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 and political events to entertainment and sports news. While 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).

1 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 $H^t_{\text{train}}$, $H^t_{\text{val}}$, and $H^t_{\text{test}}$ according to two time stamps $t^\text{split}_{\text{val}}$ and $t^\text{split}_{\text{test}}$ 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 $h_{\text{time}}^* < t^\text{split}_{\text{val}}$ are used to generate positive labels in the training data, those with $h_{\text{time}}^* \geq t^\text{split}_{\text{test}}$ 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.

2 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...
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.

### 3 Empirical Evaluation

With the formal definition of learning topical classifiers provided in Sec. 1 and the overview of our data in Sec. 2, 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. 2, 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 roughly 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. 1 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. 1. In this paper, we experiment with the following four classifiers or rankers:
Table 2: Test/Train Hashtag samples and statistics.

<table>
<thead>
<tr>
<th>Feature</th>
<th>#TrainHashtags</th>
<th>#TestHashtags</th>
<th>#TopicalTweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space</td>
<td>58</td>
<td>65</td>
<td>554</td>
</tr>
<tr>
<td>Soccer</td>
<td>98</td>
<td>91</td>
<td>8,762</td>
</tr>
<tr>
<td>FameBeat</td>
<td>1,26</td>
<td>13</td>
<td>52</td>
</tr>
<tr>
<td>Humanitarian</td>
<td>49</td>
<td>19</td>
<td>230,058</td>
</tr>
<tr>
<td>CelebrityDeath</td>
<td>28</td>
<td>19</td>
<td>163,890</td>
</tr>
<tr>
<td>SocialIssues</td>
<td>33</td>
<td>29</td>
<td>230,058</td>
</tr>
<tr>
<td>Natural Disaster</td>
<td>31</td>
<td>17</td>
<td>244,478</td>
</tr>
<tr>
<td>Epipemics</td>
<td>52</td>
<td>29</td>
<td>210,217</td>
</tr>
<tr>
<td>LGBT</td>
<td>28</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Unique Values</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mention</td>
<td>159</td>
<td>361,789</td>
<td></td>
</tr>
<tr>
<td>Hashtag</td>
<td>159</td>
<td>184,702</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>50</td>
<td>57,767</td>
<td></td>
</tr>
<tr>
<td>Term</td>
<td>50</td>
<td>317,846</td>
<td></td>
</tr>
<tr>
<td>Features (CF)</td>
<td>-</td>
<td>1,166,582</td>
<td></td>
</tr>
</tbody>
</table>

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 1, 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 $\alpha$ is tuned for $\alpha \in \{10^{-20}, 10^{-15}, 10^{-8}, 10^{-3}, 10^{-1}, 1\}$.
- **All:** Number $M$ of features selected based on highest Mutual Information (Manning, Raghavan, and Schütze 2008) [Sec. 13.5.1] with the class topic label is tuned for $M \in \{10^2, 10^3, 10^4, 10^5\}$.

We remark that many algorithms such as Naïve 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) ranking metric (Manning, Raghavan, and Schütze 2008) [Sec. 8.4] discussed 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 defined in (Manning, Raghavan, and Schütze 2008) [Sec. 8.4]:

- **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 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 2 and the effect of their attributes on learning targeted topical classifiers. We use Mutual Information (MI) (Manning, Raghavan, and Schütze 2008) [Sec. 13.5.1] as our primary metric for feature evaluation, where higher values for MI 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 for each feature type. We observe the following:
Across all topics, the Term and Location features are the most informative features on average (despite training on topics with class labels determined by hashtags).

Looking at only the Location feature, the highest MI across topics occurs for HumanDisaster, LBGT, and Soccer indicating that the content in these topics is likely to be more geographically localized than the other topics.

Looking at the overall mean MI values, the order of feature types from most to least informative is the following: Term, Location, Hashtag, Mention, From. We also remark that the mean MI values for Term and Location are an order of magnitude greater than the other feature types.

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


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

