**Technical Report for a Boil-order Classifier Using Natural Language Processing Techniques on Google News English Articles**

Chris Mantell, General Assembly, December 2017

**Introduction**

The prolonged and grossly mishandled water crisis in Flint, Michigan, starting in 2014, has shown that there is a need for transparent and easily accessible information regarding water quality issues. Although Flint is on the correct path back to a city providing clean water, other cities and towns in the United States and its territories are potentially suffering from tainted water, partially because information is not being circulated effectively as it eventually was in Flint. In 2017, the citizens of Moore Bend, Missouri, were notified of a standing boil order alert that was in place since 20131. This is just one extreme example and reason why boil orders should be tracked. This technical report describes a supervised machine learning classification model to automatically detect boil-order alerts using Natural Language Processing (NLP) on online news articles.

**Methods**

**Data Acquisition**

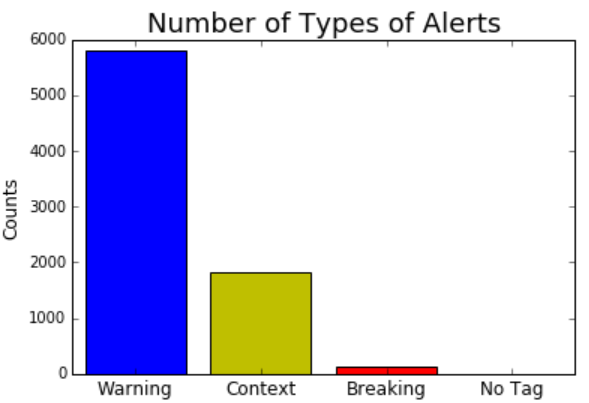
Data for this project was provided by the HealthMap, a disease outbreak and epidemiological surveillance tool. The HealthMap collects information from many various sources in their data feeds (ex: Google News, World Health Organization (WHO), Baidu News, etc.) and automatically classifies it by disease or disease category (ex: Avian Influenza H7N9, Fever, Environmental, Foodborne Illness, etc.). Information collected from the tool is curated by humans knowledgeable in public health and edits the data to increase information accuracy and/or correct information that is misclassified by the HealthMap. A .csv file of “Waterborne Illness” alerts in the United States from the Google News English feed was obtained on 10/27/17 and contained data from 5/10/10 – 10/27/17. Each row was a different location where an alert has taken place. The columns included a URL to the original link pulled from the HealthMap that held the news article referencing the location, a headline of the article, an “Issue Date” of when the article was written, and a “Smooshed Parser Extract”, which is an automated extraction of text from the article. The original dataset contained 7725 rows and 119 columns, which contained information describe above as well as many unnamed columns resulting from a mistake or an anomaly in data entry. Some articles referenced multiple locations and in this case, different rows reference the same article.

**Data Cleaning and Exploratory Data Analysis (EDA)**

After data cleaning, the dataset consisted of 7724 rows and 14 columns as described in the data dictionary (Table 1).

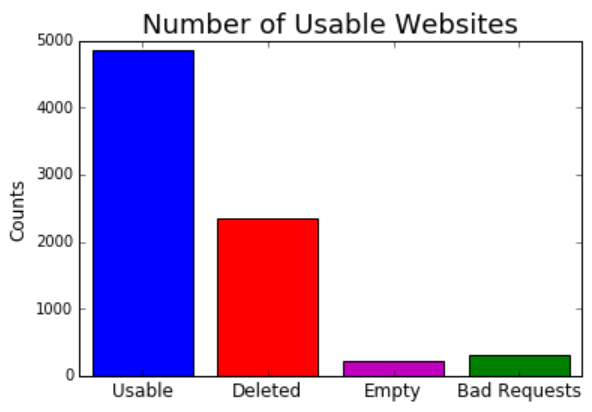
The author of this technical report was previously a data curator from 2016 – 2017 and had previous knowledge about what kind of information was brought into the “Waterborne Illness” alerts. Most alert relating to boil-orders are classified as “Warning” under the   
“Alert Tag.” Value counts of the type of alert tags are graphed as shown in Figure 1.

|  |  |
| --- | --- |
| **Data Dictionary** | |
| **Column** | **Description (Example)** |
| “Location” | Location where the alert is referencing (Delaware, United States) |
| “Country” | Country where the alert is referencing (United States) |
| “Disease” | Disease the alert is referencing (Waterborne Illness) |
| “HM Alert” | Unique HealthMap (HM) alert ID (5407609) |
| “Headline” | The Headline of the news article (‘Kutcher tells Council about water situation - Murray Ledger and Times’) |
| “URL” | URL pulled by the HM referencing the original article at the date it was pulled (http://news.google.com/news/url?sa=t&fd=R&ct2=us&usg=AFQjCNGmA0yAoTj-LEPjIRxm5NB9akVvVw&clid=c3a7d30bb8a4878e06b80cf16b898331&cid=52779651867230&ei=Cc7yWaC\_H9G7zAKcpws&url=http://murrayledger.com/news/kutcher-tells-council-about-water-situation/article\_2020569a-bac6-11e7-82f8-1f4b918f38ae.html) |
| “Issue Date” | The date and time the web article was published (10/27/17 1:11) |
| “Alert Tag” | The classification of the type of alert. This ranks the alert in importance or whether it references information relating to a disease but not necessarily an outbreak (Breaking, Warning, Content, No Tag) |
| “Dup Count” | The number of duplicate articles referencing this alert |
| “Long” | Longitude of the location of the alert |
| “Lat” | Latitude of the location of the alert |
| “Smooshed Parser Extract” | Automated text extracted from the article (‘ XXXXXX Murray Ledger and Times) |
| “Place Category” | An extra categorization column to help give more information about the details of the alert (‘Natural disaster related’) |

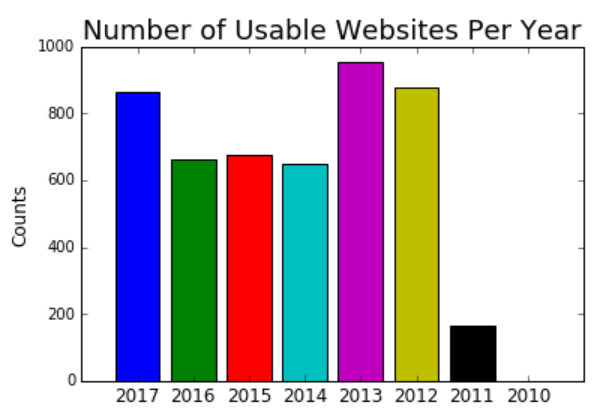
**Table 1:** A description of the different columns in the original dataset after initial cleaning.

**Figure 1:** Counts of alerts grouped by “Alert Tag”

Initially, the “Smooshed Parser Extract” column was thought to be the text of the articles in the alerts but the data turned out to be missing many rows with 5,171/7,725 (66.9%) having non-null values. Furthermore, the values present were of questionable use to determine boil-orders. To get the text of the articles, each URL was scraped to get the raw HTML/CSS and the results for each row were appended to the dataset. Each successful request that returned HTML/CSS results was then ran through custom built regular expression parser to target HTML paragraph tags and then remove undesirable text that was left, such as HTML tags and strings of non-predictive value (ex: LinkedIn, ‘’&amp’). The resulting text was used to make new features for modeling.

Only 4,849/7725 (62.8%) returned text while 2348/7725(30.4%) sites were deleted (404 or 410 errors with requests), 299/7725 (3.8%) had bad requests (400 or 500 errors other than 404 or 410), and 228/7725 (3.0%) sites had a successful request but retuned no text. The results of the web scrapping are showed in Figure 2 and the number of usable websites by the year of “Issue Date” are shown in Figure 3.

**Figure 3:** Results of scraping useable text from the dataset



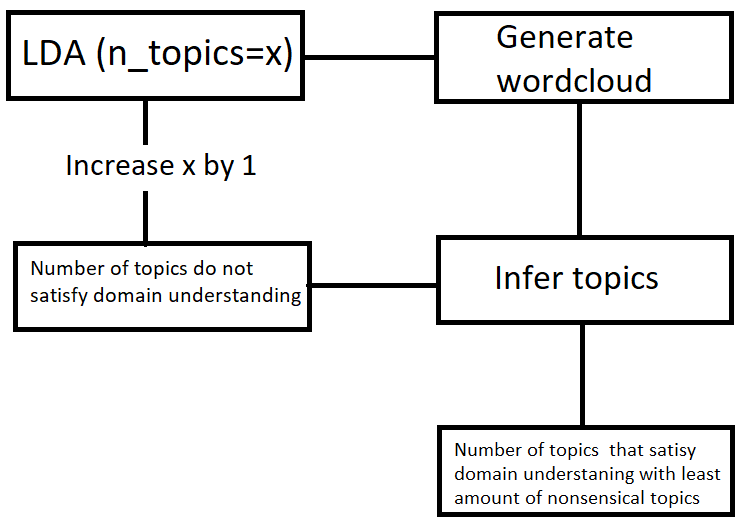
**Figure 3:** Results of scraping usable text from the dataset grouped by year of “Issue Date”

**Feature Creation and Extraction**

The results of the parsed text from the web scraping was preprocessed to remove stop words, remove punctuation, remove numbers, stem the words, and then the resulting text was put into a count vectorizer with a minimum document frequency (min\_df) of five articles. This was done set low to avoid removing potential important words that could differentiate between boil-orders and non boil-orders. 10,407 words were returned including non-descriptive and nonsensical text. This was used to run a Latent Dirichlet Allocation (LDA) unsupervised machine learning technique.

Because the results of the web scraping yielded a substantial amount of missing data and non-descriptive text, another set of features were extracted for a separate model building pipeline using the Headline text of the entire dataset. The same preprocessing steps were applied as above and then ran through a count vectorizer with a min\_df of two articles to remove non-descriptive text. The resulting text was used for a separate LDA.

LDA was performed on the web-scraped text and the headlines. Wordclouds were created for each topic. These, along with the top 25 most frequent words per topic were used to infer a category based on the author’s domain knowledge of what kind of articles are pulled into the “Waterborne Illness” section of the HealthMap. This process was repeated, each time increasing the number of computer generated topics by one, until the author could infer multiple nonsensical topics. The final number of topics was the minimum number resulting in the least number of nonsensical topics. Figure 4 outlines this process. The resulting number of topics four with the web scraped data and five for the headline data. The process assumes that LDA should be able to generate topics associated with boil-orders and other commonly reported topics relating to “Waterborne Illness” until there are only nonsensical topics left.



**Results**

**Figure 4**: Schema to infer topics with LDA from text

The results of the LDA was used to transform the text data into an array of probabilities. The results are that each data point of scraped text or headline has a list of probabilities of belonging to each LDA generated topic. The data point with the highest probability of each topic was extracted, read, and compared with the topic inferred by the author to validate the LDA generated topic.

The extracted text data did not pass this validation test but the headline data did. The headline data was further validated by appending each inferred topic to the dataset based on the highest probability of belonging to the associated LDA generated topic. Randomly selected headlines were read and compared with each associated inferred topic. Many headlines were congruent with the generated/inferred topics. This was used as a ground truth for a classification model.

Only 4,849/7725 (62.8%) returned text while 2348/7725(30.4%) sites were deleted (404 or 410 errors with requests), 299/7725 (3.8%) had bad requests (400 or 500 errors other than 404 or 410), and 228/7725 (3.0%) sites had a successful request but retuned no text. The results of the web scrapping are showed in Figure 2 and the number of usable websites by the year of “Issue Date” are shown in Figure 3.

1. Adler, P. (2017). 4-year Boil Water Order frustrates subdivision in Taney County. Retrieved December 15, 2017, from [http://www.ky3.com/content/news/Ozarks-Town-spends-4-years-under-Boil-Water-Order-415694043.html. Updated 3/9/17](http://www.ky3.com/content/news/Ozarks-Town-spends-4-years-under-Boil-Water-Order-415694043.html.%20Updated%203/9/17)