AAAI Technical Report FS-12-05
Information Retrieval and Knowledge Discovery in Biomedical Text

Multi-Tweet Summarization for Flu Outbreak Detection

Brent Wenerstrom and **Mehmed Kantardzic** and **Elaheh Arabmakki** and **Musa Hindi**
Computer Engineering and Computer Science Department
University of Louisville
Louisville, KY 40292

Abstract

Twitter provides the freshest source of data about what is happening in the lives people across the world. The publicly available streams of status updates available on Twitter have been used to track earthquakes, forest fires and most especially flu outbreaks. Current techniques for tracking flu outbreaks rely on count data for a number of keywords. However, count data alone on the noisy Twitter streams is not reliable enough for health officials to make critical decisions. We propose a semi-automatic outbreak detection system. Rather than providing only alarms backed by count data, we propose a summarization system that will allow health officials to quickly verify outbreak alarms. This will lead to higher levels of trust in the system and allow the system to be used by health organizations around the world. We experimentally verify our summarization system and have found system users to have an accuracy of 0.86 when identifying multi-tweet summaries.

Introduction

Previously systems (Achrekar et al. 2011; Culotta 2010) have been proposed for the purpose of detecing flu outbreaks. Generally these systems automatically select keywords to follow on Twitter, then count the number of tweets at each time step per keyword. Generally baselines are created for the detection of outlying count data. Additionally, the system may include a regression model for the purpose of predicting future values based on the streaming counts. Alarms may be triggered automatically to alert health officials that an outbreak may be occurring. When the right set of keywords are chosen these types of system can be highly accurate, reaching as high as a 98% correlation with actual reported flu cases. However, Twitter tweets are highly noisy.

We propose a semi-automatic outbreak detection system. Health officials will be much more confident in the output of a system that incorporates both precise algorithms and the understanding of a human expert. When our system detects a rise in activity which could predict an outbreak, our system provides a summarized view of the tweets for that particular area. This allows for a health expert to verify that the increased volume of tweets are related to sickness. Human verification will provide a higher level of trust in the system. It will be instrumental in the potential uptake of disease outbreak monitoring within organizations such as GPHIN. This work provides an approach to multi-tweet summarization with an emphasis on detecting flu outbreaks.

Semi-Automatic Flu Outbreak Detection

Our semi-automatic outbreak detection system takes the following steps: 1) Twitter data is gathered, 2) tweets are preprocessed, 3) tweets are clustered by topic, 4) summarizing tweets are selected, and 5) a summary is generated.

We obtained input datat for our system using Twitter's Streaming API (https://dev.twitter.com/docs/streaming-api). We download nearly all tweets which contain any one of these keyword phrases: flu, sore throat, cough, runny nose and headache. We then submit every unique user location to Yahoo!'s PlaceFinder (http://developer.yahoo.com/geo/placefinder/). From this online resource we obtain Yahoo!'s best guess for the city, state and country of a given Twitter user. Of the 19 million tweets we have obtained, Yahoo! is able to identify a city level location for 54% of the tweets or about 10.4 million tweets.

Before processing tweets we transform the tweet text to probability vectors using LDA (Blei, Ng, and Jordan 2003). LDA is a model that assumes that tweets are formed in a generative process. LDA represents tweets as being made of a mixture of topics. For each word in a tweet, first a topic is randomly obtained over the distribution of topics for a tweet. Then each topic provides a probability distribution over words. A word is randomly obtained from the topic distribution. The result is that each tweet is represented by a small set of features representing the probability of coming from a particular topic. Experimentally we obtained best results using 10 topics.

Tweets are clustered using agglomerative, hierarchical clustering using Ward's minimum variance method (Ward 1963). Distance between tweets is determined by the cosine distance between topic probability vectors. Each tweet begins the process in its own cluster. Clusters are joined one at a time until a single cluster results. Finally we compare the change in variance as used by Ward's method for each cluster join, when the highest percentage change in variance is found the clusters found at that step are used. A maxi-

mum of 10 clusters are used since a summary per cluster is generated, and the overall summary needs to remain small.

Clusters are summarized by the top ranking tweets. Tweets are ranked within a cluster using a graph theory metric for centrality called Closeness Centrality. Each tweet is modeled within a graph as a node and the distance between tweets form the edges. Distance is measured as one minus the Jaccard index of the bag of words for each tweet. Each cluster is then represented by the top one, two or three tweets, depending on the size of the cluster.

When summarizing a cluster we provide the number of tweets in the cluster and highlight common words in the cluster. An example of the format used can be seen in Figure 1. In this example there are 27 tweets within the cluster. Each additional summarizing tweet beyond the first is prefixed by '+' sign. Common terms are bolded. The term "headache" was seen in 25 of the 27 tweets and "giving" was seen in 3 of the 27 tweets. Both are highlighted in this case being in at least 3 tweets.

(27) Headache ??
+ When a **headache** pops up out of no where <
+ Im **giving** myself a **headache**... Its unlike me

Figure 1: Example tweet cluster summary taken from tweets in Louisville, KY for April 10, 2012.

Human Evaluation of Multi-Tweet Summaries

To evaluate our system we had human judges manually label each tweet whether it was about an individual having the flu or not. Then another judge reviewed the summaries to determine of the summarized tweets were more about individuals being sick or about other flu related topics. The idea being that good clustering and summarization will not require a user to review every tweet, but that they would obtain nearly similar results of categorization as someone who reviewed each tweet individually. We used three judges. Each judge was assigned four days worth of tweets from Louisville, Kentucky for tweets during April 2012.

Table 1: Break down of tweet level accuracy.

		Cluster Labels	
		sick	not
Ground Truth	sick	1,247	209
	not	498	478
	Accuracy		0.71

Accuracy provides a comparison of cluster labels with individual tweet labels. Overall our judges achieved an accuracy of 0.71 as seen in Table 1. This table provides the confusion matrix of tweet classifications. In total 60% of tweets have a ground truth label of "sick." This means that if we were observing the keywords for outbreak alarms chosen in this paper, then only 60% of the actual tweets relate to flu outbreaks. Through the cluster labels 72% of tweets are given a label of flu. Overall the judges' labels obtained an accuracy (0.71) very near what we predicted (0.72) in the previous subsection for word overlap plus closeness centrality.

Table 2: Break down of cluster level accuracy.

		Cluster Labels	
		sick	not
Majority	sick	68	8
	not	9	39
	Accuracy		0.86

Accuracy is directly affected by the purity of the clusters obtained. Improvements in purity will lead to higher accuracies. For example if on average a cluster contains 70% of one class and 30% of the the other, then the best accuracy that could be achieved would be 0.7. We would like to assess the cluster labels without consideration for purity. To do this we compute cluster level accuracy using the same judgments as before, by assigning the majority label to each cluster and comparing cluster labels against the majority label for that cluster. When computing cluster level accuracy, clusters with an equal number of labels in each class were ignored. The results of the cluster level accuracy is shown in Table 2. Our judges were very reliable at the cluster level and obtained a cluster level accuracy of 0.86. Of the 124 clusters only 17 of these clusters were mislabeled.

Table 3: Clustering and tweet accuracy results by judge.

Judge	Accuracy	Cluster Acc.
1	0.69	0.93
2	0.74	0.84
Medical Exp.	0.70	0.82
Total	**0.71**	**0.86**

Thus far we have only shown results of performing clustering and summarization on sets of tweets with around 100 tweets. The number of tweets one must read scales as a function of the number of clusters not the size of the data set. For example on data set of 1 million tweets, if 10 clusters were found the most tweets a user would see would be 30 tweets.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; and Liu, B. 2011. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 702–707.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Culotta, A. 2010. Detecting influenza outbreaks by analyzing twitter messages. In *Proc. 2010 Conf. on Knowledge Discovery and Data Mining*.

Ward, Jr., J. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58(301):236–244.