Using Computational Linguistics to Explore Patterns and Features to Identify Disease Outbreak signals in Social Network Data

Rastogi, Neelesh *(Senior)* & **Keshtkar,** Fazel *(Assistant Professor)*, Department of Computer Science and Mathematics, CPS

Abstract

The **CDC (Centers for Disease Control and Prevention)** currently diagnoses millions of cases of infectious diseases annually, generating population disease distributions that, while accurate, are far too delayed for real-time monitoring. The ability to instantly compile and monitor such distributions is critical in identifying outbreaks and facilitating real-time communication between health authorities and health-care providers. To offer such a solution, we introduce a novel pipeline based model to generate a near real-time, accurate depiction of infectious disease propagation over Twitter network. Our approach, a unique blend of natural language processing and supervised machine learning, is invariant to mass media hype and significantly lowers the noise introduced by the use of tweets. The correlation between our identified Twitter disease distribution and CDC data from late-2015 to mid-2016 was 0.983, thus, showing an improvement over the best model published for the 2015-16 flu season. Our model also correlates well with theoretical simulations of infection spread across airport networks, verifying its robustness and applicability in the public sphere.

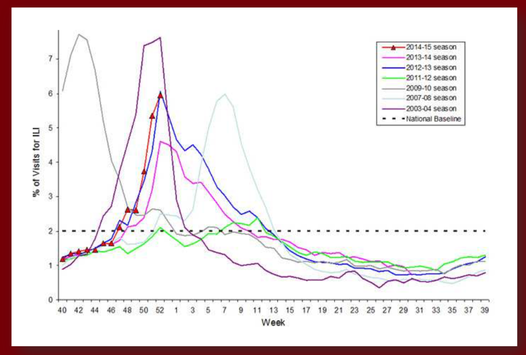


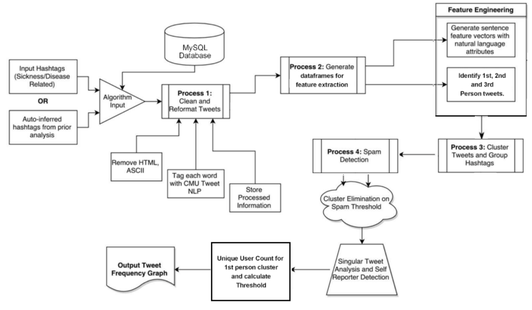
Figure 1: *The ILI distribution depicts the percentage of visits for influenza-like illness reported to the CDC by the US outpatient ILI surveillance network. Note the right-skewed nature of the curve, depicting the infection frequency increasing between months of November and January. An approximate three week delay is incurred in the generation of the disease distribution due to the time-consuming process of aggregating national patient reports.*

Model Pipeline

TWEET CATEGORY DEFINITIONS

In order to develop a robust and viable model of the CDC ILI distribution, we differentiate between three unique categories:

- **Self-Reporting Tweets:** these tweets originate from either an infected individual or someone associated with an infected individual. Tweets in this category signify that the author is likely to have a direct influence on the ILI curve.
- **Non Self-Reporting Tweets:** Non self-reporting tweets encompass tweets posted by news networks and concerned citizens not immediately affected by a sickness.
- **Spam:** As in all social networks, spam messages drastically increase distribution noise and provide no saliency when generating the ILI distribution. In this work, we consider as spam all tweets that do not refer to disease.



SOCIAL NETWORK ANALYSIS PIPELINE

$$\mathcal{P}(S) = \text{unique}(U) \times \prod_{n=1}^{3} \left(\frac{1}{1 + \#(G_n) - \text{unique}(G_n)} \right)^n$$

This part of the pipeline accepts as input either a list of hashtags or auto-inferred terms from prior analysis. Our model leverages exhaustive uninformative tweet elimination to allow for the identification of anomalies and unique disease outbreaks, thus providing prognostic significance.

Natural Language Processing

1. HASHTAG SPECIFICATION

As our pipeline accepts keywords as input to search for relevant tweets, we initially obtain hashtags linked to specific diseases (such as #influenza, #dengue, #zika, etc.) by ascertaining the popularity of disease related hashtags currently in use.

2. LINGUISTIC TERM ASSOCIATION

We use linked n-grams in order to obtain additional hashtags and keywords aside from those directly linked to disease, such as #sick and #Nyquil.

3. TERM CORPUS, TOPIC MODELING

$$\text{tf}(t,d) = 1 + \log f_{t,d}$$
$$\text{idf}(t,D) = \log\left(1 + \frac{|D|}{n_t}\right)$$
$$\text{tfidf}(t,d,D) = \text{tf}(t,d) \times \text{idf}(t,D)$$

We assign numeric feature vectors to collected tweets utilizing TF-IDF (term frequency–inverse document frequency) vectorization within corpora of hashtags.

4. TWEET CLUSTERING

Using the TF-IDF features ascertained in Step 3 and a mixed Euclidean-Cosine similarity measure, we cluster tweets according to minimal cluster RSS value via the centroid-based k-means approach.

5. SALIENT TWEET ISOLATION

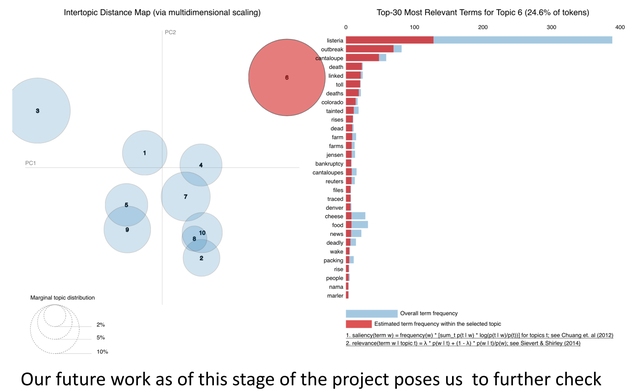
We train and apply a linguistic attribute-based random forest classifier to randomly selected subsets of each cluster, rejecting an entire cluster if its chosen subset contains a sufficiently large number of non self-reported tweets.

6. ILI ANALOG FREQUENCY DISTRIBUTION

We subsequently plot the frequency distribution of relevant tweets over time in order to model the CDC ILI curve.

7. SIMULATION & EVALUATION

CDC ILI and Tweet Distribution



Our future work as of this stage of the project poses us to further check our model within different classification algorithms. We also plan to use more various dataset on other identified disease outbreak and other social network as well. In this research using natural language processing and ma- chine learning we explored various features, over social network data, and found significant cues for certain patterns within our dataset for potential disease outbreak. We utilized various content-based features such as; unigrams, bigrams, trigrams count, Word Cloud, Sentiment Orientation, and Topic Modeling to create an ensemble of model to achieve a feasible, computational linguistics and machine learning approach towards content-based data analysis over Social Network like Twitter. (Pritchard, Stephens, and Donnelly 2000)

Results

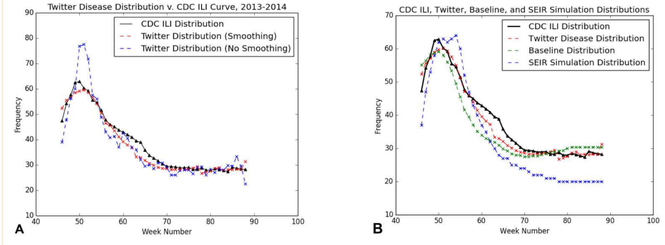


Figure 4: (**A**) A comparison of the Twitter-derived distribution with the CDC ILI curve. (**B**) A side-by-side plot of the Twitter, Baseline, SEIR, and CDC distributions.

	CDC	Twitter	Simul	Base	CDC	Twitter	Simul	Base
CDC ILI	–	0.983	0.931	0.938	–	0.003	0.014	0.005
Twitter	0.983	–	0.947	0.972	0.003	–	0.018	0.001
Simulation	0.931	0.947	–	0.898	0.014	0.017	–	0.025
Baseline	0.938	0.972	0.898	–	0.005	0.001	0.025	–