \cup Hashtag \cup Mention \cup Term \cup Location\}$ provides w.r.t. each topic label $t \in \{NaturalDisaster, Epidemics, \ldots\}$, Mutual Information is formally defined as

$$I(j,t) = \sum_{t \in \{0,1\}} \sum_{j \in \{0,1\}} p(j,t) \log \left(\frac{p(j,t)}{p(j)p(t)} \right),$$

with marginal probabilities of topic $p(t)$ and feature $p(j)$ occurrence and joint probability $p(t,j)$ computed over the sample space of all tweets, where higher values for this metric indicate more informative features j for the topic t.

In order to answer the first question regarding the best features for learning topical classifiers, we provide the mean Mutual Information values for each feature across different topics in Fig. 3. The last column in Fig. 3 shows the average of the mean Mutual Information for each feature type. From analysis of Table 3, we can make a set of observations:

- The *Term* and *Location* features are the most informative features on average.
- The Location feature provides the highest MI regarding the topics of *HumanDisaster*, *LBGT*, and *Soccer* indicating a lot of content in these topics is heavily localized.

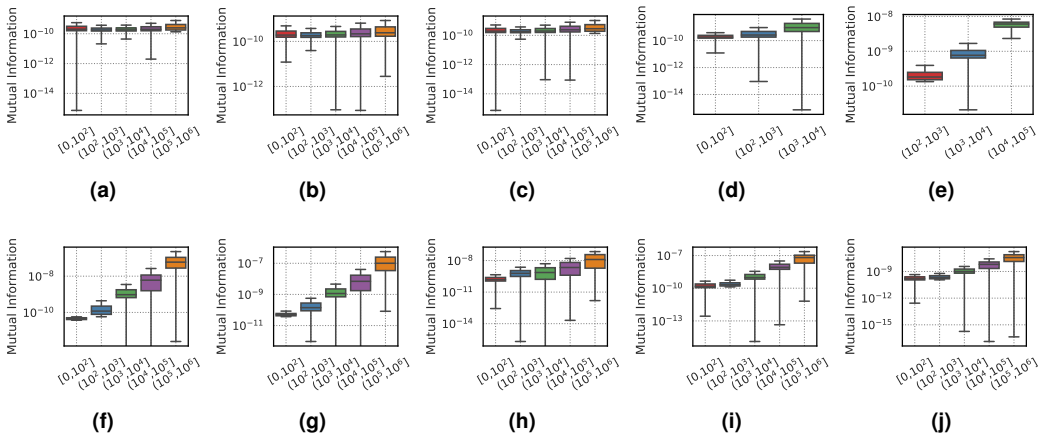


Figure 5. Violin plots for the distribution of Mutual Information values (y-axis) of different features as a function of their attribute values (binned on x-axis). Plots (a-e) respectively show attributes {favoriteCount, followerCount, friendCount, hashtagCount, tweetCount} for *From* feature. Plots (f-j) respectively show attributes tweetCount and userCount for *Hashtag*, userCount for *Location* feature, tweetCount for *Mention* and *Term* features.

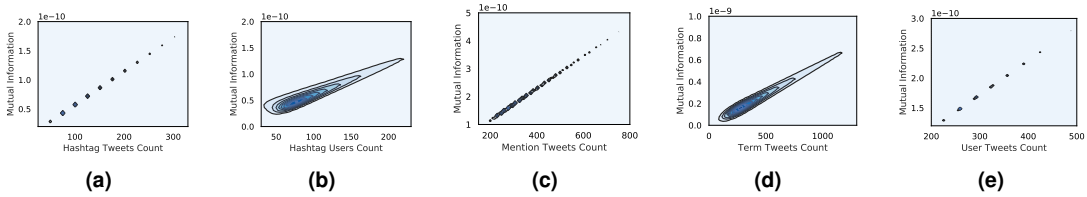


Figure 6. Density plots for the frequency values of feature attributes (x-axis) vs. Mutual Information (y-axis). Plots (a-d) respectively show attributes {favoriteCount, followerCount, friendCount, hashtagCount} for the *From* feature.

- Looking at the overall average values, the order of informativeness of feature types appears to be the following: *Term*, *Location*, *Hashtag*, *Mention*, *From*.

To further analyze the relationship between the informativeness of feature types and topics, we refer to the box plots of Fig. 4. Here we see the quartiles and outliers of the distribution rather than just the average of the MI values in order to ensure the mean MI values were not misleading our interpretations. Overall, the story is the same: term and location features dominate in terms of MI followed by the other less informative features. Furthermore, two observations are apparent: (1) terms have more outliers indicating that *the most useful individual features may be terms*, and (2) the topic has little impact on which feature is most important indicating *stability of feature type informativeness over topics*.

As anecdotal evidence to inspect which features are most informative, we refer to Table 6, 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 figure so prominently in terms of their informativeness. The top *locations are also highly relevant to most topics* indicating the overall importance of these tweet features for identifying topical tweets.

In order to answer the second question on whether any attributes correlate with importance for each feature, we provide two types of analysis. The first analysis shown in Fig. 5 analyzes the distributions of Mutual Information values for features when binned by the magnitude of various attributes of those features, outlined as follows:

- **From** vs.

  - *Favorite count:* # of tweets user has favorited.
  - *Followers count:* # of users who follow user.
  - *Friends count:* # of users followed by user.
  - *Hashtag count:* # of hashtags used by user.
  - *Tweet count:* # of tweets from user.
- **Hashtag** vs.
  - *Tweet count:* # of tweets using hashtag.
  - *User count:* # of users using hashtag.
- **Location** vs. *User count:* # of users using location.
- **Mention** vs. *Tweet count:* # of tweets using mention.
- **Term** vs. *Tweet count:* # of tweets using term.

As we can see in the Violin plots of Fig. 5, the general pattern is that the greater the number of tweets, users, or hashtag count a feature has, the more informative the feature is in general. This pattern also exists to some extent on the attributes of the *From* feature, although the pattern is less visible in general and not clear (or very weak) for the follower or friend count. In general, the informativeness of a user appears to have little correlation with their follower or friend count.

Fig. 6 provides a further analysis by showing density plots of favorite count, follower count, friends count, and hashtag count attributes of the *From* feature. Here we see an interesting phenomenon that was not clear in the Violin plots: there is a very clear bimodality of the density. On further investigation it turns out that the top mode feature occurs in at least one topical tweet whereas the bottom mode occurs in no topical tweets. While the bottom mode features may serve as good indicators of non-topicality, the top mode are inherently more indicative of topicality, which justifies feature selection by mutual information.

RELATED WORK

Twitter Topic Classification

Topic classification for social media aims to detect and track general topics such as ”Baseball” or ”Fashion”. In previous work, researchers have collected labeled data either by using a single hashtag for each topic (Lin et al., 2011), a user-defined query for each topic (Magdy and Elsayed, 2014), or co-training based on the URLs and text of the tweet (Yang et al., 2014). We expand on (Lin et al., 2011)’s work and use a set of hashtags instead of a single hashtag. Similarly, we extract features consisting of hashtags, mentions, unigram terms, and authors as done in this prior work, but also add location as another feature, which has shown to be the second most important feature for topic classification after unigram terms. Furthermore, we provided a novel learning and evaluation paradigm based on splitting both the data and hashtags along temporal boundaries to generate train, validation and test datasets in order to evaluate long-term generalization of trained topic classifiers. In contrast, we remark that (Lin et al., 2011) only evaluated over 1 week, (Magdy and Elsayed, 2014) over 4 days, and (Yang et al., 2014) did not explicitly mention the data duration or that their study was intended to assess long-term performance. Hence these previous studies do not permit one to assess the long-term topic classification performance of topic classifiers for Twitter as intended by the 2 year longitudinal study performed in this article.

Related Applications of Classifiers for Social Media

Aside from highly related work on supervised topic classifiers for Twitter Lin et al. (2011); Yang et al. (2014); Magdy and Elsayed (2014) that motivated this study as discussed previously, there are many other uses of classifiers for social media. While we argue no prior work has performed a longitudinal analysis of supervised Twitter topical classifiers as done in this article, these alternative applications of classifiers for social media may broadly benefit from the insights gained by our present study. We cover these

related uses below along with important differences with the present work, divided into the following four subareas: (1) trending topic detection, (2) tweet recommendation, (3) friend sensors, and (4) specific event detection such as earthquake or influenza sensors.

Trending Topic Detection represents one of the most popular types of topical tweet detector and can be subdivided into many 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 et al. (2010); Ishikawa et al. (2012); Phuvipadawat and Murata (2010); Becker et al. (2011); O'Connor et al. (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 et al. (2012); Aiello et al. (2013). The third category of methods extends the previous categories by additionally exploiting network structure properties Budak et al. (2011). Despite this important and very active area of work that can be considered a type of topical tweet detector, trending topic detection is intrinsically unsupervised and not intended to detect targeted topics. In contrast, the work in this article is based on supervised learning of a specific topical tweet detector trained on the topical set of hashtags provided by the user.

Tweet Recommendation represents an alternate use of tweet classification 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 et al. (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 et al. (2013); Xu and Yang (2012); Petrovic et al. (2011). Despite the fact that 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 classifiers 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 topical tweet detectors as we do in this work but focuses on highly specific events such as 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 et al. (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 et al. (2011), e.g, by exploiting knowledge of users' proximity and friendship along with the contageous nature of influenza Sadilek et al. (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 topic-based classifiers for Twitter as analyzed in this work.

Friend Sensors are a fourth and final class of social sensors 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 classifiers 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 classifier suggesting that general topical classifiers can benefit from a wide variety of features well beyond those of author features alone.

CONCLUSIONS

This work provides a long-term study of topic classifiers on Twitter that further justifies classification-based topical filtering approaches while providing detailed insight into the feature properties most critical for topic classifier performance. Our results suggest that these learned topical classifiers generalize well to unseen future topical content over a long time horizon (i.e., one year) and provide a novel paradigm for the extraction of high-value content from social media. Furthermore, an extensive analysis of features and

feature attributes across different topics has revealed key insights including the following two: (i) largely independent of topic, generic terms are the most informative features followed by topic-specific locations, and (ii) the number of unique hashtags and tweets by a user correlates more with their informativeness than their follower or friend count.

Among many interesting directions, future work might evaluate a range of topical classifier extensions: (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, (3) identification of long-term temporally stable predictive features, and (4) utilizing more social network structure as graph-based features. Altogether, we believe these insights will facilitate the continued development of effective topical classifiers for Twitter that learn to identify broad themes of topical information with minimal user interaction and enhance the overall social media user experience.

REFERENCES

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

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*, pages 1481–1484.

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, pages 661–670. ACM.

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, 2012*, pages 1794–1798.

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

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, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477.

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.

Iman, Z., Sanner, S., Bouadjenek, M. R., and Xie, L. (2017). A longitudinal study of topic classification on twitter. In *ICWSM*, pages 552–555.

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*, pages 1–5.

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, pages 53–54, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

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.

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 *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 422–429. ACM.

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 University Press, New York, NY, USA.

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*, pages 1155–1158.

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*, pages 41–48. AAAI Press.

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*, pages 189–198.

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, 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, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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*, pages 120–123.

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, 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, 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, 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*, pages 46–50.

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, pages 516–525, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang, S.-H., Kolcz, A., Schlaikjer, A., and Gupta, P. (2014). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1907–1916. ACM.

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.