Towards detecting influenza epidemics by analyzing Twitter messages

Aron Culotta
Southeastern Louisiana University
Department of Computer Science
Hammond, LA 70402
culotta@selu.edu

ABSTRACT

Rapid response to a health epidemic is critical to reduce loss of life. Existing methods mostly rely on expensive surveys of hospitals across the country, typically with lag times of one to two weeks for influenza reporting, and even longer for less common diseases. In response, there have been several recently proposed solutions to estimate a population's health from Internet activity, most notably Google's Flu Trends service, which correlates search term frequency with influenza statistics reported by the Centers for Disease Control and Prevention (CDC). In this paper, we analyze messages posted on the micro-blogging site Twitter.com to determine if a similar correlation can be uncovered. We propose several methods to identify influenza-related messages and compare a number of regression models to correlate these messages with CDC statistics. Using over 500,000 messages spanning 10 weeks, we find that our best model achieves a correlation of .78 with CDC statistics by leveraging a document classifier to identify relevant messages.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms

Algorithms, Experimentation, Measurements

Keywords

data mining, regression, classification, social media

1. INTRODUCTION

There has been growing interest in monitoring disease outbreaks using the Internet, typically either by mining newspaper articles mentioning flu illnesses [9, 18, 1, 23, 3, 15], mining the frequency of visits to health-related websites [12] or mining search engine logs for flu-related queries [6, 22, 8]. 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.

Initial work in this direction includes Ritterman et al. [24], 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 and Kostova [5] 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. However, to our knowledge, no one has measured the correlation between patterns in Twitter messages and national health statistics.

In this paper, we investigate several models to analyze Twitter messages in order to predict rates of influenza-like illnesses in a population. 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]

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). To alleviate this burden, Ginsberg et al.[8] show that an accurate regression model of ILI rates can be estimated using the proportion of flu-related queries submitted to the Google search engine over the same time period. Central to the approach is a method to select which keywords to monitor, done by computing the correlation coefficient of each keyword on held-out data.

While the model in Ginsberg et al. exhibits a compelling correlation with CDC statistics over a long time period, there are several reasons to consider a model based on Twitter messages:

- Full messages provide domain experts with more descriptive information than queries to characterize an outbreak.
- Twitter profiles often contain semi-structured metadata (city, state, gender, age), enabling a more detailed statistical analysis. (Note that techniques to infer demographics from search history [27] could be applied here as well. It is worth investigating whether

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

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

the predictions will be more accurate using Twitter messages.)

- Ginsberg et al. note that "an unusual event, such as a drug recall ... could cause ... a false alert." Indeed, the final queries used in their regression model are not published over concerns that "users may be inclined to submit some of the queries out of curiosity, leading to erroneous future estimates of the ILI percentage." While our proposed method is certainly not immune to false alerts, we demonstrate that a simple document classification algorithm can filter out many erroneous messages.

- 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]

- All our data sources are publicly available, enabling reproducibility and follow-on research.

We collect over 500,000 Twitter messages from a 10 week period and develop several regression models that predict ILI rates based on the frequency of messages containing certain keywords. We find the following results:

1. Aggregating keyword frequencies using separate predictor variables (i.e., multiple linear regression) outperforms aggregating keyword frequencies into a single predictor variable (i.e., simple linear regression).

2. Selecting keywords based on residual sum of squares appears to be more effective than selecting keywords based on correlation coefficient.

3. A simple bag-of-words classifier trained on roughly 200 documents can effectively filter erroneous document matches, resulting in better model predictions. This final model achieves a .78 correlation with CDC statistics over 5 weeks of validation data.

2. MODELING INFLUENZA RATES

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.

We propose estimating P from $Q(W, D)$ using linear regression. We consider several competing systems, which vary based on the number of regression coefficients, how the keywords W are chosen, and whether matching documents D_W are filtered for false matches.

2.1 Regression Models

We consider two common regression models, Simple Linear Regression and Multiple Linear Regression.

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

2.1.1 Simple Linear Regression

Following Ginsberg et al.[8], we first consider a simple linear model between the log-odds of P and $Q(W, D)$:

$$logit(P) = \beta_1 logit(Q(W, D)) + \beta_2 + \epsilon \quad (1)$$

with coefficients β_1, β_2, error term ϵ, and logit function $logit(X) = \ln(\frac{X}{1-X})$.

In this model, there is a single Q variable. We refer to this as *Simple Linear Regression*.

2.1.2 Multiple Linear Regression

Because W contains more than one keyword, it is natural to consider expanding Simple Linear Regression to include separate parameters for each element of W. This results in the *Multiple Linear Regression* model:

$$logit(P) = \beta_1 logit(Q(\{w_1\}, D)) + \dots + \beta_k logit(Q(\{w_k\}, D)) + \beta_{k+1} + \epsilon \quad (2)$$

where $w_i \in W$. Ginsberg et al.[8] found Multiple Regression to overfit in initial experiments, and so only report results using Simple Linear Regression. We investigate whether this holds for Twitter data as well.

Both Simple and Multiple regression models are fit using ridge regression with regularization parameter of $1.0E-5$.

2.2 Keyword Selection

We now turn to the problem of selecting W, the keywords likely to be contained in messages reporting ILI symptoms. We consider two methods of generating W, which will determine the value of $Q(W, D)$ used in each of the regression models.

2.2.1 Selection by Correlation Coefficient

We adopt the method of Ginsberg et al.[8], which creates W from a larger set W' by iteratively adding terms in order of their correlation with held-out validation data. For each word, we fit a Simple Linear Regression model to observed data using only that word as W. We then evaluate each model on validation points, from which we can calculate the correlation between the predicted and true values. Keywords are then added to W in order of decreasing correlation.

In more detail, given a set of n ILI values $\mathbf{P} = \{P_1 \dots P_n\}$ used for training data, we estimate the correlation of each term $w_i \in W'$ as follows:

1. Loop for $j = 1$ to n:

   (a) Reserve P_j for validation, and fit a Simple Linear Regression Model on the remaining points $\mathbf{P} \setminus P_j$ using $W = \{w_i\}$.

   (b) Predict $\hat{P}_j$ using the learned regression model.

2. Collect all predicted $\hat{P}_j$ values into $\hat{\mathbf{P}}$, and compute the Pearson correlation coefficient between $\hat{\mathbf{P}}$ and $\mathbf{P}$.

Note that it is possible to perform an additional cross-validation procedure to determine when to stop adding terms to W, but in this work we simply report results over a range of sizes for W.

2.2.2 Selection by Residual Sum of Squares

We repeat the same term selection procedure as above, but replace the Pearson correlation coefficient with the residual sum of squares (RSS):

$$RSS(\mathbf{P}, \hat{\mathbf{P}}) = \sum_i (P_i - \hat{P}_i)^2 \quad (3)$$

2.3 Keyword Generation

The Keyword Selection techniques in the previous section first require a larger set of keywords W' from which to select. We propose two methods to generate the initial candidate list of words W'.

2.3.1 Hand-chosen keywords

We first consider a simple set of four keywords consisting of {*flu, cough, sore throat, headache*}[3]. Besides the word *flu*, the additional three terms are commonly known flu symptoms.

2.3.2 Most frequent keywords

To expand this candidate set, we search for all documents containing any of the hand-chosen keywords. We then find the top 5,000 most frequently occurring words in the resulting set. The idea is to seed the selection process with terms that are likely to correlate with ILI rates without manually tuning this set.

2.4 Document Filtering

As Ginsberg et al.[8] point out, a principal drawback of using keyword matching to calculate Q is that it is susceptible to fluctuations in term frequency that are unrelated to influenza. For example, the keyword *Tylenol* may be a valuable term, but the recent recall of several *Tylenol* products led to a spike in its term frequency, without a corresponding spike in ILI rates.

We propose a first-step to mitigate this problem by training a classifier to label whether a message is reporting an ILI-related event or not. This is related to problems such as *sentiment analysis* [21] and *textual entailment* [7], 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 of Section 2.3.1. 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:

$$p(y_i = 1|\mathbf{x}_i; \theta) = \frac{1}{1 + e^{(-\mathbf{x}_i \cdot \theta)}} \quad (4)$$

We learn θ using L-BFGS gradient descent [16] as implemented in the MALLET machine learning toolkit[4]. We use the default regularization parameter $\Lambda = 1$.

[3]Note that the "sore throat" query is treated as a conjunction of terms, not a phrase.

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

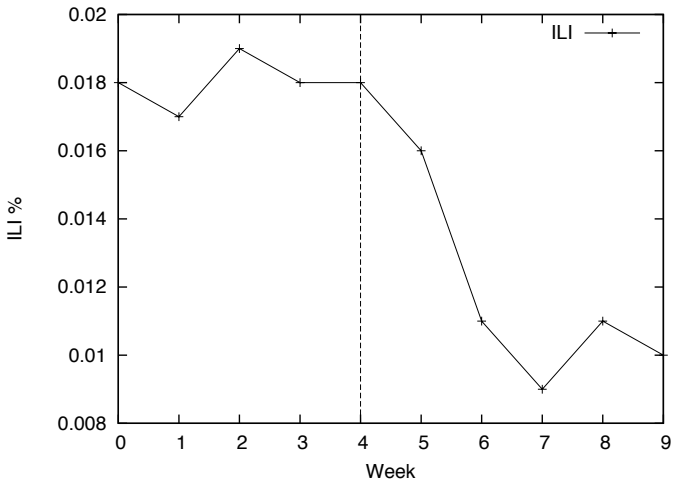

Figure 1: Ground truth ILI rates, as obtained from the CDC's weekly tracking statistics. Weeks 0-4 are used for training; weeks 5-9 are used for testing.

Number of messages	574,643
Number of tokens	12.3 million
Vocabulary size	582,419
Average number of tokens per message	21.4

Table 1: Characteristics of collected Twitter messages.

We incorporate the classifier into a Simple Regression Model by defining the *expected fraction* of positively classified documents as

$$Q_p(W, D) = \frac{\sum_{d_i \in D_W} p(y_i = 1|\mathbf{x}_i; \theta)}{|D|} \quad (5)$$

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. With an accurate classifier, we can reduce the impact of erroneous matches by limiting their impact on the computation of $Q_p(W, D)$.

3. EXPERIMENTS

Below we describe the data and systems used for document filtering and ILI rate prediction.

3.1 Data

We collected 574,643 Twitter messages for the 10 weeks from from February 12, 2010 to April 24, 2010 (Table 1). Note that we are restricted to messages that are set to "public" visibility by the poster. In order to obtain a pseudo-random sample of messages, we searched for messages containing any of the following common words: {a, I, is, my, the, to, you}. We note that this is in contrast to de Quincey and Kostova [5], who search specifically for flu-related Twitter messages. In order to regress with ILI rates, which are percentages, we need to estimate the percentage of Twitter messages that report an ILI. This requires a reliable denominator that counts all Twitter messages.

Corresponding ILI rates were obtained directly from the

Positive Examples
Headace, 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 swin. #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 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 2.4.

Accuracy	F1	Precision	Recall
84.29 (1.9)	90.2 (1.5)	92.8 (1.8)	88.1 (2.0)

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

CDC website[5]. The true ILI values are displayed in Figure 1, with a horizontal line indicating the split between training and testing points.

To generate labeled data for the classification model of Section 2.4, we searched for Twitter messages containing any of the four hand-chosen keywords defined in Section 2.3.1, making sure the messages were posted outside of the date range of the previously collected messages. (The classification messages were collected from April 25 to May 4.) This resulted in 206 messages, which were manually categorized into 160 positive examples and 46 negative examples. (Examples are shown in Table 2.)

3.2 Systems

Table 4 displays the 10 different systems we evaluate. For the systems that use keyword selection, we indicate how many words it selects in parentheses. For example `simple-hand-rss(2)` selects two keywords to compute Q.

It is important to note that the classification methods (`classification-hand`, `classification-freq`) use the *entire* set of respective keywords to identify candidate messages, which are then assigned confidence values by the classifier to calculate $Q_P(W, D)$ as in Equation 5.

4. RESULTS

Below we report both the accuracy of the document filter as well as the correlation and residuals of the regression models.

4.1 Document Filtering

We first evaluate the accuracy of the message classifier. We perform 10-fold cross validation on the 206 labeled messages, obtaining a mean accuracy of 84.29% (Table 3). Inspecting the learned parameters of the model reveals large positive weights assigned to words like *headache* (0.88), *n* (0.45) (a common abbreviation for *and*), *have* (0.40), *throat* (0.29), and *sick* (0.24). Note that because all the labeled examples already match one of {*flu, cough, sore throat, headache*}, the classifier is learning only to classify this subset of documents. It would not be expected to perform well on messages in general. This explains why common words like *n* and *have* receive high weights, as in "cough **n** sneezin" and "I **have** a headache".

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

4.2 Regression

Table 5 compares the results of each of the 10 systems. Note that only the best performing keyword selection systems are displayed in this table. More detailed results are displayed in subsequent figures. We make the following four observations of the results:

(1) Multiple Regression outperforms Simple Regression. Contrary to results reported in Ginsberg et al.[8], simple linear regression performs much worse than multiple linear regression. In fact, only one simple linear regression model produces a positive correlation with the testing data, and that model uses only a single keyword (*flu*). Adding a disjunction of additional terms only reduces the effectiveness of the simple regression model.

A possible explanation for why the reverse was found for the query data in Ginsberg et al. is that queries exhibit less ambiguity (e.g., "flu symptoms", "influenza symptoms"). In contrast, individual keywords can be highly ambiguous (e.g., "cough", "sick") and can have frequent idiomatic uses (e.g., "cough it up", "I'm sick of politics"). Combining these disparate queries into one fraction $Q(W, D)$ likely results in a noisy estimate of true ILI reports.

Multiple regression can alleviate this problem by placing separate weights on each keyword. For example, system *multi-hand-rss(2)* selects keywords *flu* and *sore throat* with coefficients 498 and -190, respectively.

Despite the fact that multiple regression outperforms simple regression, multiple regression does indeed begins to overfit when too many keywords are added. In Figure 8, we can see a dramatic drop in correlation after five keywords are added. This is not unexpected, as the number of parameters for the multiple regression model (6) is now greater than the number of training points (5).

(2) Keyword selection is prone to overfitting. The best performing keyword selection system that selected from the large set of 5,000 terms is `multi-freq-rss(3)`, which achieves a correlation of .582. However, inspecting this system reveals that it selects the terms *net, rely, leave*, which do not have any obvious connection to influenza. The term *net* occurs most frequently with hyperlinks and *leave* occurs most frequently in phrases like *I have to leave*. Examining the training correlation and RSS reveals that these models that can choose from such a large set of keywords often over-

	Regression	Classification	Keyword Generation	Keyword Selection
`classification-hand`	Simple	yes	hand-chosen	no
`classification-freq`	Simple	yes	most-frequent	no
`simple-hand-rss`	Simple	no	hand-chosen	RSS
`simple-hand-corr`	Simple	no	hand-chosen	correlation
`simple-freq-rss`	Simple	no	most-frequent	RSS
`simple-freq-corr`	Simple	no	most-frequent	correlation
`multi-hand-rss`	Multiple	no	hand-chosen	RSS
`multi-hand-corr`	Multiple	no	hand-chosen	correlation
`multi-freq-rss`	Multiple	no	most-frequent	RSS
`multi-freq-corr`	Multiple	no	most-frequent	correlation

Table 4: The 10 different systems evaluated. Each system varies by the type of regression model, whether it uses classification, the method to generate candidate keywords, and the method of selecting keywords from that candidate set.

	Train			Test		
system	r	RSS	$\hat{\sigma}$	r	RSS	$\hat{\sigma}$
`classification-hand`	.838	5.96e-7	4.4e-4	.780	2.47e-4	.00907
`classification-freq`	.742	8.96e-7	5.5e-4	-.396	3.24e-4	.01040
`simple-hand-rss(1)`	.125	1.97e-6	8.10e-4	.498	2.51e-4	.00914
`simple-freq-rss(8)`	.997	1.10e-8	6.07e-5	-.034	2.17e-4	.00850
`simple-hand-corr(4)`	.186	1.93e-6	8.02e-4	-.761	2.42e-4	.00899
`simple-freq-corr(8)`	.997	1.10e-8	6.07e-5	-.034	2.17e-4	.00850
`multi-hand-rss(2)`	.703	1.01e-6	5.81e-4	.739	2.78e-4	.00914
`multi-freq-rss(3)`	.998	6.10e-9	4.51e-5	.582	1.90e-4	.00710
`multi-hand-corr(1)`	.858	5.28e-7	4.19e-4	-.911	2.51e-4	.00914
`multi-freq-corr(3)`	.998	6.10e-9	4.51e-5	.582	1.90e-4	.00710

Table 5: Pearson's regression coefficient (r), residual sum of squares (RSS), and standard regression error ($\hat{\sigma}$) of each system. For each of the keyword selection systems, we have displayed the system with the highest correlation on the testing data, with the number of keywords chosen shown in parentheses. Additional results by number of keywords are displayed in Figures 6-9.

fit – they obtain the highest training correlation and lowest training RSS. To obtain more reliable results from keyword selection, it is likely necessary to obtain a greater number of training points.

(3) Classification appears to be an effective method of removing erroneous messages. The system achieving the highest correlation on the testing data is `classification-hand` (.78). Comparing this result to that of `simple-hand-rss(1)` (.498) suggests that using the classifier to compute Q results in a more reliable indicator of ILI rates.

Inspecting the classifier predictions on the unlabeled data shows that the most common types of errors involve a colloquial use of "cough", as in "U know that ability where you re able to sense if someone is engaged in a conversation or not? Some people dont have it. **cough** mymom **cough**".

While the system `multi-hand-rss(2)` is competitive with `classification-hand` (.739 vs. .78 correlation), it is important to note that `multi-hand-rss` is very sensitive to the number of keywords selected. The other results are `multi-hand-rss(1)` = .498, `multi-hand-rss(3)` = .732, `multi-hand-rss(4)` = −.991. On the other hand, `classification-hand` uses all four hand-chosen keywords, relying on the classifier to determine how much each message should be weighted in computing Q.

(4) Classification is sensitive to the set of training messages. This result is not too surprising, but it should be noted that the classifier's performance diminishes considerably when applied to documents with a very different word distribution. System `classification-freq` obtains a poor −.396 correlation on the testing data because it must classify documents that match any of the 5,000 keywords created by the *Most Frequent Keywords* method described in Section 2.3.2. This differs from `classification-hand`, in which it classifies documents that match the hand-chosen keywords, which is the same type of documents that were labeled to train the classifier. (Note, though, that there were no actual overlapping documents, since the labeled documents were drawn from a different time range).

Finally, while the best model exhibits a high correlation with true ILI rates (.78), its residual term is still likely too high for practical use. There are at least two reasons for this. First, the 10 weeks of data only results in 10 ILI data points. More data will likely enable both a better estimate of performance as well as a reduction in overfitting observed in the keyword selection procedures. We are in the process of collecting more messages, and fortunately, this data should be available soon, as Twitter has agreed to donate to the U.S. Library of Congress every public Twitter message since its inception (March 2006) [25]. Secondly, the characteristics of the train-test split make prediction particularly difficult. As Figure 3 shows, while the five training points exhibit limited fluctuation, a sharp drop in ILI rates begins with

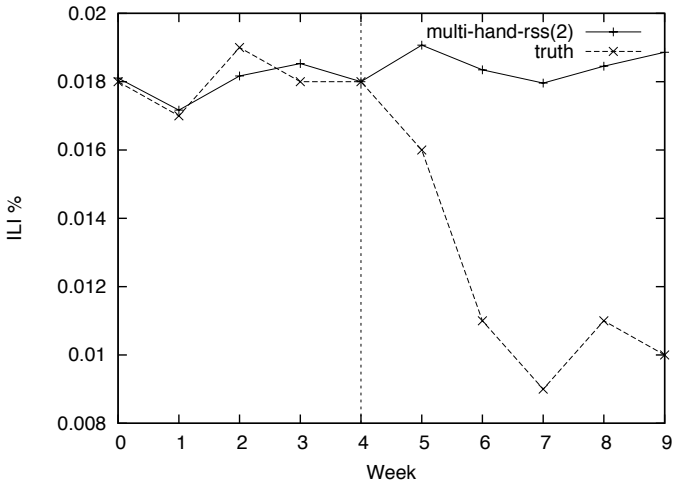

Figure 2: Results for `multi-hand-rss(2)` on training and testing data points.

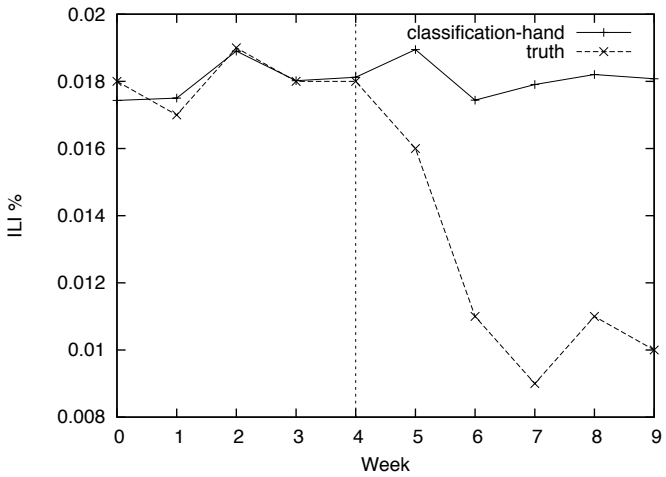

Figure 3: Results for `classification-hand` on training and testing data points.

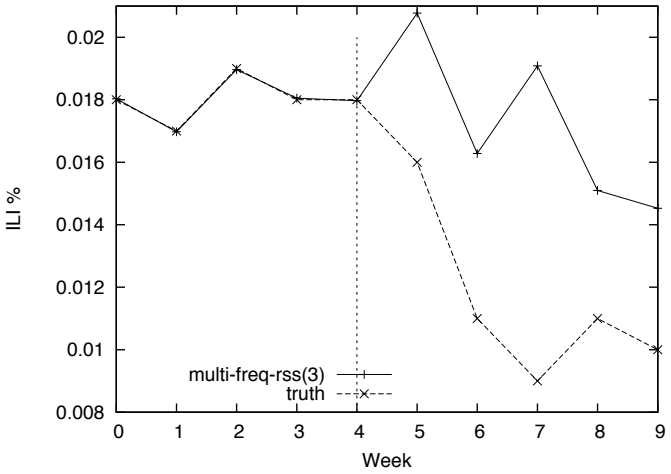

Figure 4: Results for `multi-freq-rss(3)` on training and testing data points.

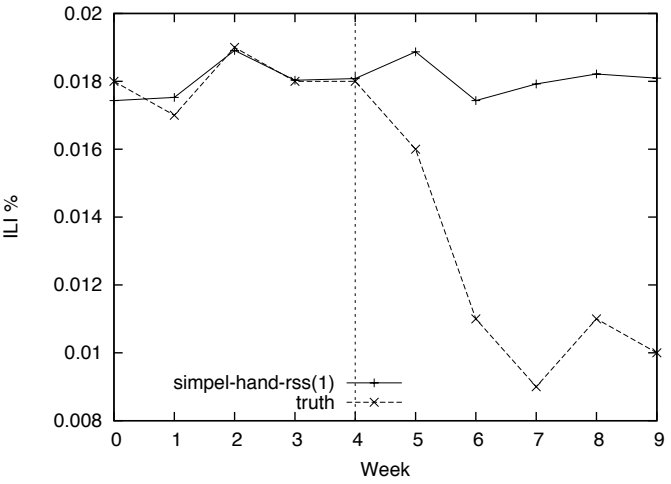

Figure 5: Results for `simple-hand-rss(1)` on training and testing data points.

the first test point as flu-season winds down. A training sample covering an entire flu season would likely result in significant error reductions.

5. RELATED WORK

There have been a number of proposals to monitor emerging health threats through an automated analysis of online news sources [18, 1, 23, 15]. While many rely on keyword matching or document classification, some apply more complex linguistic analysis such as named-entity recognition and topic modeling [9, 3]. A similar correlation was found in flu-related queries submitted to search engines [12, 6, 22, 8]. Our methodology most closely follows that of Ginsberg et al. [8]. The principal differences are the use of Twitter data, multiple regression, and document filtering.

Other Web mining techniques applied to user-generated Web content generally rely on a descriptive analysis derived by analyzing keywords and link structures. For example, Mishne et al. [19] track the mood of the blogosphere by tracking frequencies of terms like *tired* or *happy*, and Chau & Xu [2] analyzes link patterns to discover communities of hate groups. While this type of analysis is useful for generating descriptive trends of online data, our work aims to map these Internet trends to *real-world* statistics, and to accumulate these into actionable knowledge to be used by decision makers.

Most other related work in analyzing user-generated Web content is motivated by marketing applications, in which a client determines the impact of a marketing strategy by mining changes in consumer reception of the product from blog posts, product reviews, or Twitter messages [11, 14, 13].

A few researchers have applied regression on the output of a sentiment analysis system to predict real world values. For example, Gruhl et al. [10] use hand-crafted keywords applied to Web documents to predict product sales. Similarly, Liu et al. [17] perform sentiment classification on blog posts, using the output to estimate sales performance of a product. They use a similar computation as our *expected fraction* in

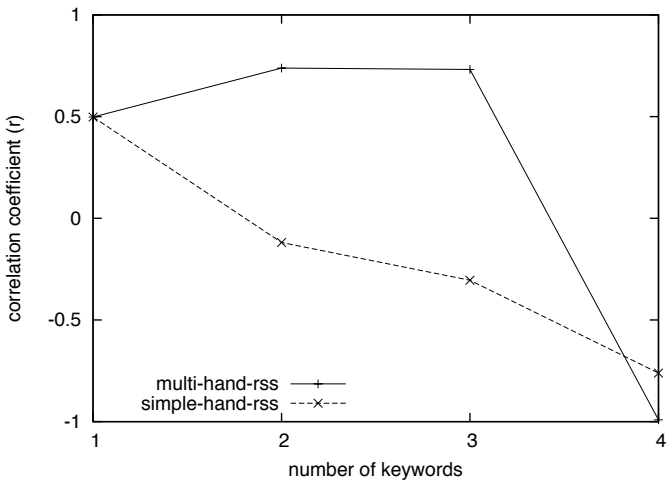

Figure 6: Correlation results for `simple-hand-rss` and `multi-hand-rss` on training and testing data points as a function of the number of selected keywords.

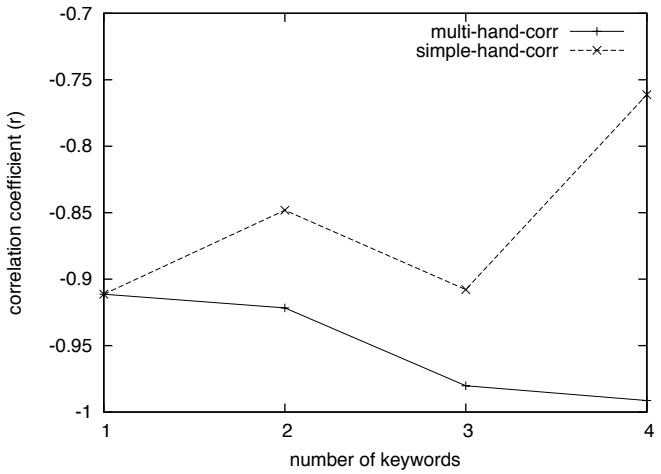

Figure 7: Correlation results for `simple-hand-corr` and `multi-hand-corr` on training and testing data points as a function of the number of selected keywords.

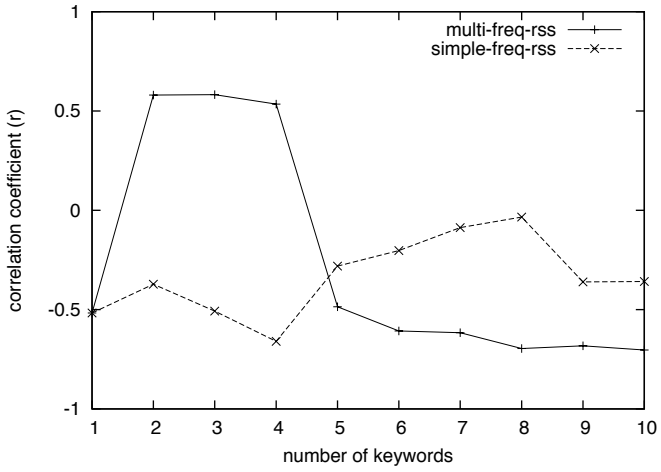

Figure 8: Correlation results for `simple-freq-rss` and `multi-freq-rss` on training and testing data points as a function of the number of selected keywords.

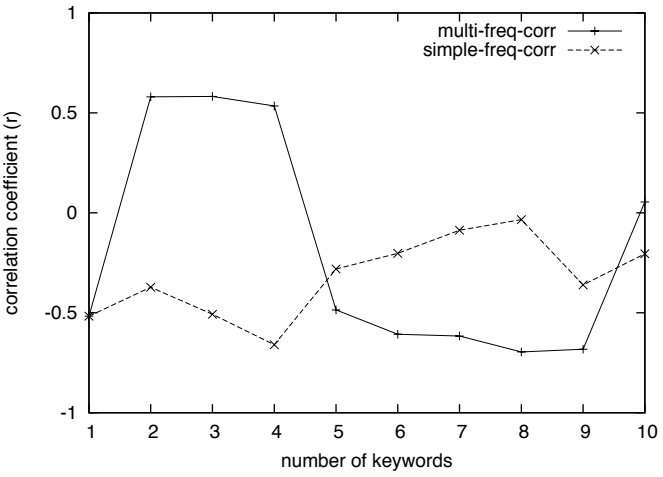

Figure 9: Correlation results for `simple-freq-corr` and `multi-freq-corr` on training and testing data points as a function of the number of selected keywords. Note that correlation and RSS resulted in the same keywords, so there is no difference between these results and those in Figure 8.

Equation 5. Additionally, de Choudhury et al. [4] show promising results predicting changes in stock market value based on properties of blogs.

Most recently, O'Connor et al. [20] performed perhaps the most comprehensive Twitter analysis to date, collecting one billion messages over two years, revealing that the frequency of certain hand-selected keywords correlates highly with public opinion polls. The results of Tumasjan et al. [26] suggest that it might be possible to use Twitter to predict election results.

6. CONCLUSIONS & FUTURE WORK

In this paper, we have explored the possibility of detecting influenza outbreaks by analyzing Twitter data. It appears that different approaches are required than those used for analyzing query log data. In particular, the longer messages allow us to use simple classification techniques to prevent erroneous messages from overwhelming model estimates.

Additionally, it is likely that supplying the classifier with more sophisticated linguistic features (n-grams, synonyms, etc.) will improve its accuracy. Perhaps more importantly, given the informal syntax and number of spelling mistakes in Twitter messages, it is likely that a more sophisticated pre-processing stage could improve the quality of analysis.

For future work we plan not only to analyze a larger number of messages, but also to associate geographic information with each message in order to perform a more fine-grained regression with region-specific ILI data reported by the CDC.

7. ACKNOWLEDGMENTS

We thank Troy Kammerdiener for helpful discussions. 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.

8. REFERENCES

[1] J. Brownstein, C. Freifeld, B. Reis, and K. Mandl. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine*, 5:1019–1024, 2008.

[2] M. Chau and J. Xu. Mining communities and their relationships in blogs: A study of online hate groups. *Int. J. Hum.-Comput. Stud.*, 65(1):57–70, 2007.

[3] N. Collier, S. Doan, A. Kawazeo, R. Goodwin, M. Conway, Y. Tateno, H.-Q. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi. BioCaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 2008.

[4] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann. Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 55–60, 2008.

[5] E. de Quincey and P. Kostkova. Early warning and outbreak detection using social networking websites: the potential of twitter, electronic healthcare. In *eHealth 2nd International Conference*, Instanbul, Tirkey, September 2009.

[6] G. Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA: Annual symposium proceedings*, pages 244–248, 2006.

[7] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June 2007.

[8] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457, February 2009.

[9] R. Grishman, S. Huttunen, and R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246, 2002.

[10] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 78–87, New York, NY, USA, 2005. ACM.

[11] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. 10th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2004.

[12] H. Johnson, M. Wagner, W. Hogan, W. Chapman, R. Olszewski, J. Dowling, and G. Barnas. Analysis of web access logs for surveillance of influenza. *MEDINFO*, pages 1202–1206, 2004.

[13] J. Kessler and N. Nicolov. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *3rd Int'l AAAI Conference on Weblogs and Social Media*, San Jose, CA, May 2009.

[14] S. Kim and E. Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *ACL Workshop on Sentiment and Subjectivity in Text*, 2006.

[15] J. Linge, R. Steinberger, T. Weber, R. Yangarber, E. van der Goot, D. Khudhairy, and N. Stilianakis. Internet surveillance systems for early alerting of health threats. *Eurosurveillance*, 14(13), 2009.

[16] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Programming*, 45(3, (Ser. B)):503–528, 1989.

[17] Y. Liu, X. Huang, A. An, and X. Yu. ARSA: A sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007.

[18] A. Mawudeku and M. Blench. Global public health intelligence network (GPHIN). In *7th Conference of the Association for Machine Translation in the Americas*, 2006.

[19] G. Mishne, K. Balog, M. de Rijke, and B. Ernsting. MoodViews: Tracking and searching mood-annotated blog posts. In *International Conference on Weblogs and Social Media*, Boulder, CO, 2007.

[20] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 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.

[21] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[22] P. Polgreen, Y. Chen, D. Pennock, and N. Forrest. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47:1443–1448, 2008.

[23] A. Reilly, E. Iarocci, C. Jung, D. Hartley, and N. Nelson. Indications and warning of pandemic influenza compared to seasonal influenza. *Advances in Disease Surveillance*, 5(190), 2008.

[24] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Media*, 2009.

[25] R. Stross. When history is compiled 140 characters at a time. New York Times, April 2010.

[26] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media*, Washington, D.C., 2010.

[27] I. Weber and C. Castillo. The demographics of web search. In *Proceedings of the 33th annual international ACM SIGIR conference on Research and development in information retrieval*, 2010.