Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages

Aron Culotta

Received: date / Accepted: date
This is a preprint: The final version available at springer.com

Abstract We analyze over 570 million Twitter messages from an eight month period and find that tracking a small number of keywords allows us to estimate influenza rates and alcohol sales volume with high accuracy. We validate our approach against government statistics and find strong correlations with influenza-like illnesses reported by the U.S. Centers for Disease Control and Prevention ($r(14) = .964$, $p < .001$) and with alcohol sales volume reported by the U.S. Census Bureau ($r(5) = .932$, $p < .01$). We analyze the robustness of this approach to spurious keyword matches, and we propose a document classification component to filter these misleading messages. We find that this document classifier can reduce error rates by over half in simulated false alarm experiments, though more research is needed to develop methods that are robust in cases of extremely high noise.

Keywords social media, regression, classification

1 Introduction

There has been growing interest in monitoring disease outbreaks using the Internet. Previous approaches have applied data mining techniques to news articles [13,21,1,29,5,19], blogs [6], search engine logs [9,27,12], and Web browsing patterns [15]. The recent emergence of *micro-blogging* services such as Twitter.com presents a promising new data source for Internet-based surveillance because of message volume, frequency, and public availability. The principal advantages over traditional data collection approaches are lower cost and more rapid results. For example, to obtain an estimate of the influenza rate, the U.S. Centers for Disease Control surveys thousands of hospitals which is both costly and typically has a

Dr. Aron Culotta
Department of Computer Science & Industrial Technology
Southeastern Louisiana University
Hammond, LA 70402
E-mail: culotta@selu.edu

reporting lag of one to two weeks. Additionally, since not all infected people are admitted to a hospital, more informal means of data collection may provide greater insight into the spread of the disease. Providing a real-time estimate of influenza rates may provide public health agencies with an early-warning system, which can help inform decisions such as allocation of medical resources and public messaging campaigns.

Initial work in this direction includes Ritterman et al. [30], who show that Twitter messages can improve the accuracy of market forecasting models by providing early warnings of external events like the H1N1 outbreak. More recently, de Quincey & Kostkova [28] have demonstrated the potential of Twitter in outbreak detection by collecting and characterizing over 135,000 messages pertaining to the H1N1 virus over a one week period, though no attempt is made to estimate influenza rates.

Two similar papers were recently published that estimate national influenza rates from Twitter messages [17, 7]. Both use linear regression to detect keywords that correlate with influenza rates, then combine these keywords to estimate national influenza rates. Lampos & Cristianini [17] train and evaluate on a much larger data set (28 million messages) than used in Culotta [7] (500K messages), which likely contributes to the differing quality of the estimates (.97 correlation with national statistics on held-out data in Lampos & Cristianini [17], .78 correlation in Culotta [7]).

In contrast to previous methods that rely on computationally intensive models, the main result of this paper is that simple keyword matching techniques can result in accurate estimates of influenza rates. This result is in large part due to the immense volume of Twitter messages posted each day.

We report results of our analysis of over 570 million Twitter messages collected in the 8 months from September 2009 to May 2010. This data was originally collected as part of the work of O'Connor et al. [24], in which a strong correlation is revealed between certain Twitter messages and political opinion polls.

The contributions of this paper are as follows:

- We find that simple keyword matching produces a surprisingly high correlation with national statistics. For example, the proportion of Twitter messages containing flu-related keywords produces a .95 held-out correlation with weekly influenza-like-illness statistics reported by the U.S. Centers for Disease Control and Prevention.
- We replicate this result in the domain of alcohol sales volume estimation. A simple model based on the frequency of the word "drunk" produces a .93 correlation with sales estimates from the U.S. Census Bureau.
- Despite these strong correlations, we find that the methodology of selecting keywords based on their correlation with national statistics can sometimes be problematic because of the likelihood of detecting false correlations. For example, the phrase "flu shot" has a correlation greater than .90, but certainly this is not a good term to monitor, as it may spike in frequency without a corresponding spike in influenza rates. We propose a method to estimate robustness to false alarms by simulating false outbreaks like those described above. Using this measure, we show that by adding a document classification component to remove spurious keyword matches, we can reduce the severity of false alarms while preserving accurate forecasting estimates.

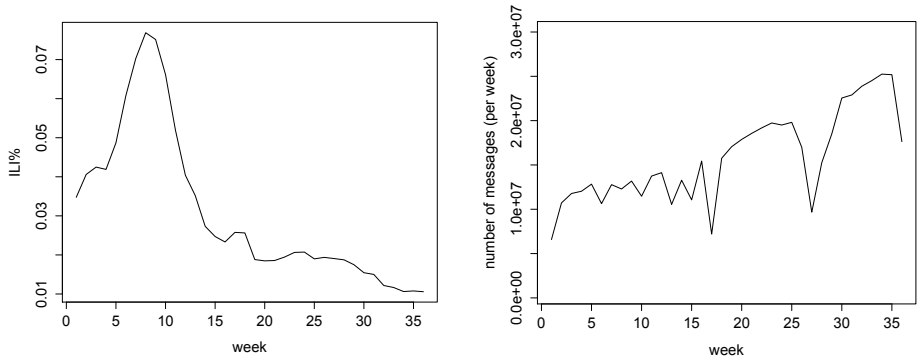

Fig. 1 The left figure shows the ILI rates as obtained form the CDC's weekly tracking statistics. Week 1 ends on September 5, 2009; week 36 ends on May 8, 2010. The right figure displays the number of Twitter messages collected per week over the same time period.

In Section 2, we first describe the national influenza statistics as well as the Twitter dataset. Then in Section 3, we describe the methodology of correlating Twitter messages with national statistics, and report correlations on a range of keyword combinations. In Section 4, we discuss the impact of spurious keywords on correlation results. In Section 5 we introduce and evaluate a document classifier to filter spurious messages, which we empirically validate in Section 6. Section 7 replicates these results on a different time span, and Section 8 compares with a more complex regression method (ϵ-SVR). In Section 9, we apply the same methodology to estimate alcohol sales volume, and we conclude with a discussion of related (Section 10) and future (Section 11) work.

2 Influenza Data

We begin with a description of the data used in all influenza-tracking experiments.

2.1 Influenza Monitoring in the United States

The U.S. Centers for Disease Control and Prevention (CDC) publishes weekly reports from the US Outpatient Influenza-like Illness Surveillance Network (ILINet). ILINet monitors over 3,000 health providers nationwide to report the proportion of patients seen that exhibit influenza-like illnesses (ILI), defined as "fever (temperature of 100 ° F [37.8 ° C] or greater) and a cough and/or a sore throat in the absence of a known cause other than influenza."[1] Figure 1 shows the ILI rates for the 36 week period from August 29, 2009 to May 8, 2010.

While ILINet is a valuable tool in detecting influenza outbreaks, it suffers from a high cost and slow reporting time (typically a one to two week delay). The goal of this line of research is to develop methods that can reliably track ILI rates in real-time using Web mining.

[1] `http://www.cdc.gov/flu/weekly/fluactivity.htm`

2.2 Twitter Data

Twitter.com is a micro-blogging service that allows users to post messages of 140 characters or less. Users can subscribe to the *feeds* of others to receive new messages as they are written. As of April 2010, Twitter reports having 105 million users posting nearly 65 million message per day, with roughly 300,000 new users added daily [25].

There are several reasons to consider Twitter messages as a valuable resource for tracking influenza:

- The high message posting frequency enables up-to-the-minute analysis of an outbreak.
- As opposed to search engine query logs, Twitter messages are longer, more descriptive, and (in many cases) publicly available.
- Twitter profiles often contain semi-structured meta-data (city, state, gender, age), enabling a detailed demographic analysis.
- Despite the fact that Twitter appears targeted to a young demographic, it in fact has quite a diverse set of users. The majority of Twitter's nearly 10 million unique visitors in February 2009 were 35 years or older, and a nearly equal percentage of users are between ages 55 and 64 as are between 18 and 24.[2]

The Twitter messages used in this paper are a subset of those used in O'Connor et al. [24], restricted to the 2009-2010 flu season from September 2009 to May 2010. O'Connor et al. [24] gathered the messages through a combination of queries to Twitter's public search API as well as messages obtained from their "Gardenhose" stream, a pseud-random sample of all public Twitter messages.

Figure 1 shows the number of Twitter messages obtained per week for the same time frame as the ILI percentages. The average number of messages per week is 15.8 million. Due to difficulties in data collection, there are a few anomalies in the data – i.e., the drop in messages for weeks 17 and 27 was due to a bug in data collection, not a drop in actual Twitter usage. However, even the smallest sample (week 1) contains 6.5 million messages.

3 Correlating keywords with ILI rates

In this section, we describe a methodology to correlate Twitter messages with ILI rates. The three stages of the approach are: (1) identify the proportion of Twitter messages in week i in which users describe having flu-like symptoms; (2) fit a linear regression model in which the independent variable is the Twitter proportion from (1) for week i and the dependent value is the ILI rate for week i; (3) use the regression model to predict ILI rates on future weeks, validating against the CDC data.

This basic approach is used in Ginsberg et al. [12] to estimate influenza rates from query log data. Note that this approach predicts each week in isolation. While there are many well-known time-series prediction techniques that incorporate previous time steps into the regression model, we choose this relatively simple model

[2] *Twitter older than it looks*. Reuters MediaFile blog, March 30th, 2009.

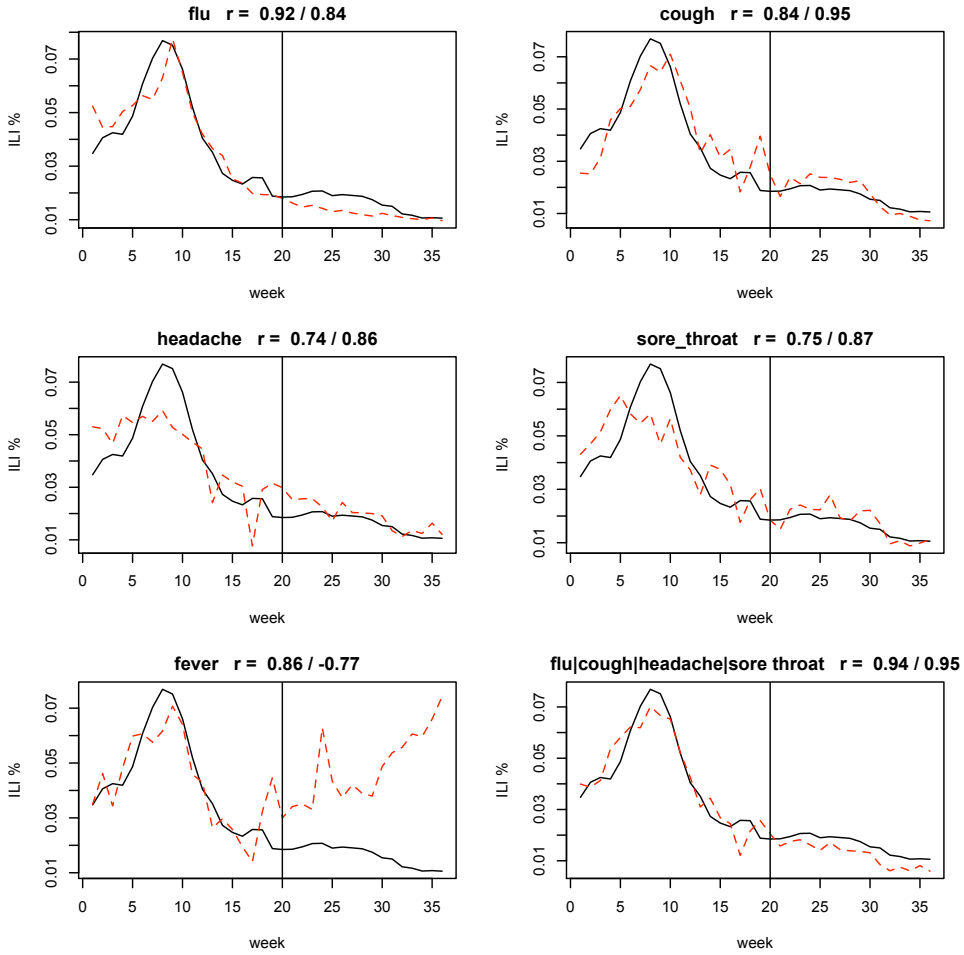

Fig. 2 Fitted and predicted ILI rates using regression over query fractions of Twitter messages. Black (solid) lines are the true ILI rates, red (dotted) lines are the fitted/predicted values, with the vertical line splitting the training and evaluation data.

for two reasons: (1) it allows us to confirm the transferability of the approach of Ginsberg et al. [12] to this new data source; and (2) it allows us to more precisely test our hypothesis that the content of Twitter messages at time i correlates with the ILI rate at time i. As we shall see, this simple approach works well in these domains. We leave for future work the exploration of more sophisticated time-series models.

We now describe the method more formally. Let P be the true proportion of the population exhibiting ILI symptoms. In all experiments, we assume P is the value reported by the CDC's ILINet program.

Let $W = \{w_1 \dots w_k\}$ be a set of k keywords, let D be a document collection, and let D_W be the set of documents in D that contain at least one keyword in W. We define $Q(W, D) = \frac{|D_W|}{|D|}$ to be the fraction of documents in D that match W, which we refer to as the *query fraction*.

Following Ginsberg et al. [12], we first consider a simple linear model between the log-odds of P and $Q(W, D)$:

$$logit(P) = \beta_1 logit(Q(W, D)) + \beta_0 + \epsilon \quad (1)$$

with coefficients β_1, β_0, error term ϵ, and logit function $logit(X) = \ln(\frac{X}{1-X})$.

Figure 2 displays the result of this regression for a number of keywords. We fit the regression on weeks 1-20, and evaluate on weeks 21-36. In each figure, the black line is the true ILI rate, the red line is the fitted/predicted rate. The vertical line indicates the transition from training to evaluation data. The title of each plot indicates the query used as well as the training and evaluation correlation values.

These results show extremely strong correlations for all queries except for *fever*, which appears frequently in figurative phrases such as "I've got Bieber fever" (in reference to pop star Justin Bieber).

We note these results are competitive with those found in Lampos & Cristianini [17] using U.K. data, who obtain a .97 correlation with the U.K.'s Health Protection Agency statistics using 73 keywords and a more sophisticated keyword weighting scheme (See Section 10 for more discussion.) Note that direct comparisons are difficult due to differences in time span and location (U.K. versus U.S.). The conclusion we draw from these results is that even extremely simple methods can result in quite accurate models of ILI rates from Twitter data.

4 Analysis of spurious matches

While these strong correlations are encouraging, we must be careful about the conclusions we draw. For example, a number of messages containing the term "flu" are actually discussing "flu shots", "flu vaccines", or are simply referencing news stories about the flu. While these type of messages may correlate with ILI rates, they are likely not the types of messages researchers have in mind when they report these correlations. That is, people get flu shots without having the flu; so these terms track flu-related events, but not necessarily flu symptoms. Instead, the system would ideally track mentions of people reporting having the flu or flu-like symptoms, as opposed to simply mentioning the flu in passing.

These spurious correlations can leave keyword-based methods vulnerable to false alarms. For example, a recall of a flu vaccine, a governmental policy announcement regarding flu, or a release of a new flu shot will all lead to a spike in messages containing the word flu.

Figure 3 displays regression results for a number of potential spurious keywords, such as *swine or H1N1*, *shot*, *vaccine*, *season*, and *http* (to heuristically filter messages that are simply linking to stories about the flu). Because of the large amount of noise introduced by discussion of the H1N1 virus, we have filtered those terms from all results in this figure. Table 1 shows the total number of messages matching each of the queries.

We make two observations concerning Figure 3. First, notice that removing messages containing the terms "swine" and "H1N1" greatly improves correlation on both the training and evaluation data over using the query "flu" alone (training correlation improves from .93 to .97, evaluation improves from .84 to .91).

Second, notice that removing the other spurious terms does not obviously result in a better fit of the data. In fact, the training correlation declines by .02, and the

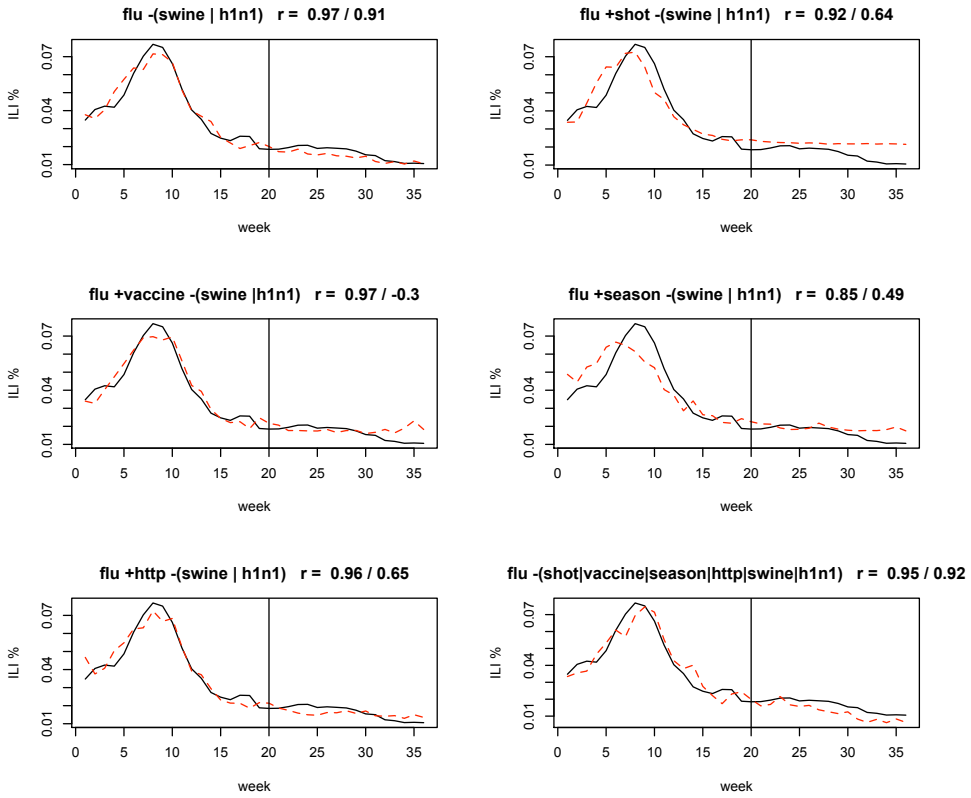

Fig. 3 Correlation results with refinements of the flu query.

Table 1 Number of messages containing the keyword "flu" and a number of keywords that might lead to spurious correlations, along with training and testing correlations. The final row uses messages without those spurious words.

query	**# messages**	r **(training)**	r **(testing)**
flu	387,405	0.92	0.84
flu -(swine | h1n1)	198,149	0.97	0.91
flu +shot -(swine | h1n1)	16,042	0.92	0.64
flu +vaccine -(swine | h1n1)	6,561	0.97	-0.3
flu +season -(swine | h1n1)	7,007	0.85	0.49
flu +http -(swine | h1n1)	43,928	0.96	0.65
flu -(swine | h1n1 | shot | vaccine | season | http)	126,194	0.95	0.92

evaluation correlation improves by .01. Thus, methods that use held-out correlation to select keywords may still be vulnerable to spurious terms. This result emphasizes the need to *explicitly* test for robustness in the presence of false alarms, since other measures do not penalize these spurious terms. We propose such a measure in Section 6.

5 Filtering spurious matches by supervised learning

We propose mitigating the spurious message problem by training a document classifier to label whether a message is reporting an ILI-related event or not. This is related to problems such as *sentiment analysis* [26] and *textual entailment* [10], which in their most general form can be quite difficult due to the ambiguities and subtleties of language. We limit this difficulty somewhat here by only considering documents that have already matched the hand-chosen ILI-related terms *flu, cough, headache, sore throat*. We choose these words due to their high correlation with CDC statistics, as shown in the previous sections. The classifier then calculates a probability that each of these messages is in fact reporting an ILI symptom.

We train a bag-of-words document classifier using logistic regression to predict whether a Twitter message is reporting an ILI symptom. Let y_i be a binary random variable that is 1 if document d_i is a positive example and is 0 otherwise. Let $\mathbf{x}_i = \{x_{ij}\}$ be a vector of observed random values, where x_{ij} is the number of times word j appears in document i. We estimate a logistic regression model with parameters θ as:

$$p(y_i = 1|\mathbf{x}_i; \theta) = \frac{1}{1 + e^{(-\mathbf{x}_i \cdot \theta)}} \tag{2}$$

We learn θ using L-BFGS gradient descent [20] as implemented in the MALLET machine learning toolkit[3].

5.1 Combining filtering with regression

We consider two methods to incorporate the classifier into the regression model in Equation 1. The first method, which we term **soft classification**, computes the *expected fraction* of positively classified documents as

$$Q_s(W, D) = \frac{\sum_{d_i \in D_W} p(y_i = 1|\mathbf{x}_i; \theta)}{|D|} \tag{3}$$

This procedure can be understood as weighting each matched document in D_W by the probability that it is a positive example according to the classifier.

The second method, which we term **hard classification**, simply uses the predicted label for each document, ignoring the class probability. For the binary case, this simply counts the number of documents for which the probability of the positive class is greater than 0.5:

$$Q_h(W, D) = \frac{\sum_{d_i \in D_W} \mathbf{1}(p(y_i = 1|\mathbf{x}_i; \theta) > 0.5)}{|D|} \tag{4}$$

For both methods, we substitute $Q(W, D)$ in Equation 1 with the corresponding classification quantity from Equation 3 or 4.

[3] `http://mallet.cs.umass.edu`

Table 2 Six Twitter messages labeled as positive or negative examples of an ILI-related report. A total of 206 messages were labeled to train the classifier of Section 5.

Positive Examples
Headache, cold, sniffles, sore throat, sick in the tummy.. Oh joy !! :' (
me too... i have a headache my nose is stopped up and it hurts to swallow :/
im dying , got flu like symptoms, had to phone in ill today coz i was in yest n was ill and was making mistakes :(
Negative Examples
swine flu actually has nothing to do with swine. #OMGFACT to the point where they tried to rename the virus
Links between food, migraines still a mystery for headache researchers `http://ping.fm/UJ85w`
are you eating fruit breezers. those other the yummy ones haha. the other ones taste like well, cough drops haha.

Table 3 Results of 10-fold cross validation on the message classification task, with standard errors in parentheses.

Method	**Accuracy**	**F1**	**Precision**	**Recall**
Logistic Regression	83.83% (3.2)	89.46 (2.5)	85.31 (3.6)	94.89 (2.2)
SVM	83.98% (1.2)	90.01 (0.9)	94.38 (2.2)	86.63 (1.4)
Decision Tree	81.48% (2.4)	86.53 (2.4)	93.40 (3.1)	81.07 (2.8)

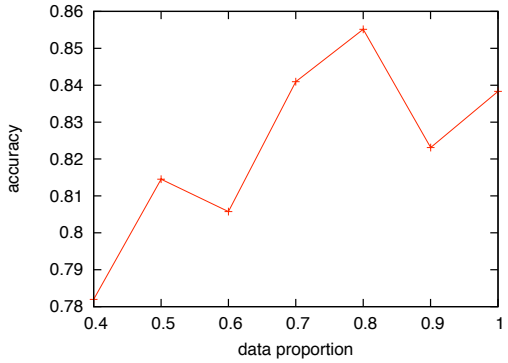

Fig. 4 Logistic regression 10-fold cross validation accuracy as the number of labeled examples increases.

5.2 Filtering Results

To generate labeled data for the classification model of Section 5, we sampled Twitter messages containing any of *flu, cough, headache, sore throat*, making sure the messages were posted outside of the date range of the previously collected messages (from the end of May, 2010). We choose these words because of their observed high correlation to the CDC statistics, as reported above. The messages were sampled using Twitter's Search API. We sampled 206 messages, which were manually categorized into 160 positive examples and 46 negative examples. We labeled as positive examples messages that contain a user reporting flu-like symptoms, otherwise the message is labeled as a negative example. Examples are shown in Table 2.

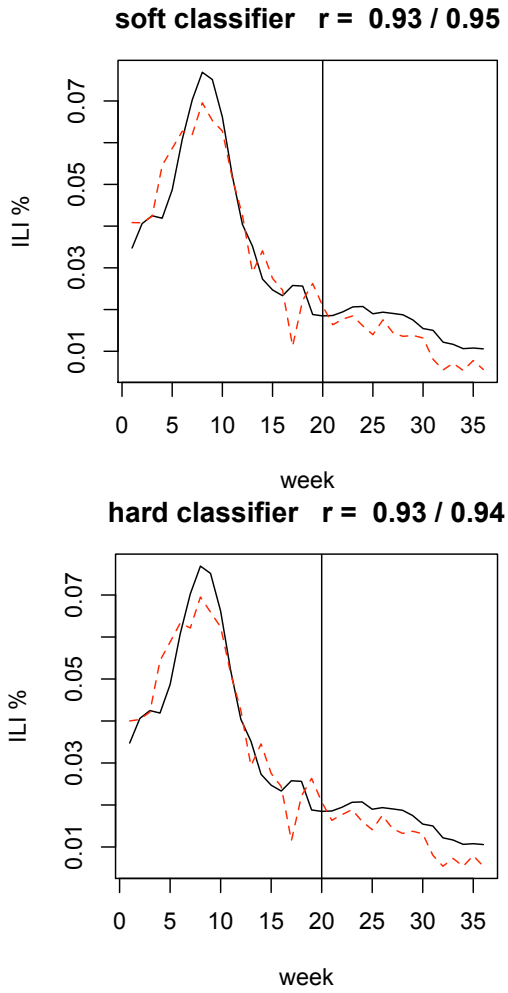

Fig. 5 Correlation results using both classification strategies to filter spurious messages from the query *(flu or cough or headache or sore throat)*. Filtering does not significantly affect the correlation results obtained by the original query, as seen by comparing with the final graph in Figure 2.

In addition to the logistic regression classifier, we also compare with a support vector machine using a linear kernel[4] and an ID3 decision tree[5]. For all experiments, we use a bag-of-words document representation (i.e., multinomial distributions over words), ignoring case and punctuation. Results of 10-fold cross-validation on this data are shown in Table 3. Here, precision and recall are computed for the positive class.

[4] We use LibSVM [2] with a linear kernel and the default parameter settings.

[5] We use MALLET's (`http://mallet.cs.umass.edu`) implementation with the default parameter settings.

Table 4 Evaluation of hard classification. **Total** is the total number of messages matching the query *flu or cough or headache or sore throat*. **Filtered** is the number removed by the hard classifier. 100 documents were sampled uniformly at random and manually annotated to estimate classifier accuracy.

Total	Filtered	Est. Accuracy	Est. F1	Est. Precision	Est. Recall
992,735	248,969	79.00%	85.72%	79.75%	92.65%

Logistic regression and SVM exhibit comparable performance, though with different precision-recall trade-offs. The higher recall of logistic regression suggest it is a less aggressive filter of negative examples.

Notice that despite the small sample size, the classifier can distinguishes positive and negative messages with fairly high accuracy. To determine the impact of training size on accuracy, we plot a learning curve in Figure 4. This shows how the cross-validation accuracy of logistic regression varies as a function of the labeled data size. This figure suggests that accuracy is plateauing around 80-85%, as accuracy does not increase after using 70% of the data. Thus, adding additional labeled examples would appear to have limited impact on accuracy for this task.

We next embed the logistic regression classifier in the regression model to determine the effect of filtering on estimates of flu rates. Figure 5 shows the results of both filtering strategies. Table 4 displays statistics for the hard classifier. Note that the initial set of documents is collected using the query of the final graph in Figure 2 (*flu or cough or headache or sore throat*). By comparing these two graphs with the corresponding one in Figure 2, we can see that all three methods perform similarly on this data. This suggests that filtering out spurious messages does not hurt performance, and that there likely exists a true underlying correlation between Twitter messages and ILI rates. The fact that filtering does not improve performance may suggest that there are no significant spikes in messages that would lead to a false alarm in the evaluation data. That is, there were not sufficient number of negative examples in the data to significantly harm the accuracy of the influenza rate estimates.

It is important to note that spikes in spurious terms are rare but impactful events. They may not happen often, but when they do, it is important that the system is robust enough to avoid predicting false spikes in flu rates. We explore this further in the next section.

6 Evaluating false alarms by simulation

In this section, we evaluate how well the methods proposed in the previous section filter spikes in spurious messages. Because false alarms are by definition rare events, it is difficult to use existing data to measure this. Instead, we propose simulating a false alarm as follows:

– We first sample 1,000 messages deemed to be spurious. We do this by searching all the data for messages containing *flu or cough or headache or sore throat* sent by users that were news services (e.g., Reuters News). Additionally, we searched for messages containing *associated press, AP*, or *health officials*. These messages were then manually evaluated to determine that they were all negative examples. Details of the spurious message are in Table 5.

Table 5 Details of sampled spurious messages.

# messages	1000	**total words**	25,188	**unique words**	4,792
frequent tri-grams					
"seasonal flu vaccine" (41), "health officials say" (35), "a headache for" (32) , "air travel headaches" (24), "recovering in hospital" (21)					

Table 6 Results of simulated false alarms.

Filtering method	**Mean-squared error**
none	0.077
classify-soft	0.035
classify-hard	0.023

- Next, we sample with replacement an increasing number of these spurious messages. Each sample is added to the original messages collected for May 2, 2010 (3,452,968 messages total). The resulting dataset contains 5 batches of simulated data, ranging from a batch with 0 additional false messages to a batch with 100,000 additional false messages.
- Finally, we use the same trained regression models from Figure 5 to estimate the ILI rates for each of these synthetic datasets.

Table 6 and Figure 6 shows the result of this approach for both classification methods as well as the original keyword based method (which does no filtering). We make three observations of these results. First, either classification method appears to improve over no filtering method. While the results are most pronounced after 10,000 spurious messages, notice that this would not be uncommon, and is precisely the scenario we aim to protect against. In a sample of 3 million messages, it would be quite easy for a new trend (e.g., "Bieber fever") to be introduced into Twitter and spread rapidly.

Second, hard classification appears to do a better job than soft classification, most likely because removing any document with probability below 0.5 results in a much more aggressive filter. In contrast, the soft classifiers allows a large number of spurious messages to influence the results, even if they all have low classification scores.

Third, it is clear that none of the methods are completely adequate under extreme conditions. An additional 100,000 spurious messages overwhelms all approaches and produces an invalid spike in ILI estimates. Since no classification method will ever achieve 100% accuracy, it is difficult to guard against such extreme cases. We leave investigation of further improvements for future work.

7 Predicting the peak of flu season

While Figures 2 and 3 appear to predict well during the tail end of flu season, we are more interested in predicting the peak of the season. Unfortunately, the collected data only spans one flu season. To simulate predicting during the peak time, we simply reverse the order of the weeks and regenerate Figures 2 and 3. That is, the linear model is trained on the final 20 weeks of flu season and tested on the first 16 weeks. This experiment is obviously chronologically unrealistic, but it does provide insight into system accuracy during peak season.

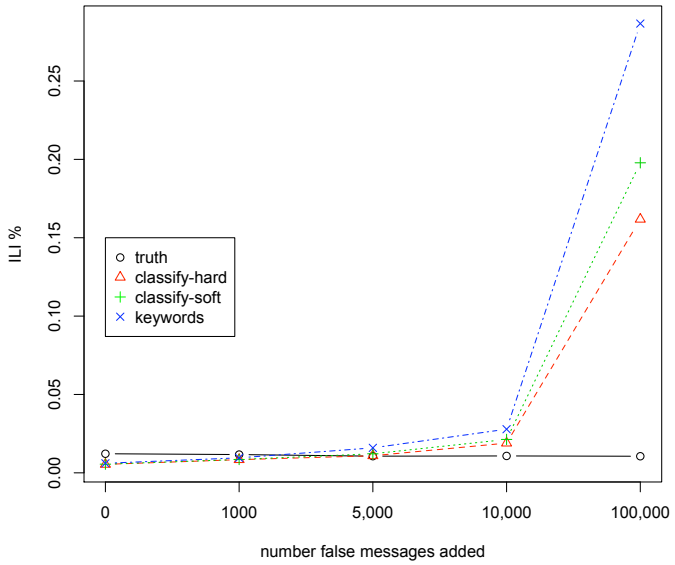

Fig. 6 Results of the false alarm simulation experiment. Corresponding mean-squared errors are **keywords=.077, classify-soft=.035, classify-hard=.023**. Hard classification appears to be more robust than soft classification in the presence of spurious messages, although all approaches are overwhelmed when the number of spurious messages reaches 100,000.

Figures 7 and 8 mostly replicate the results in Figures 2 and 3, with a few notable exceptions. First, the removal of the keywords *swine* and *h1n1* appears even more critical to accurately predicting the peak of the season. For example, using only the keyword *flu* correlates 0.9 on the testing set, whereas removing *swine* and *h1n1* improves this further ($r(14) = .964$, $p < .001$). Secondly, *flu +vaccine* surprisingly exhibits a very high testing correlation (0.97). However, examining the graph in Figure 8 shows that while there is a strong correlation, the mean squared error is much larger than *flu - (swine | h1n1)* (.0007084 vs .000036).

In summary, it appears that the query *flu -(swine | h1n1)* is the most reliable in terms of training fit and testing prediction on both the original dataset and the reversed dataset.

8 Support Vector Regression

Thus far, we have used a standard linear regression model to fit the ILI data. We have chosen this model because of its simplicity, its success in related work [12], and because it allows us to measure the utility of the Twitter signal in isolation. However, numerous more advanced regression methods exist. In this section, we consider one such advanced technique, Support Vector Regression (SVR) [8]. Signorini et al. [31] have had recent success modeling influenza rates using SVRs. However, in our experiments, SVRs do not exhibit any significant improvements over simple linear regression.

We use ϵ-SVR with a radial-basis kernel, as implemented in LibSVM [2]. (Please see Drucker et al. [8] for full details of the algorithm.) As in the previous section,

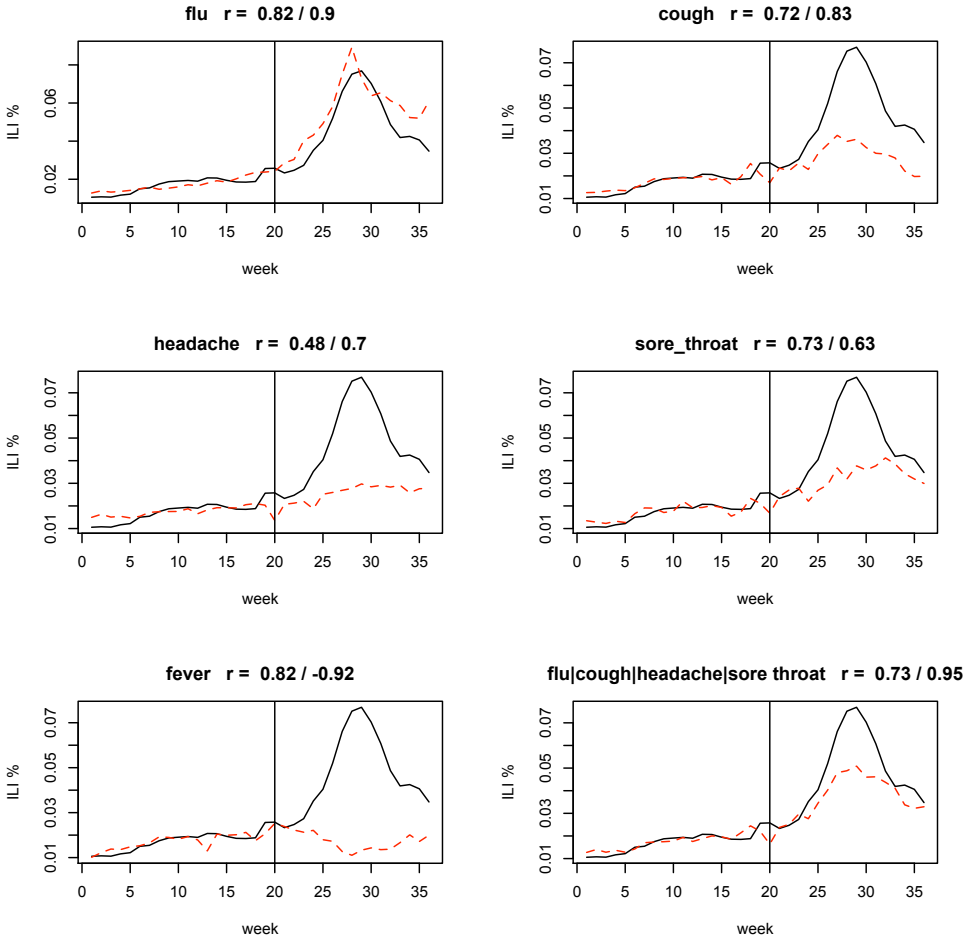

Fig. 7 Fitted and predicted ILI rates using regression over query fractions of Twitter messages. *The weeks have been reversed from those in Figure 2.* Black (solid) lines are the true ILI rates, red (dotted) lines are the fitted/predicted values, with the vertical line splitting the training and evaluation data.

we again train on the final 20 weeks and predict on the first 16 to predict the peak of flu season. We use the grid search optimization procedure in LibSVM to optimize SVR parameters C, ϵ (complexity parameters) and γ (the width of the radial basis kernel). This optimization was done on the training set. (We observe that this optimization can have a noticeable impact on model quality.)

The ϵ-SVR model learns weights for each input variable, similar to linear regression. We consider three different sets of inputs for the ϵ-SVR model:

– `svr:flu-no-swine`: The single input variable corresponds to the *flu -(swine | h1n1)* query fraction used in the experiments in Figure 8. Thus, the only difference from the previous linear regression model is the use of ϵ-SVR to determine the regression parameters.

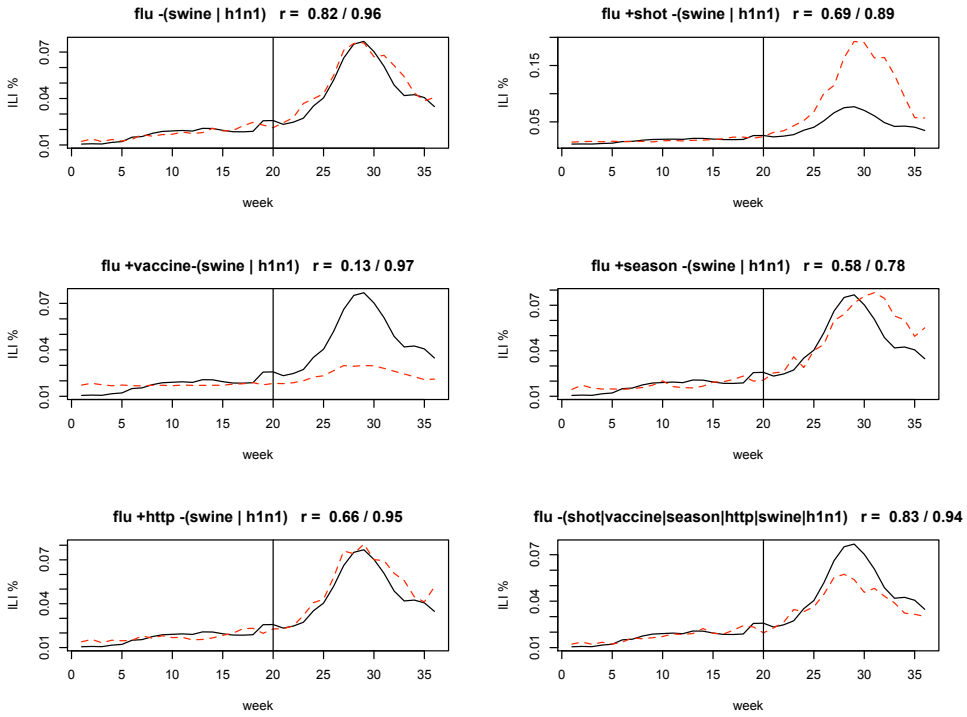

Fig. 8 Fitted and predicted ILI rates using regression over query fractions of Twitter messages. *The weeks have been reversed from those in Figure 3.* Black (solid) lines are the true ILI rates, red (dotted) lines are the fitted/predicted values, with the vertical line splitting the training and evaluation data.

Table 7 Comparison between ϵ-SVR and linear regression. The mean-squared error of the fourth system is significantly worse than the other three using a Wilcoxan rank sum test ($p < 0.05$). The remaining differences are not statistically significant.

Method	Correlation	Mean-squared error
lr:flu-no-swine	.964	.000036
svr:flu-no-swine	.963	.000090
svr:flu-no-swine-cough	.957	.000038
svr:flu-no-swine-cough-sore-throat	.763	.000616

- `svr:flu-no-swine-cough`: We add an additional regression variable for the query fraction of the term *cough*.
- `svr:flu-no-swine-cough-sore-throat`: We add an additional regression variable for the query fraction of the term *sore throat*.

In Figure 9 and Table 7, these models are compared with the original linear regression model (`lr:flu-no-swine`).

Our first observation is that ϵ-SVR provides no improvement in either correlation or mean-squared error over simple linear regression. We hypothesize that due to the small number of training instances (20) and small number of input variables (1-3), the additional complexity of SVRs is not necessary. Second, we note that performance drops substantively when a third input variable (sore throat) is

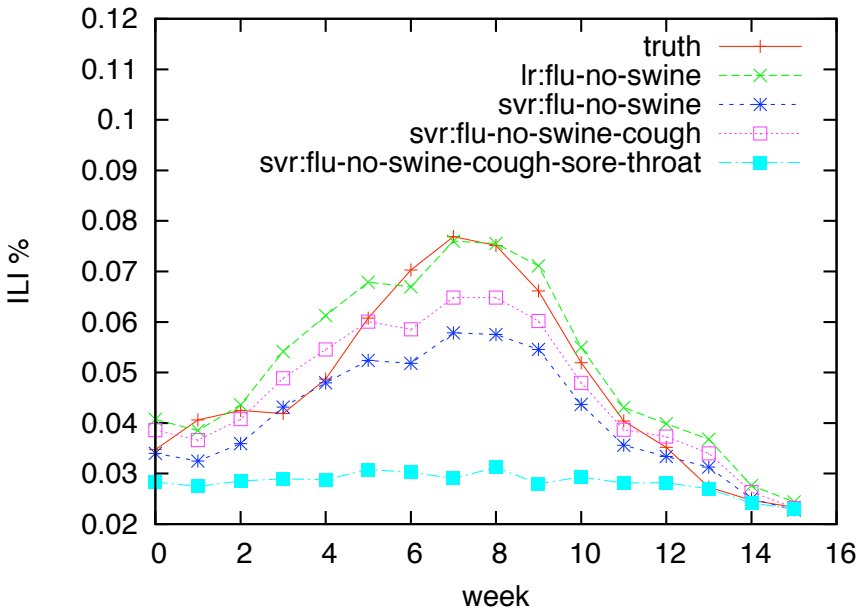

Fig. 9 Experiments comparing linear regression with ϵ-SVR, with various input variables. ϵ-SVR shows no improvement over linear regression (`lr:flu-no-swine`). The models were all trained on weeks 20-36 and tested on weeks 1-16 (shown).

added. We again attribute this to the small number of training examples. As the number of parameters increases, so does the likelihood of overfitting the training data. We conclude from these experiments that simple linear regression is adequate when the number of training instances is small.

9 Alcohol Sales Volume

In this section, we consider applying the previously described methodology in a new domain: alcohol sales. We choose this domain for three reasons: (1) The U.S. Census Bureau tracks monthly sales volume[6], which we use as validation data; (2) understanding patterns and causes of alcohol consumption is an important mission of the National Institute on Alcohol Abuse and Alcoholism (NIAAA); (3) alcohol-related messages are common on social networking sites.

Excessive alcohol consumption is the third leading preventable cause of death in the United States, accounting for approximately 15,000 deaths per year [22, 16]. It is therefore necessary for researchers to understand the causes and patterns of dangerous consumption levels; however, collecting such data can be difficult. This difficulty can even be observed in the extensive Behavioral Risk Factor Surveillance System (BRFSS)[7], the world's largest telephone health survey system, conducted by the U.S. Centers for Disease Control and Prevention (CDC). Not only does

[6] `http://www.census.gov/retail/`
[7] `http://www.cdc.gov/brfss`

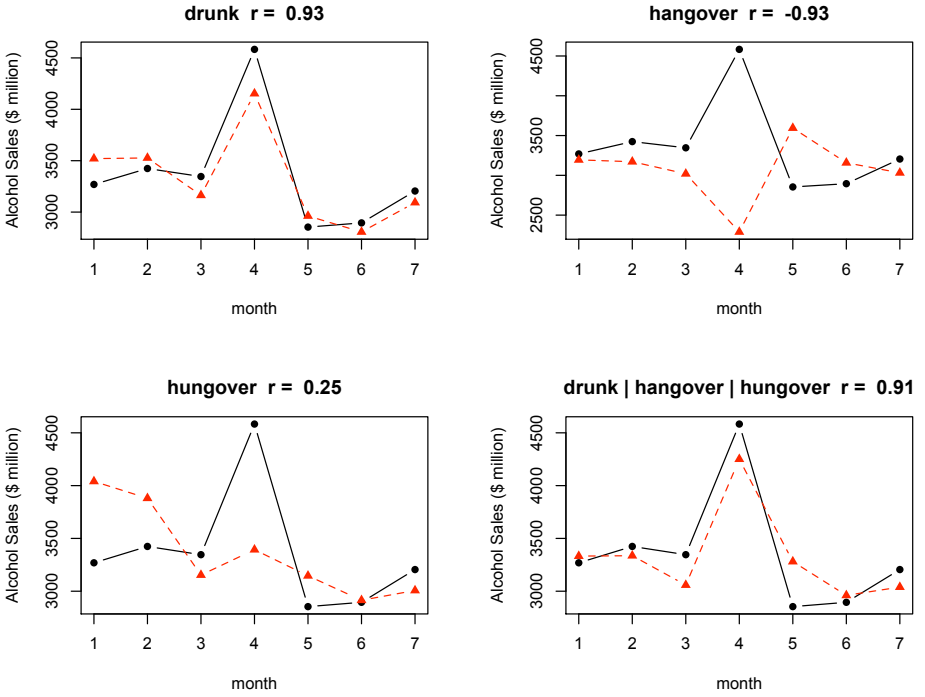

Fig. 10 Results of alcohol sales volume estimation. The black (solid) line is the validation data from the U.S. Census, the red (dotted) line is the estimated value, computed using linear regression with leave-one-out estimation.

the BRFSS survey suffers from low response rates (52.5%) [16], but, perhaps more importantly, BRFSS is likely to greatly underestimate binge drinking in arguably the most critical demographic: 18–34 year olds. This is because BRFSS does not collect data from people living in institutional settings (e.g., college dormitories), and only collects data through landlines, which few people in that age range possess[8]. In this section, we consider the viability of Twitter as an alternate source of such statistics.

After a brief review of alcohol-related Twitter messages, we selected the keywords (*drunk* or *hangover* or *hungover*) as potential indicators of alcohol consumption. We searched for these keywords in Twitter messages from September 2009 to March 2010, and we collected the sales volume for each month in that time span from the U.S. Census. We selected the unadjusted monthly sales volume for business categorized as "Beer, wine, and liquor stores" (NAICS code 4453).

Because we are correlating Twitter mentions of alcohol *consumption* with alcohol *sales*, it is to be expected that sales values will pre-date Twitter-derived consumption estimates. That is, people must first purchase the alcohol before they consume it. Thus, we introduce a 7-day lag in the Twitter estimation. For example, the Twitter messages mined to compute an estimate for sales volumes in October span from the second week in October through the first week in November.

[8] The complementary Youth Risk Behavior Surveillance System only partially solves this problem, since it is restricted to surveys of high school students (`http://www.cdc.gov/yrbs`).

Table 8 Alcohol volume estimates using leave-one-out training for the model using the single keyword "drunk". **Msgs** is the total number of Twitter messages collected per month, and **"Drunk"** is the number of messages matching the keyword "drunk". Note that the 7-day lag is reflected in these values.

Month	Msgs	"Drunk"	Sales ($ mil)	Estimate	Error
9/09	50,242,252	54,143	3,269	3,520	251
10/09	56,533,746	61,324	3,424	3,527	103
11/09	56,159,982	56,466	3,346	3,164	-182
12/09	57,637,948	76,830	4,583	4,153	-430
1/10	83,504,599	78,631	2,854	2,962	108
2/10	66,261,459	60,686	2,895	2,807	-88
3/10	90,300,192	88,997	3,205	3,093	-112

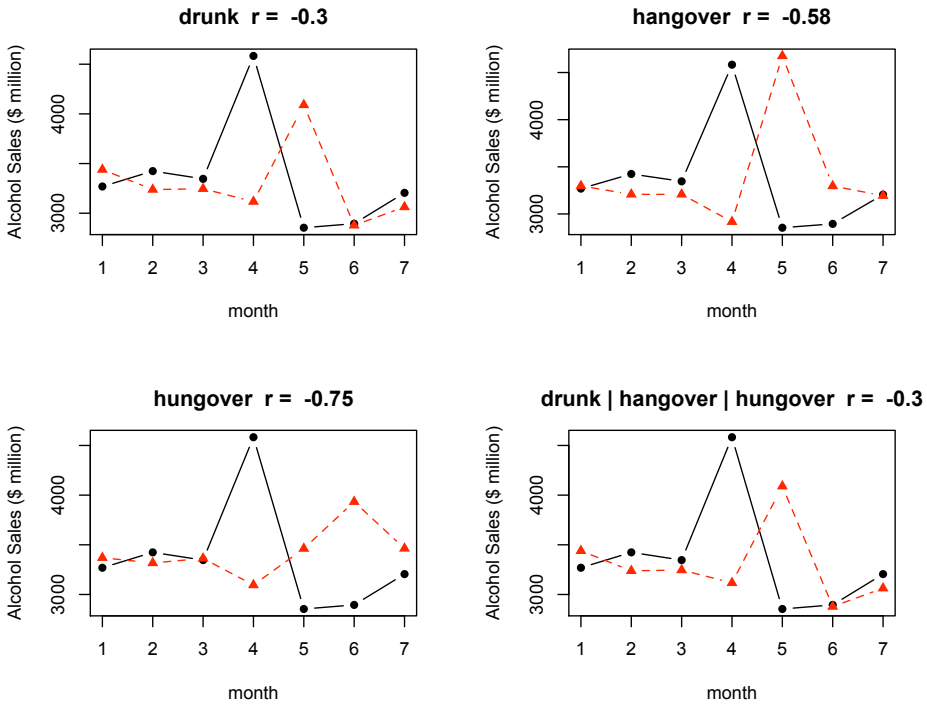

Fig. 11 Results of alcohol sales volume estimation **without** the 7-day lag. The black (solid) line is the validation data from the U.S. Census, the red (dotted) line is the estimated value, computed using linear regression with leave-one-out estimation.

As the U.S. Census only computes sales volume once per month, there are 7 validation points. To evaluate the model, we use leave-one-out training to learn the linear regression model. That is, the model prediction for month m is obtained by first fitting the model on all months other than m, then predicting m.

Figure 10 compares the model estimates with the U.S. Census statistics for the 7-month span. The black (solid) line is the validation data from the U.S. Census, the red (dotted) line is the estimated value. We run experiments for each keyword in isolation as well as their union. The most accurate estimate uses only the keyword *drunk* ($r(5) = .932$, $p < .01$). Table 8 provides details on this model.

We note that the keyword *hangover* produces poor estimates. We believe this is in part due to the motion picture *The Hangover*, which was released in June 2009. We observed a number of Twitter messages related to it, and it is possible that a filtering stage may help here.

Figure 11 shows the need for the 7-day lag. Without it, the predictions are clearly shifted forward. This is especially problematic for December 2009-January 2010, which spans New Year's Eve, a lucrative time for liquor stores.

In summary, it appears that tracking a single Twitter keyword (*drunk*) is a viable way to estimate alcohol sales. While the estimates lag by one week, we believe that these positive results suggest that Twitter may be a useful source for researchers to explore trends in alcohol consumption.

10 Related Work

There has been a number of recent papers related to Twitter and influenza [17, 7, 31, 4]. Lampos & Cristianini [17] perform a similar analysis of Twitter message to track influenza rates in the U.K. They learn a "flu-score" for each document by learning weights for each word by their predictive power on held-out data. Using a set of 41 hand-chosen "markers", or keywords, they obtain a correlation of .92 with statistics reported by the U.K.'s Health Protection Agency. Additionally, they obtain a correlation of .97 by using automated methods to select additional keywords to track (73 in total), similar to the methodology of Ginsberg et al. [12]. In this paper, we have presented a simpler scheme to track flu rates and have found a comparable level of correlation as in Lampos & Cristianini [17]. A principal distinction of this paper is our attempt to address the issue of false alarms using supervised learning.

In an earlier version of our work [7], we perform a similar analysis as in Lampos & Cristianini [17], also experimenting with automated methods to select keywords to track. We also report an improved correlation in simulation experiments by using a classifier to filter spurious messages. In this paper, we have used similar techniques on a much larger dataset (570 million vs. 500K). We have also more closely evaluated the impact of false alarms on these types of methods.

This paper, as well as Lampos & Cristianini [17] and Culotta [7], are similar in methodology to Ginsberg et al. [12], who track flu rates over five years by mining search engine logs, obtaining a .97 correlation with ILI rates on evaluation data. Thus, while estimating influenza rates from the Web is not new, extending these methods to Twitter, blogs, and other publicly available resources has the benefit of measuring how different data sources affect the quality of predictions. It also allows us to study how vulnerable different data sources are to spurious matches, which is critical to deploying this technology.

Corley et al. [6] track flu rates by examining the proportion of blogs containing two keywords (*influenza* and *flu*), obtaining a correlation of .76 with true ILI rates. It is possible that the brevity of Twitter messages make them more amenable to simple keyword tracking, and that more complex methods are required for blog data.

Classifying Twitter messages is often a part of *sentiment mining* [26]. This work is similar to research that has shown strong correlations between sentiment

mined from online media and other external values such as stock prices [18,11], product sales [14,11], mood levels [23], and political polls [24].

11 Conclusions and Future Work

In this paper, we have provided evidence that relatively simple approaches can be used to track influenza rates and alcohol sales volume from a large number of Twitter messages, exhibiting a strong correlation to held-out data. We have also proposed a supervised learning approach to reduce the burden of false alarms, and through simulation experiments we have measured the robustness of this approach. These results suggest that while document classification can greatly limit the impact of false alarms, further research is required to deal with extreme cases involving a large number of spurious messages.

We believe this line of research has the potential to aid the tracking of a number of public health statistics using less money and time than existing approaches. Given the time-critical nature of such statistics, accurate, real-time tracking provides considerable benefit for minimizing health risks.

While the data show that this simple approach works well, in future work we plan to investigate methods to improve its accuracy using more sophisticated natural language processing approaches to document filtering. Additionally, we will investigate whether temporal models can allow us to forecast farther into the future than our current approach.

Acknowledgements We would like to thank Brendan O'Connor from Carnegie Mellon University for providing access to the Twitter data and Troy Kammerdiener of Southeastern Louisiana University for helpful discussions in early stages of this work. This work was supported in part by a grant from the Research Competitiveness Subprogram of the Louisiana Board of Regents, under contract #LEQSF(2010-13)-RD-A-11.

References

1. Brownstein, J., Freifeld, C., Reis, B., Mandl, K.: Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Medicine **5**, 1019–1024 (2008)
2. Chang, C., Lin, C.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3) 27:1–27:27 (2011) *Software available at* `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
3. Chang, C., Lin, C.: Training ν-Support Vector Regression: Theory and Algorithms. Neural Computation **14**(8), 1959–1977 (2002)
4. Chew, C., Eysenbach, G.: Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE **5**(11), (2010)
5. Collier, N., Doan, S., Kawazeo, A., Goodwin, R., Conway, M., Tateno, Y., Ngo, H.Q., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K.: BioCaster: detecting public health rumors with a web-based text mining system. Bioinformatics (2008)
6. Corley, C., Cook, D., Mikler, A., Singh, K.: Text and structural data mining of influenza mentions in web and social media. International Journal of Environmental Research and Public Health **7**(2), 596–615 (2010)
7. Culotta, A.: Towards detecting influenza epidemics by analyzing twitter messages. In: Workshop on Social Media Analytics at the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2010)
8. Drucker, H., Burges, C., Kaufman L., Smola A., Vapnik V.: Support Vector Regression Machines. In: Advances in Neural Information Processing Systems 9, pp. 155–161. (1996)

9. Eysenbach, G.: Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In: AMIA: Annual symposium proceedings, pp. 244–248 (2006)
10. Giampiccolo, D., Magnini, B., Dagan, I., Dolan, B.: The third pascal recognizing textual entailment challenge. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 1–9. Prague (2007)
11. Gilbert, E., Karahalios, K.: Widespread worry and the stock market. In: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, D.C. (2010)
12. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457** (2009)
13. Grishman, R., Huttunen, S., Yangarber, R.: Information extraction for enhanced access to disease outbreak reports. Journal of Biomedical Informatics **35**(4), 236–246 (2002)
14. Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp. 78–87. ACM, New York, NY, USA (2005)
15. Johnson, H., Wagner, M., Hogan, W., Chapman, W., Olszewski, R., Dowling, J., Barnas, G.: Analysis of web access logs for surveillance of influenza. MEDINFO pp. 1202–1206 (2004)
16. Kanny, D., Liu, Y., Bewer, R.: Binge drinking - united states, 2009. *Morbidity and Mortality Weekly Report*, **60**(01):pp. 101–104 (2011).
17. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: 2nd IAPR Workshop on Cognitive Information Processing (CIP 2010), pp. 411–416 (2010)
18. Lavrenko, V., Schmill, M.D., Lawrie, D., Ogilvie, P., Jensen, D., Allan, J.: Language models for financial news recommendation. In: Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM). Washington, D.C. (2000)
19. Linge, J., Steinberger, R., Weber, T., Yangarber, R., van der Goot, E., Khudhairy, D., Stilianakis, N.: Internet surveillance systems for early alerting of health threats. Eurosurveillance **14**(13) (2009)
20. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Programming **45**(3, (Ser. B)), 503–528 (1989)
21. Mawudeku, A., Blench, M.: Global public health intelligence network (GPHIN). In: 7th Conference of the Association for Machine Translation in the Americas (2006)
22. McGinnis, J., Foege, W.: Actual causes of death in the united states. *JAMA*, **270**:2207–2012, (1993).
23. Mishne, G., Balog, K., de Rijke, M., Ernsting, B.: MoodViews: Tracking and searching mood-annotated blog posts. In: International Conference on Weblogs and Social Media. Boulder, CO (2007)
24. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From Tweets to polls: Linking text sentiment to public opinion time series. In: International AAAI Conference on Weblogs and Social Media. Washington, D.C. (2010)
25. Oreskovic, A.: Twitter snags over 100 million users, eyes money-making. Reuters (2010)
26. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1–2), 1–135 (2008)
27. Polgreen, P., Chen, Y., Pennock, D., Forrest, N.: Using internet searches for influenza surveillance. Clinical infectious diseases **47**, 1443–1448 (2008)
28. de Quincey, E., Kostkova, P.: Early warning and outbreak detection using social networking websites: the potential of twitter, electronic healthcare. In: eHealth 2nd International Conference. Instanbul, Turkey (2009)
29. Reilly, A., Iarocci, E., Jung, C., Hartley, D., Nelson, N.: Indications and warning of pandemic influenza compared to seasonal influenza. Advances in Disease Surveillance **5**(190) (2008)
30. Ritterman, J., Osborne, M., Klein, E.: Using prediction markets and Twitter to predict a swine flu pandemic. In: 1st International Workshop on Mining Social Media (2009)
31. Signorini, A., Polgreen, P. M., Segre, A.M.: The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic PLoS ONE, **6**:(5) (May 4, 2011).