Harnessing Machine Learning to Anticipate Infectious Disease Threats with Pandemic Potential

Chandra Swarathesh Addanki
Department of Computer Science
Lakehead University
Thunder Bay, Ontario

***Abstract* — With the massive emergence and exponential increase in the number of infectious diseases with pandemic potential spreading across the globe, we need a modern approach to find these sort of infectious diseases before they reach their pandemic potential , Our objective is to address this challenge by creating an IVR (interactive Voice Response) system which takes a voice input from user then queries, filters and classifies using big data whether the disease is pandemic and informs the user the results of the prediction and keeps track of it.**

Keywords— Pandemic Disease, Disease Classifier, Interactive Voice Recorder, Naive Bayesian, Machine Learning, Natural Language Processing, Big Data.

I. INTRODUCTION.

With the rapid development of the modern civilization the number of infectious diseases with deathly consequences also increased with it and keeping track of such kind of infectious diseases.

It is necessary because if they are kept unchecked, they may reach their pandemic potential and consequences will lead to a severe human disaster.

However, there is no any distinct system which monitors these kinds of diseases from user using IVR, with big data which holds tremendous promise for infectious diseases research with pandemic potential, surveillance, and prevention we can create a system to create a solution.

The objective of this paper is to address this challenge through the means of automated feature extraction of symptoms from the conversation of the user or the potential patient with our IVR system (Interactive Voice Response) and use machine learning techniques on the feature extracted which are nothing but the symptoms to classify and predict the type of disease the user has and also to facilitate knowledge management and early detection of infectious disease with pandemic disease .

II. BIG DATA

The term big data is used to refer to data sets that are too large or complex or maybe both which are optimal for traditional data-processing application software to deal with, Data which consists of rows offer greater statistical power, while data with [1] more attributes or columns may lead to a higher bias rate.

The challenges [2] Big data face includes capturing data, data storage, data analysis, search, sharing, transfer visualization, querying, updating, information privacy and data source.

Big data has three primary key concepts:

- Volume.
- Variety.
- Velocity.

Other concepts which are included later which are attributed with big data are veracity [3] [4] and value. [5]

The term "big data" tends to refer to the use of predictive analytics, logical analytics, or certain other advanced data analysis method that predicts disease from the given symptom, and seldom to a particular size of data set in our case, we use the data set consisting of diseases and their associated symptoms. [6]

Analysis of data sets can find new correlations to "predict diseases, prevent diseases and so on."[7].

III. BIG DATA IN DETECTION OF INFECTIOUS DISEASES.

The Big data analytics has helped in improving detection of infectious disease by providing predictive analytics based upon on the symptoms, clinical risk intervention.

The amount of data generated in healthcare systems is not trivial or small.

This includes electronic health record data, imaging data, patient-generated data, sensor data, and other forms data which are hard to process. [18]

The words “Big Data“ often refers to ‘dirty data' and the fractional bias of data .

Inaccuracies increase with data volume growth. Manually inspecting at the big data scale is impossible and there is a desperate need in pandemic surveillance for intelligent tools for accuracy and believability control and handling of information missed. [19] The use of big data in disease prediction has raised significant ethical challenges ranging from risks for individual privacy and autonomy to transparency.

Data about diseases and outbreaks are outsourced not only through online by government agencies but also through other channels that are informal, like press reports, blogs, chat rooms, analyses of Web searches. When we aggregate all such kind of data.

These sources provide a big picture of global health that is fundamentally much different from that yielded by the reporting of the traditional public health infrastructure.

But these data are often biased because they come from unreliable sources, using such kind of data leads to predictions which are biased.

IV. BIG DATA FOR INFECTIOUS DISEASE AND MODELING.

The Big data which is gathered from social media like Twitter and Facebook, the internet and other digital sources have much potential to provide timelier and required information on infectious disease threats or outbreaks than traditional surveillance methods.

But the problem with such kind of data is that they may sometimes be biased or entirely fake using such kind of data leads to severe consequences.

Like Google Flu Trends Disaster- influenza pandemic in 2009, this disaster is caused by data bias towards the search results.

Typical traditional infectious disease surveillance system which is typically based on laboratory tests and epidemiological data collected is up to the standard.

But it should be noted that it can include the lag of time and is expensive to produce and typically and mostly lacks the local resolution needed for predictive and accurate monitoring.

Further, it can be cost-effective in low-income countries.

In contrast, big data streams from the internet, for example, is available in real time and can track disease activity in real time but have their own biases because of the data.

Hybrid tools which is a combination of both the traditional surveillance and big data sets may provide a way forward better prediction of diseases, serving to complement existing methods rather than replacing them.

The ultimate goal of the pandemic surveillance system is to be able to forecast the size, peak or trajectory of an outbreak before it happens and also in order to provide a better response to infectious disease threats.

Combining big data in surveillance is a primary step toward this long-term goal to prevent a pandemic outbreak.

The potential of big data must be tempered with attention, Non-traditional data sources may lack demographic identifiers.

Furthermore, the social media source's may not always be constant origin of data, as they can die if there is a loss of concern, financing or maybe fake entirely.

Most importantly, any unique data source must be validated against established infectious disease surveillance data and systems.

V. ENSURING DATA PRIVACY

The Big data offers a lot of possibilities to provide more information for infectious disease surveillance, but the purpose is decades behind other fields such as climatology and marketing.

E-health records with personal identifying information removed, for example, may be a resource to monitor infectious diseases outcomes.

Applying the personal identification data to surveillance has been very slow.

This is because of concerns regarding ethical reasoning such as and not limited about patient privacy.

VI. HARNESSING SPATIAL BIG DATA

To predict or determine the origin of an outbreak or where future ones may occur.

For example, the need for spatial data rises, social media such as tweets and Facebook posts and mobile phones have the to supply topological knowledge gaps.

But it should be noted that there are technical, practical and ethical issues that should be addressed.

The required restrictions and possible solutions to protect privacy, such as masking individual-level information should be implemented.

This includes reviews of two nontraditional sources for monitoring influenza and other diseases

- Crowdsourced data.
- Direct conversation with the potential patient.

VII. CONNECTING MOBILITY TO INFECTIOUS DISEASES

With appropriate restrictions and rules implemented to ensure anonymity, call conversations from mobile phones may provide researchers "an unprecedented opportunity" to determine how to predict the diseases and to monitor them.

Studies [21] of malaria and rubella in Kenya showed how call data improved the understanding of the transmission of those diseases.

Because mobile phone data has biases young children are not likely to be represented.

For example: more research is needed to determine if mobility patterns derived from call data records are representative of general travel patterns.

VIII. CASE STUDY: CONSEQUENCES OF USING BIASED DATA

On February 2013, Google Flu Trends (GFT) made headlines because of the flu tracking system, the news source Nature reported through their news source that GFT was predicting more than double the amount of doctor visits for influenza-like illness rather than the Centers for Disease Control and Prevention, which bases its results on surveillance reports from laboratories.

This happened even though GFT was built to foretell CDC reports, this is due to biased searches the user did, resulting in prediction which is biased.

The problems identified here are not restricted to GFT. Analysis on whether a search or social media can predict pandemic diseases

Although the studies [22] shows the value of this kind of data, they are far from a origin where they can replace more traditional methods or theories.

In this paper we explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics.

- Big data hubris is the assumption that big data are a substitute, rather than a supplement for the traditional data collection and analysis for analytic purposes.

- The beta and the first version of GFT was a particularly problematic combination of big and small data.

- Primarily, the ideology was to find the best matches among 50 million search terms to fit 1152 data points.

- The chances of finding search terms that match the competence of the flu but are biased and totally unrelated.

- Although not widely published until 2013, the new GFT has been persistently exaggerating flu predominance for a much longer time period.

- GFT also abstained by a very wide margin in the 2011–2012 flu season and has missed hundred out of one hundred and eight weeks starting with August 2011.

- These errors are not randomly assigned. For example, first week's errors predict this week's and degree of error varies with the time of year.

- These exemplars mean that GFT overlooks important information that could be obtained by conventional statistical methods.

- Even with the update to GFT in 2009, the comparative of the algorithm as a stand-alone flu monitor is questionable. Because they are fractionally biased.

IX. TRANSPARENCY, GRANULARITY.

A glimpse at the flu forecast on your phone sets you straight if there’s a [22] predicting about a current spike of cases nearby.

So, the potential patient may head to the clinic rather than risk a feverish week in bed.

Epidemiologists sincerely predict such a future, in which they can trace infectious diseases with the same determination as meteorologists mapping the weather.

But those making forecasting of this type, face a serious dilemma. There is just not a huge amount of observational data in the disease world.

X. METHODOLOGY

In this paper, we implement an interactive voice response system. This system will have an interactive conversation with the user, or the potential patient and conversations are then filtered and extracted for the symptoms and using the symptoms extracted from the conversation we predict the type of disease the user or the potential patient is suffering.

The methodology consists of four phases they are as following:

1. Phase one: Interaction with the help-line
2. Phase Two: Speech to text conversion
3. Phase Three: Feature extraction
4. Phase Four: Disease classification

1. Phase One: Interaction with the helpline

In this module the user uses calls the system hosting the interactive voice response system through the hotline we provide , the interactive voice response system starts the conversation with the user or the potential patient , the interactive voice response system asks the user or the potential patient to describe the symptoms the potential patient is suffering from , the interactive voice response system also notifies that the call is being recorded for quality purposes and analytic purposes.

Here is the basic diagram for the flow of the conversations.

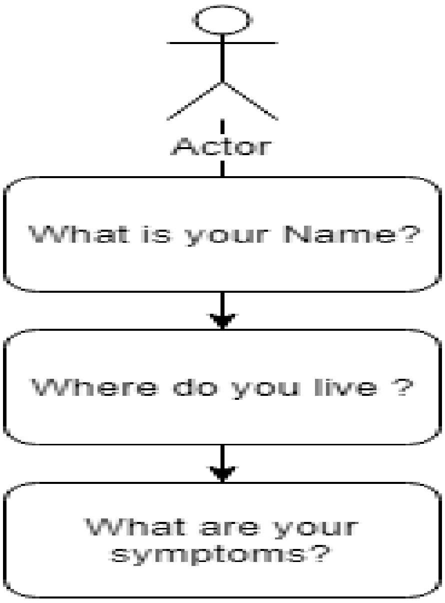

Fig 1: Flow of conversation with IVR

The flow of the conversation is as following in chronological order.

1. The user will be asked consent to record their conversation for training purpose, and also for quality purposes.
2. Later they are asked about their personal details, like address and country they live, this is an optional question the user or the potential patient can skip over this question if they want to.
3. Then they are asked to describe their symptoms in detail. The user or the potential patient can describe their symptoms like the following:

"I feel chill and nauseated "

2. Phase Two: Speech to text conversion

In two phase two, the recorded conversation is converted to text so that we can analyze it using the machine learning techniques and predict the diseases. But before we do the conversion the raw audio should be filtered for noise and silence after removing the noise and silence.

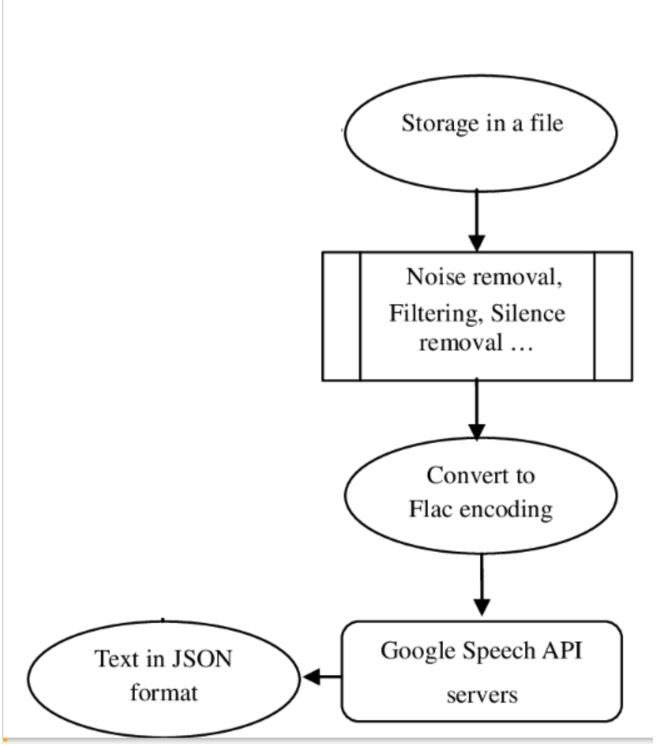
Storage in a file
Noise removal, Filtering, Silence removal …
Convert to Flac encoding
Google Speech API servers
Text in JSON format

Fig 2: Block diagram for call transcription.

The following steps take place in chronological order.

1. First, we retrieve the audio file which we store in our database.
2. Then we remove all the unwanted elements in our audio file such as noise and silence this is done by using Noise removing algorithms.
3. Then we encode the results into FLAC Free Lossless Audio Codec format, it is a lossless audio format,

meaning the audio is compressed in FLAC without any loss in quality.

4. Then we upload the resultant FLAC encoded audio format to the Google speech API.

5. Then we use the GET method to retrieve the JSON formatted output from the Google speech API which consists of the transcribed text of the conversation.

6. Store the resultant transcribed text from the audio conversation into our database

3. *Phase Three: Feature Extraction.*

In phase three we extract the feature which in our case is nothing but the symptoms , these symptoms are extracted from the transcribed text from the conversation that the user has with the Interactive voice system, then we check if the conversation which the user or the potential patient had with the interactive voice system is genuine or not this is done by using simple "bag of words", this model just checks If the conversation that the user or the potential patient had with the Interactive voice response system is sticking to the context of disease or not. To predict whether the conversation is genuine or not we use Fake or Not classifier, this classifier is developed using Gaussian Naïve Bayes

The following steps take place in chronological order to determine a conversation is genuine or not:

1. In this phase, we extract the symptoms out of the text.

2. But before we do that, we need to know whether the call we got is genuine or not.

3. So, we train "bag of words" model and use naive Bayesian classifier on it to predict if the call we got is genuine or not.

4. If Gaussian naïve Bayes classifier predicts that it is true, we go to next step, that is to extract the symptoms out the transcribed text which we get from the conversations.

We use the following pseudocode to implement the above steps:

```
Pseudocode:

Procedure FakeOrNotClassifierTrainer()

classifier = GaussianNB()

classifier.train(calls,fake/true)

End Procedure
```

In the above pseudocode, we use the Fake or Not Classifier Trainer procedure to train the naïve Bayes classifier with the call database which we have each call conversation in our database can be categorized as fake or genuine.

```
Procedure FakeOrNotClassifier()

call_predict = classifier.predict(call)

End Procedure
```

In the above pseudocode, we use the Fake or Not Classifier procedure to check if the call transcription is fake or not

If the call is genuine and the conversation sticks to the context of diseases, then we move on to the next step to extract the symptoms of the diseases

a. Symptom Extractor

After checking whether the call we received is genuine or not we extract the symptoms from the transcribed call text , but before we analyze the text we need to preprocess the text so that that unwanted elements in the text can be removed and we can extract only the useful elements out of the text, this is done in the following order first we use the stop words dictionary to compare it with our transcribed call text and find the similar words in the stop word directory and remove it if there are similar ones in both the stop words directory and our transcribed call text.

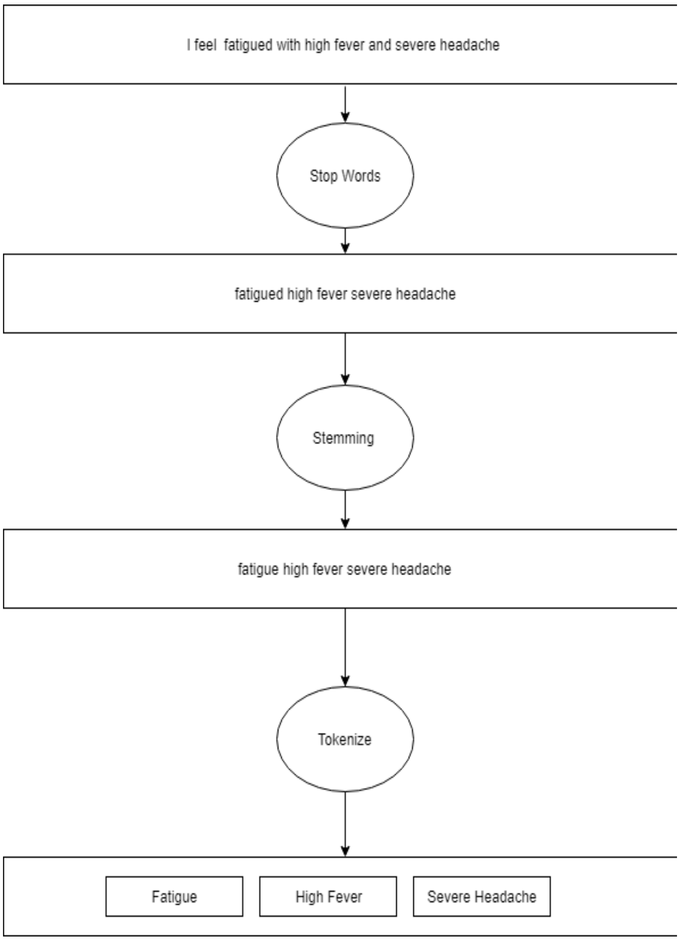
I feel fatigued with high fever and severe headache
Stop Words
fatigued high fever severe headache
Stemming
fatigue high fever severe headache
Tokenize
Fatigue
High Fever
Severe Headache

Fig 3: Symptom Extractor

The following steps take place in chronological order for feature extraction:

1. First, we load the text which we want to analyze.

   "I feel fatigued with high fever and severe headache"

2. Next, we use the stop words directory to remove all the unwanted words in our text after removing all the stop words in our text the following will be the result.

   "fatigued high fever severe headache".

3. After removing all the stop words from the text, next we stem the words to convert them into their root words this is done by using stemming. stemming is a process of converting of the words to their roots, by removing prefix and postfix attacked to them.
   I.e. The word "fatigued "becomes "fatigue".

   So, the final result after stemming becomes

   "Fatigue high fever severe headache"

4. Now we need to tokenize the text into separate words so can be used for analytical purposes, we use the tokenizer to do It

   After the tokenization the result will be

Fatigue, High fever, Severe Headache

The pseudocode for the feature extraction is as following

```
Prodedure SymptomExtractor()

  While(End oF file)

     if word in RAW_TEXT not in STOP_WORD :
     symptoms = word

   End While

 PorterStemmer(RAW_TEXT)
 Tokenize (n-grammer) the symptoms

End Procedure
```

4. Phase five Disease Classification:

In this phase, we use the resultant of the phase four which are the extracted symptoms from the conversation to predict the type of disease, in this phase we use gaussian naïve Bayes classifier to predict the type of disease

The pseudocode we use for the disease classification is as following

```
Procedure DiseaseClassifierTrainer()
  classifier = GaussianNB()
   classifier.train(diseases,symptoms)
End Procedure
```

In the above pseudocode, we create a naïve Bayes classifier and train it on the disease dataset consisting of both diseases and its symptoms

```
Procedure DiseaseClassifier()
  Pridicted_disease = classifier.predict(symptoms)
End Procedure
```

In the above pseudocode we use the trained classifier to predict the disease based upon their symptoms.

a. BAYESIAN CLASSIFICATION:

The set of classes (e.g. Diseases) C: = {Influenza, Ebola, ...,}, and the document consisting of Symptoms D: = {Chill, fever, headache ...}.

The objective of the classifier is to ascertain the probability that the Symptoms belong to some class C consisting of the diseases.

This is given that some set of training data associating documents and classes.

By using Bayes' Theorem, we propose that:

$$P(C|D) = P(D|C) * P(C/P(D).$$

The LHS (left-hand side) is the probability that the document consisting of diseases belongs to class C consisting of the symptoms given the document and the classifier will calculate this probability and outputs the most likely class (disease) for this document C consisting of symptoms.

P(C) refers to as the "prior" probability [23], or the probability that a Symptom belongs to C (Disease).

P(D|C) is the probability of seeing such a symptom, given that it belongs to C (disease). we assume that symptoms appear independently in documents, this being the "naive" assumption we can estimate

$$P(d|c) \sim= P(w\_1|c\_j)*...*P(w\_k|c\_j)$$

here P(w_i|c_j) is the probability of seeing the given symptom in a document of the given class (disease).

Finally, P(D) is a scaling factor and is relevant to classification, we introduce this factor to normalize the resulting

XI. CONCLUSION

A simple Interactive voice response system was proposed through which the user or the potential patient can contact and start a conversation with the system, the system will transcribe the entire conversation it had with the user or the potential patient into text format and analysis on the text is done using NLP and the symptoms are extracted, then machine learning techniques are used on extracted symptoms to predict the disease and it is recorded.

XII. REFERENCES

[1] "The World's Technological Capacity to Store, Communicate, and Compute Information". MartinHilbert.net. Retrieved 13 April 2016.

[2] Breur, Tom (July 2016). "Statistical Power Analysis and the contemporary "crisis" in social sciences". Journal of Marketing Analytics. 4 (2–3): 61–65. doi:10.1057/s41270-016-0001-3. ISSN 2050-3318.

[3] Laney, Doug (2001). "3D data management: Controlling data volume, velocity and variety". META Group Research Note. 6 (70).

[4] Goes, Paulo B. (2014). "Design science research in top information systems journals". MIS Quarterly: Management Information Systems. 38 (1): –.

[5] Marr, Bernard (6 March 2014). "Big Data: The 5 Vs Everyone Must Know".

[6] boyd, dana; Crawford, Kate (21 September 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.

[7] "Data, data everywhere". The Economist. 25 February 2010. Retrieved 9 December 2012.

[8] "Community cleverness required". Nature. 455 (7209): 1. 4 September 2008. Bibcode:2008Natur.455....1.. doi:10.1038/455001a. PMID 18769385.

[9] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.

[10] Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1.

[11] Jump up to:a b Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". Science. 332 (6025): 60–65. Bibcode:2011Sci...332...60H. doi:10.1126/science.1200970. PMID 21310967.

[12] "IBM What is big data? – Bringing big data to the enterprise". www.ibm.com. Retrieved 26 August 2013.

[13] Sh. Hajirahimova, Makrufa; Sciences, Institute of Information Technology of Azerbaijan National Academy of; str., B. Vahabzade; Baku; AZ1141; Azerbaijan; Aliyeva, Aybeniz S. "About Big Data Measurement Methodologies and Indicators". International Journal of Modern Education and Computer Science. 9 (10): 1–9. doi:10.5815/ijmecs.2017.10.01.

[14] Reinsel, David; Gantz, John; Rydning, John (13 April 2017). "Data Age 2025: The Evolution of Data to Life-Critical" (PDF). seagate.com. Framingham, MA, US: International Data Corporation. Retrieved 2 November 2017.

[15] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", December 2012

[16] Huser, Vojtech; Cimino, James J. (2016). "Impending Challenges for the Use of Big Data". International Journal of Radiation Oncology*Biology*Physics. 95 (3): 890–894. doi:10.1016/j.ijrobp.2015.10.060. PMC 4860172. PMID 26797535.

[17] O'Donoghue, John; Herbert, John (1 October 2012). "Data Management Within mHealth Environments: Patient Sensors, Mobile Devices, and Databases". Journal of Data and Information Quality. 4 (1): 5:1–5:20. doi:10.1145/2378016.2378021.

[18] Mirkes, E.M.; Coats, T.J.; Levesley, J.; Gorban, A.N. (2016). "Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes". Computers in Biology and Medicine. 75: 203–216. arXiv:1604.00627. doi:10.1016/j.compbiomed.2016.06.004. PMID 27318570.

[19] Murdoch, Travis B.; Detsky, Allan S. (3 April 2013). "The Inevitable Application of Big Data to Health Care". JAMA. 309 (13): 1351–2. doi:10.1001/jama.2013.393. ISSN 0098-7484. PMID 23549579.

[20] David Lazer1,2,*, Ryan Kennedy1,3,4, Gary King3, Alessandro Vespignani5,6,3 "The Parable of Google Flu: Traps in Big Data Analysis"

[21] Big data for infectious disease surveillance, modeling

[22] The Parable of Google Flu: Traps in Big Data Analysis

[23] "Naive Bayes text classification" Stanford NLP - Bayes Classifier