# A Longitudinal Study of Topic Classification on Twitter

**Authors**
Affiliations

**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 (**?**; **?**; **?**), 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, 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 (1) such classifiers can indeed generalize to novel topical content with high precision over a year or more after training, (2) simple terms and locations are the most informative feature classes (despite the intuition that hashtags may be the most topical feature class), and (3) 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.

## 1 Learning Topical Social Sensors

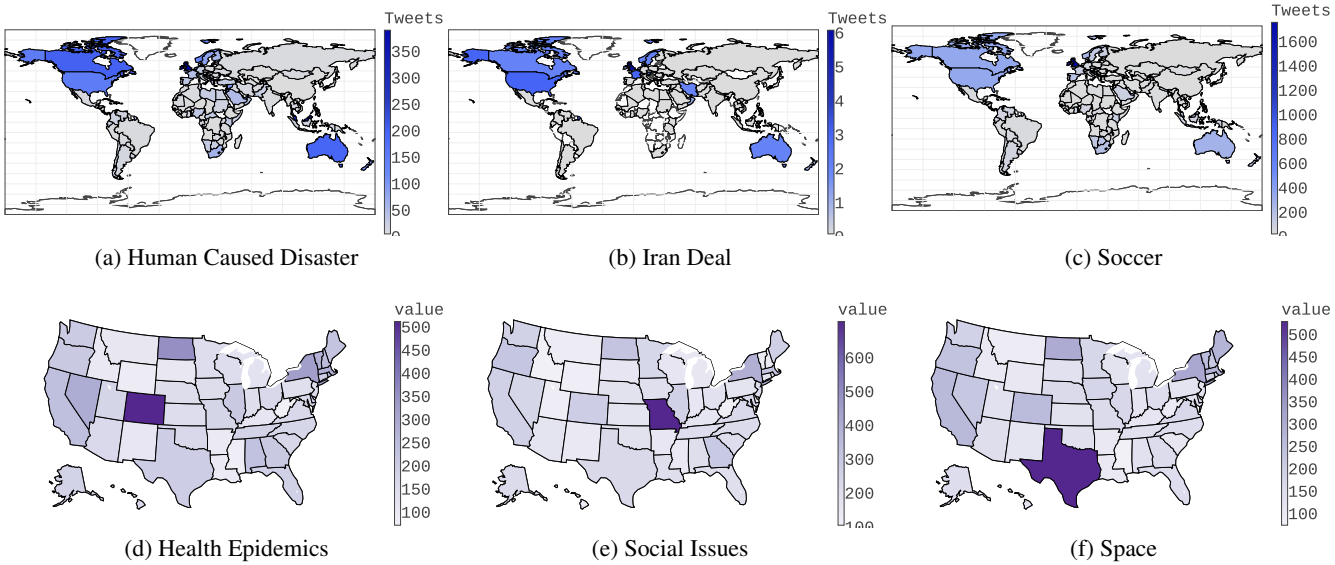
Our objective is to build a binary classifier that can label a previously unseen tweet as topical (or not) by training on topically labeled historical tweets. Following the approach of (**?**), for each topic $t \in T$, we leverage a (small) set of user-curated topical hashtags $H^t$ to efficiently provide a large number of supervised topic labels for social media content. 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 $H^t_{\text{train}}$, $H^t_{\text{val}}$, and $H^t_{\text{test}}$ 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 to generate positive labels in the test data and the remainder are used to label the validation data.

The critical insight here is that we not only partition the train, validation, and test data temporally, but we also divide the hashtag 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.

Next we proceed to train our classifier and select hyperparameters to optimize Average Precision (AP) (**?**) (a ranking metric) on the validation data (and hashtag labels). Because all of our classifiers provide a real-valued score (e.g., logistic regression provides the probability of the class), the classifiers can be used to rank. In our results, we report final evaluations on the held-out test data with their own disjoint set of topical hashtag labels.

## 2 Data Description

Now we provide 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




(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.

**#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.

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 **??**. 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. **??** 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.

## 3 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 **??** 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. 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 **??**. 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 $D_i^+$

| | 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.

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 **??**. 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 to generate a score function $f^t$ for learning topical classifiers as defined in Sec. **??**. In this paper, we experiment with the following four state-of-the-art supervised classification and ranking methods:

1. **Logistic Regression** using LibLinear (**?**)
2. **Bernoulli Naïve Bayes** (**?**)
3. **Rocchio** (**?**)
   (a centroid-based classifier)
4. **RankSVM** (**?**)

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

- *Logistic Regression*: $L_2$ regularization constant $C$ is tuned for $C \in \{1E-12, 1E-11, ..., 1E+11, 1E+12\}$.
- *Naïve Bayes*: Dirichlet prior $\alpha$ is tuned for $\alpha \in \{1E-20, 1E-15, 1E-8, 1E-3, 1E-1, 1\}$.
- *All Classfiers*: The number of top features $M$ selected based on their Mutual Information is tuned for $M \in \{1E2, 1E3, 1E4, 1E5, 1166582 \text{ (all features)}\}$.

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 (where it is possible to select all features). Hyperparameter tuning is done via exhaustive grid search and using the Average Precision (AP) (**?**) 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.

## 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 **??** 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 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 **??**. We've annotated tweets with symbols as follows:

| | | 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 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.

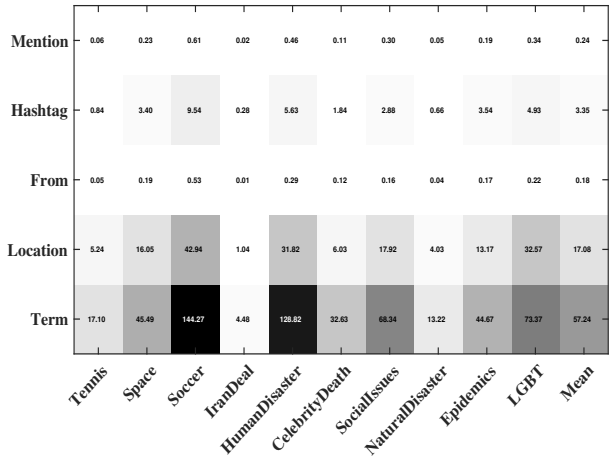



Figure 2: Matrix of mean Mutual Information values for different feature types vs. topics. The last column as average of mean values across all topics. All values should be multiplied by $1E + 10$.)

- ✓: 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 **??** 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.

# 4 Feature Analysis

In this section, we analyze the informativeness of our defined features in Sec **??** 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?

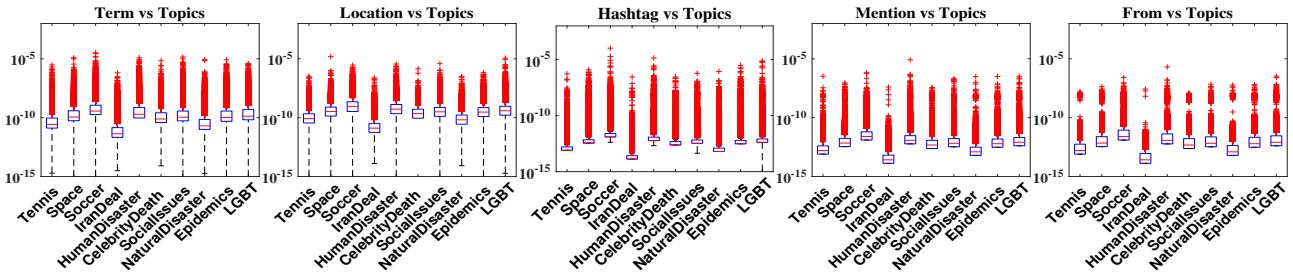
To answer these questions, we use Mutual Information (MI) (**?**) as our primary metric for feature evaluation. Mutual Information is a general method for measuring the amount of information one random variable contains about another random variable and is used to select predictive features in machine learning. To calculate the amount of information that each feature $j \in \{\textit{From} \cup \textit{Hashtag} \cup \textit{Mention} \cup \textit{Term} \cup \textit{Location}\}$ provides w.r.t. each topic label $t \in \{NaturalDisaster, Epidemics, ...\}$, 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. **??**. The last column in Fig. **??** shows the average of the mean Mutual Information for each feature type. From analysis of Table **??**, 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.




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

Figure 3: Box plots of Mutual Information values (y-axis) per feature type across topics (x-axis labels).

| 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.

- 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. **??**. 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 **??**, 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. **??** 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. **??**, 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. **??** 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.

## 5 Conclusions and Future Work