Investigating Twitter as a Source for Studying Behavioral Responses to Epidemics

Alex Lamb, Michael J. Paul,* Mark Dredze
Human Language Technology Center of Excellence
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
{alamb3,mpaul19,mdredze}@jhu.edu

Introduction

Public health policies combatting the spread of infectious disease rely on bio-surveillance systems for coordinating responses such as vaccination programs. Classical epidemiology assumes that population behavior remains constant during an epidemic. However, the public response to an epidemic can have a serious impact on an epidemic's course.

As a result, modeling the influence of population behavior has become a recent trend in epidemiology research (Funk, Salathé, and Jansen 2010; Bisset et al. 2009), and there have been a few proposed mathematical models to explain individual responses to an outbreak, including game theoretic models (Reluga 2010) and agent based models (Epstein et al. 2008). Model validation has relied on simulated epidemic data because of a lack of empirical data. Consider influenza (seasonal flu) which, as the season progresses, leads to an increased awareness of infection, which may cause individuals to respond by obtaining a flu vaccination or limiting social contact. Fear of an illness can itself be a contagion (Epstein 2009), and disease propagation models could be applied. However, research will be limited without high quality empirical data reflecting population behavioral responses.

One promising source of such data are social media, such as Twitter. Recent studies have shown an ability to track influenza rates from Twitter (Paul and Dredze 2011; Aramaki, Maskawa, and Morita 2011; Signorini, Segre, and Polgreen 2011; Lampos, De Bie, and Cristianini 2010; Culotta 2010) since Twitter users tweet illnesses ("i am home sick with the flu"). However, users may also tweet concerned awareness of illness ("don't want to get sick, need a flu shot"). Identifying these messages can support proposed computational epidemic response models.

We present preliminary results for mining concerned awareness of influenza tweets. We describe our data set construction and experiments with binary classification of data into influenza versus general messages and classification into concerned awareness and existing infection.

Data Collection

We began by searching a collection of over 2 billion tweets (collected between May 2009 and October 2010) (O'Connor et al. 2010) for terms related to concerned awareness of influenza, including "flu", "worried", "worry", "scared", and "scare". The data from this period coincided with the 2009 outbreak of the H1N1 (swine flu) virus. To find data pertinent to our study, we manually annotated a subset of these tweets using Amazon Mechanical Turk (Callison-Burch and Dredze 2010), a crowdsourcing tool. Annotators were asked to describe tweets using four categories of interest:

- **Concerned Awareness:** People expressing concerned awareness are fearful that s/he or a friend/colleague/family member will contact the flu, but not that they already have the flu. These messages may reflect a heightened interest in taking preventative measures for the flu, jokes about someone being worried about the flu, etc.
- **Existing Infection:** Concern about existing infection: People expressing concern that a person may already have the flu. These messages may list symptoms or reasons to believe that the author, friend, colleague, etc. has the flu.
- **Recognition of Flu Coverage in the Media:** Users sharing news stories, linking to news articles about the flu, or saying that flu has been in the news.

Additionally, we included a fourth **Other** category for tweets that did not fit into any of the other categories. Tweets can be labeled with multiple categories.[1] Using these guidelines, we manually annotated 100 tweets to use as *gold standard* labels. Each MTurk task was comprised of ten tweets, including one gold standard tweet, against which we evaluated annotator performance. Each set of tweets was labeled by three different annotators to increase the robustness of the data. In total, 5,990 tweets were annotated.

Experiments

We experimented with using our annotated corpus as training data for supervised classifiers. The long term goal is to automatically identify relevant tweets for analysis.

Rejection Rates We checked if gold standard tweets were annotated correctly (i.e. the tweet was labeled with at least one of the categories from the gold annotations) to identify low quality annotators. Annotations by annotators whose

*Supported by an NSF Graduate Research Fellowship.
Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] For example, a tweet can be both concern awareness and existing infection: "my son has the flu and I'm worried I'll get it."

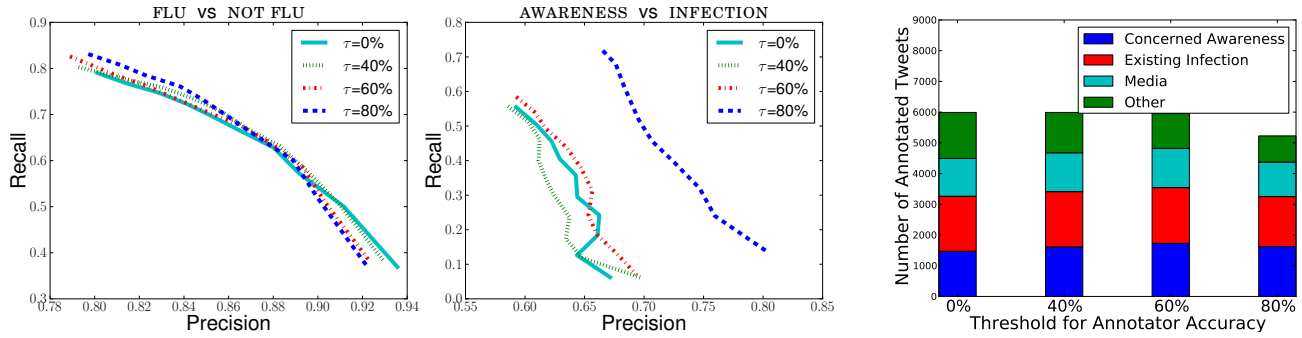

Figure 1: Precision-recall curves for two tasks (left) and the distribution of labels (right) at different rejection thresholds τ.

accuracy was below some threshold τ were rejected – different training sets were generated by using various acceptance thresholds τ (0%, 40%, 60%, and 80%.) The distribution of labels in the different data sets is shown in Figure 1. At these τ thresholds, the rejection rates (the percentage of annotations which were removed) are, respectively: 0%, 11%, 23%, and 60%. Although some annotations are removed, as long as a tweet still had at least two annotations, we included the tweet in the training data, so removing a few bad annotators did not significantly reduce the size of the data set. Each tweet's label was determined as the most frequent label among its available annotations.

Classification We considered two different binary classification tasks. First, we wanted to classify tweets as either FLU (either concerned awareness or existing infection) vs NOT FLU (shared media or other). Second, among the FLU tweets we classified them as either concerned AWARENESS or existing INFECTION.

Classification was performed with logistic regression (MaxEnt) using the MALLET library (McCallum 2002) with 1- and 2-gram word features. We replaced URLs with a URL feature. An instance was labeled as "positive" if its probability was above a certain threshold – by varying this threshold from 0.0 to 1.0, precision-recall curves were produced for both tasks (Fig. 1), measured over 10-fold cross validation. (In both tasks, the first label is considered the positive label when computing precision and recall.)

With an annotator accuracy threshold of 80% (the highest quality data), we achieve a maximum F-score of 0.81 for the FLU vs. NOT FLU task and 0.69 for the AWARENESS vs. INFECTION task. The labels FLU and NOT FLU essentially denote whether tweets are relevant or irrelevant to our study – we are not interested in tweets that do not express direct concern of influenza. We find that this task is fairly easy with little difference in the classifier performance among the different thresholds.

The difference between AWARENESS and INFECTION is a subtler distinction, and we find this to be a much harder task. Not only is the F-score lower for this task, but unlike the first task, we find that there is a substantial drop in performance if we include training data by annotators with less than 80% accuracy. This indicates that many annotators had difficulty recognizing this distinction. In future work, we hope to improve this with richer features beyond n-grams.

Conclusion

We have investigated a more nuanced view of influenza tweets than previous surveillance work. Our preliminary results suggest that it is possible to learn to distinguish different categories of influenza awareness. Moreover, that actual mentions of infection are less common than mere awareness, suggesting that influenza detection through the Web would benefit from distinguishing these message types.

References

Aramaki, E.; Maskawa, S.; and Morita, M. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *EMNLP*.

Bisset, K.; Feng, X.; Marathe, M.; and Yardi, S. 2009. Modeling interaction between individuals, social networks and public policy to support public health epidemiology. In *Winter Simulation Conference (WSC), Proceedings of the 2009*, 2020–2031. IEEE.

Callison-Burch, C., and Dredze, M. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *NAACL Workshop on Creating Speech and Language Data With Mechanical Turk*.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *ACM Workshop on Soc.Med. Analytics*.

Epstein, J.; Parker, J.; Cummings, D.; and Hammond, R. 2008. Coupled contagion dynamics of fear and disease: Mathematical and computational explorations. *PLoS ONE* 3(12).

Epstein, J. 2009. Modelling to contain pandemics. *Nature* 460(7256):687–687.

Funk, S.; Salathé, M.; and Jansen, V. 2010. Modelling the influence of human behaviour on the spread of infectious diseases: a review. *Journal of The Royal Society Interface* 7(50):1247–1256.

Lampos, V.; De Bie, T.; and Cristianini, N. 2010. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases* 599–602.

McCallum, A. 2002. MALLET: A machine learning for language toolkit.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*.

Paul, M. J., and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*.

Reluga, T. 2010. Game theory of social distancing in response to an epidemic. *PLoS computational biology* 6(5):e1000793.

Signorini, A.; Segre, A.; and Polgreen, P. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza a H1N1 pandemic. *PLoS One* 6(5):e19467.