E-055

Estimating Outbreak of Influenza Like Diseases Using Social Media

Muhammad Asif Hossain Khan[1], Masayuki Iwai[1], Kaoru Sezaki[1,2]

Abstract- **Early detection of the onset and outbreak of influenza like contiguous diseases followed by timely intervention by authorities with rapid response can greatly reduce the damage caused by them. Unfortunately, until now actual data arrives in the hands of authorities with a lag time of ten days to two weeks. In this research we have used social media, namely Twitter, to predict the state of such disease outbreak by analyzing tweets made by the members of the community. We have used natural language processing techniques to identify and quantify flu related tweets from normal tweets .**

I. Introduction

Influenza like illness (ILI) is increasing in an alarming rate all over the world causing from 250 thousands to 500 thousands deaths each year. In our research, we are trying to use social media to gather information regarding influenza like diseases. The objective is to get early indication about the state of the disease in a community so that authorities can take appropriate actions to subdue its impact. Similar efforts has been made by Google by using search engine query data [1]. The ground truth data about the spread of influenza like diseases collected and compiled by different government organizations still lag by about two weeks. We are trying to develop an automated system named ILIAD (Influenza Like Illness – Advance Detector), that would be able to detect the outbreak and spread of influenza like contagious diseases in a community based on the information shared by people of that community through social media like Twitter with a lag time of one day. The work is still ongoing, but we have completed some modules of the system. In the following sections we present the architecture of our system, our methodologies and obtained results .

II. System Architecture

Our developed system ILIAD would be composed of four modules. The representativeness verification module [1] would be responsible to verify that the community is responsive enough about ongoing events in the society like the outbreak of a disease.

The keyword identification module would be responsible for identifying disease related keywords and assigning weights to them. As Twitter imposes character limitations (140 characters / tweet), people use a lot of jargons, vernaculars and colloquial in their tweets, which also varies from region to region. We are working on identifying region specific keywords.

The disease onset and outbreak detection module is responsible for identifying the intensity of disease outbreak in the community. Sub modules in this module are responsible for filtering out flu related tweets from the daily twitter corpus, identifying the emotional polarity in the flu related tweets and using sentence dependency structure to accurately quantify the information in the tweets. In future, we shall consider information from external sources like sudden environmental changes, periodic seasonal impact etc.

The propagation projection module would project the propagation direction of the disease taking human mobility model of the region into consideration.

III. Methodologies

We have completed the implementation of the Representativeness Verification module of ILIAD [2]. We are working on the Keyword Identification module and the Disease Onset and Outbreak detection module simultaneously. We have implemented some parts of both the modules and in this section we describe the methodologies we have adopted thus far.

Identifying Flu Related Keywords

We have crawled forums and news sites where users had shared their flu symptoms and flu related experiences. After removing html tags, web references, stop words and rare words, we stemmed the rest of the words using Porter's Algorithm [3]. However, we found that no stemming algorithm could perfectly deal with all words. For example, Porter stemmer wrongly stemmed the word 'Probability' to 'probabl' and the word 'saw' when used as a verb, i.e. past tense of the verb 'see', was mistaken with the noun 'saw' and thus wrongly assumed to be already stemmed. The second sort of mistakes is done because stemmer algorithms do not consider the part-of-speech of a word in the sentence and so cannot distinguish a noun from a verb. Therefore, we decided to use libraries developed by Stanford's NLP group [4] for lemmatizing tweets instead of stemming. We developed frequency histograms of the lemmatized words and used top 50 high frequency words that are related to flu symptoms as our final set of keywords. To identify colloquial and jargons we used a probabilistic generative model called Latent Dirichlet Allocation (LDA) [5] that performs latent semantic analysis. We applied LDA on the daily twitter corpus of New York on several days of their last flu season (February 10 to February 28 2011) to identify the colloquial people used when twitting on flu.

Processing Flu Related Tweets

Thus far, we have used the naïve bag of words technique to filter flu related tweets from daily twitting corpus. However, we are now considering using machine learning algorithms for document classification. We want to start with Multinomail Naïve Bayes classifier and gradually compare the performance of other machine learning techniques for accurately identifying flu related tweets.

[1] Institute of Industrial Science, The University of Tokyo

[2] Center for Spatial Information Science, The University of Tokyo

After filtering the flu related tweets, we have determined their sentence dependency structure and part-of-speech for each word appearing in the tweet (Figure 1). We shall use this information as features for training data for the machine learning algorithms.

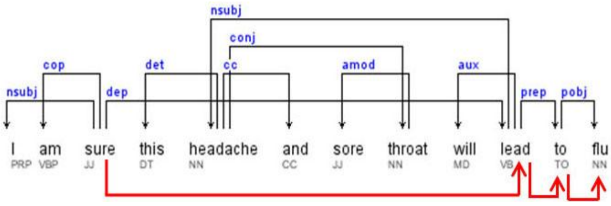

Figure 1. Identified dependency graph and part-of-speech of a tweet

Identifying Emotional Polarity in Flu Related Tweets

Diseases like flu certainly have impacts on the mental state of the infected person. We assume that when a person will start to believe that s/he or her/his close acquaintances are vulnerable to infection, this concern will be reflected in tweets s/he makes. Therefore, we tried to classify the daily flu related tweets into one of the two mood polarities – positive and negative. We have used AFINN-111 [6], a list of subjective lexicon of 2477 words, where each word has been rated with an integer from -5 to +5 depending on how much positive/negative sentimental notion the word carries. We have quantified each flu related tweet with a mood score (emotional polarity) depending on the lexicons they contain. Each tweet was assigned a positive score and a negative score. The summation of positive scores and negative scores of all flu related tweets of each day was calculated. The ration of positive vs. negative mood scores constitutes the emotional polarity of a day.

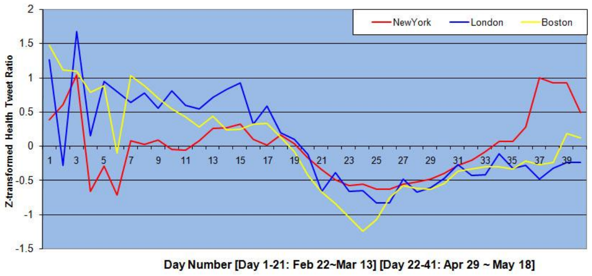

Figure 2: Ratio (z-transformed) of flu related tweets vs.non-flu related tweets in three cities

IV. Results

For the cities of New York, Boston and London we filtered out flu related tweets from other tweets. Then we calculated the ration of flu related tweets vs. non-flu related tweets for each city. We have used z-transformation on the daily time series of this ration to get a distribution with mean 0 and standard deviation 1. Then we took a 7-point moving average on the time series data to diminish any impact of any particular day of the week, e.g. the weekends. Figure 2 shows the results for the three cities. We are working on developing a model that will be able to predict the ration of infected people in a community using this tweet ration and environmental parameters of the region into consideration. Figure 3 shows the daily ration of mood scores for the three cities we are observing. Boston seems to have a relatively smooth change in mood valence relative to New York or London. We are still to use this information in our analysis. However, we believe that the ratio will be getting more negative as the intensity of the flu epidemic increases and will start to move towards positive scales when flu season will be closing to an end. This hypothesis is still to be verified.

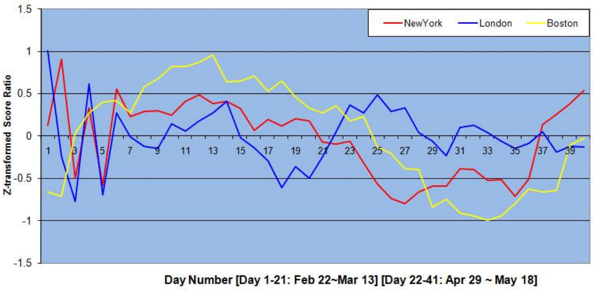

Figure 3: Ratio of mood scores (z-transformed) for three cities (7-point moving average applied)

V. Conclusion

We are trying to develop an automated system that when deployed in a region for detecting the activity of influenza like illness, would customize itself for the region. It would extract the flu information from social media used by the inhabitants of the region. We have completed implementing some modules of our model and are working on the remaining modules. In the paper we have presented our methodologies and obtained results regarding some of our developed sub-modules like Representativeness verification module, Keyword identification sub-module, Tweet classification sub-module and Emotional polarity identification sub-module. We believe that our system would be able to accurately infer the onset and outbreak of influenza like illnesses in a region with a lag time of few hours to one day.

References

[1] Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S., and Brilliant L. “Detecting influenza epidemics using search engine query data” . *Nature*, Vol. 457(19), pp. 1012-1014, 2009

[2] Muhammad Asif Hossain Khan, Tomohiro Sakamaki, Masayuki Iwai, Yoshito Tobe and Kaoru Sezaki, 2011 (March). “Using Entropy and Observed Twitting Behavior to Identify Large Events”. In *proceeding of 73rd Annual Convention IPS Japan*, 5D-1, Tokyo, Japan.

[3] Porter M. “ An algorithm for suffix stripping ” *Program*, vol. 14(3), pp. 130‑137, 1980.

[4] http://nlp.stanford.edu

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *JMLR*, vol. 3, pp. 993-1022, 2003.

[6] http://arxiv.org/abs/1103.2903