A Framework for Detecting Public Health Trends with Twitter

Jon Parker [1, 2], Yifang Wei[1], Andrew Yates[1], Ophir Frieder[1], and Nazli Goharian[1]

[1]Department of Computer Science
Georgetown University
Washington DC, USA
{jon, yifang, andrew, ophir, nazli}@cs.georgetown.edu

[2]Department of Emergency Medicine
Johns Hopkins University
Baltimore, USA
jparker5@jhmi.edu

Abstract—**Traditional public health surveillance requires regular clinical reports and considerable effort by health professionals to analyze data. Therefore, a low cost alternative is of great practical use. As a platform used by over 500 million users worldwide to publish their ideas about many topics, including health conditions, Twitter provides researchers the freshest source of public health conditions on a global scale. We propose a framework for tracking public health condition trends via Twitter. The basic idea is to use frequent term sets from highly purified health-related tweets as queries into a Wikipedia article index – treating the retrieval of medically-related articles as an indicator of a health-related condition. By observing fluctuations in frequent term sets and in turn medically-related articles over a series of time slices of tweets, we detect shifts in public health conditions and concerns over time. Compared to existing approaches, our framework provides a general a priori identification of emerging public health conditions rather than a specific illness (e.g., influenza) as is commonly done.**

Keywords—Twitter, health surveillance, item-set mining, Wikipedia

I. INTRODUCTION

Social media allows users to be both active consumers and producers of information. This new style of communicating has shown unprecedented levels of uptake and growth. For example, in 2012 the number of Twitter users broke the 500 million users threshold [3], and the number of tweets published per day reached fifty million in early 2010 [2] a number that is still growing rapidly. Much of the content published on social media, and Twitter in particular, contains personal opinion on trending topics. This characteristic enables Twitter to provide instant access to public opinions on trending topics at a global scale.

Twitter has been shown to be a reliable source for tracking public opinion about topics that range from political issues [24, 30], to natural disasters [28], and even brand sentiments [20]. Personal health is also actively discussed in social media. People with chronic diseases like cancer are using social media to discuss their health, share stories, and provide peer-to-peer help with increasing frequency [10]. A recent survey revealed that 26% of "online" US adults discussed their health issues online in the past 12 months, and 42% of them use social media to post or seek information about health conditions [1]. These facts suggest that social media content reflects, at least in part, public health conditions and can potentially serve as a foundation for public health surveillance systems.

Traditional public health surveillance systems are typically managed by professional health institutions. For example, the Center for Disease Control and Prevention (CDC), gives early warnings of epidemic outbreaks that typically incur a one-to-two week reporting delay [17]. These systems also require regular clinical reports and considerable effort by health professionals to analyze data. A more recent alternative approach is the Global Public Health Intelligence Network (GPHIN). GPHIN captures epidemic outbreaks by monitoring global media sources, essentially news websites, to supply approximately 40% of the World Health Organization’s (WHO) early warnings [22]. Given that GPHIN's success is largely attributed to the incorporation of comprehensive information from global news websites [22], it is reasonable to infer that social media – and Twitter in particular – could enable the creation of a variety of low-cost (as compared to non-automated approaches) alternative public health indicators that serve as the basis for public health surveillance systems.

This thinking is supported by a number of Twitter-based public health monitoring approaches [6, 13, 14, 17, 18, 20, 26, 27, 32]. Nevertheless, most efforts focus on detecting a pre-established health condition (e.g., influenza or insomnia) based on an existing assumption that the condition is present. In contrast, we propose a general framework for identifying emerging health conditions without prior knowledge of a condition’s existence. In other words, while other approaches answer the question "Is such-and-such illness a prevalent health condition?", we answer the more general question "What health conditions are prevalent?" (with an answer that may include, but is not limited to the condition presupposed by other approaches).

Our framework consists of: health-related tweet extraction from a large comprehensive corpus of tweets, frequent word set generation, frequent word set trend tracking over time, connecting frequent word sets to Wikipedia articles, and filtering Wikipedia articles according to health relevance.

We start by extracting health-relevant tweets from a tweet corpus. The initial corpus, created by Paul and Dredze [26] and used herein, consists of 2 billion tweets, filtered three times to yield 1.6 million health condition specific tweets. From the health-related tweets, our method finds frequent word sets. Then, a Wikipedia article index is used to evaluate the relevance of each of the obtained frequent word sets to health, while monitoring the fluctuation of the health-relevant word-sets that might indicate trending health conditions.

Ideally, a long-term goal of creating an automated general purpose public health trend detector is to make a concrete impact on health outcomes. This goal necessitates an efficient detection method so that planners and decision makers can get "in front of" a health crisis. There is a vast body of disease simulation literature that seeks to clarify public health decision like "Should schools be closed?" [10] and "Should international travel restrictions be put in place" [19]. Simulation techniques are now fast enough [29] that if accurate and timely disease surveillance data were available then better public health decisions could be made with less angst. Note, however, that regardless of which disease surveillance methods are used, there will always be public health officials vetting and inspecting the surveillance data.

The most salient features of the proposed framework are:

- The ability to capture emerging health conditions without a priori knowledge of condition existence.
- Simplicity and efficiency to be of practical use.

II. RELATED EFFORTS

Many efforts focus on tracking epidemics with tweets. Most of these efforts target the detection of influenza. Early work by Corley, et al. directly correlate occurrence of text which contain manually picked influenza-related words with official data [13] (e.g., correlating the occurrences of the blog posts containing influenza or flu with Influenza Like Illness (ILI) rates). Similarly, Ginsberg, et al. [17] show compelling evidence of correlation between the occurrence of search queries containing flu-related words and ILI rates.

To reduce human involvement and explore the entire feature space, Culotta [14] proposed a model for automatically selecting textual features useful for labeling tweets as health-related, which are later employed in tracking ILI rates. An improved version by Lampos, et al. [20] employs a bootstrapping algorithm to extract a set of textual features from a tweet corpus using different feature selection principles. Additionally, Aramaki, et al. [6] train a support vector machine to label tweets as flu-related or flu-unrelated, and then evaluate the correlation of flu rates and flu-related tweets.

Rather than correlating the occurrence of flu rates and flu-related tweets, Wenerstorm et al. [32] proposed a summarization method for flu-related tweets. According to their method, each flu-related tweet is represented with a vector of probabilities, each component of which corresponds to the tweet's probability of coming from a particular topic. A pairwise similarity value between tweets is derived from tweets' probability vectors, based on which tweets are clustered in a hierarchical or an agglomerative way. Tweets within the same cluster are ranked using closeness centrality, and common words of top ranking tweets summarize the cluster. When a Twitter monitoring system based on counting flu-related tweets signals a flu outbreak alarm, the summarization system will allow health officials to quickly verify outbreak alarms.

Besides assisting in the influenza detection system, Twitter is employed to study and monitor other ailments and health concerns. Jamison-Powell, et al. [18] conducted a thematic analysis of insomnia-related tweets to reveal the degree to which people are using Twitter to discuss their mental health and how exactly they are doing it. Nakhasi, et al. [23] investigated patient perspectives on medical errors by exploring Twitter messages for self-reported adverse medical events. Diaz-Aviles, et al. [15] presented a personalized tweet ranking algorithm that could provide users a personalized, short list of tweets based on his or her own tweet context. Zhu and Goharian [35] also report personalize twitter information. White et al. [33] analyzed web search logs as opposed to personal twitter feeds to extract information about adverse drug reactions.

While all the above research targets a specific illness or health concern, a system capable of monitoring multiple ailments and health concerns is of more practical use. One appealing class of techniques for extracting information spanning across multiple health conditions is probabilistic topic modeling. Techniques within this class include Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and Non Negative Matrix Factorization (NNMF). These methods model the association of terms with hidden topics, and view documents as a multinomial mixture of hidden topics [9]. Nevertheless, the topics discovered from these topic-modeling approaches still need to be manually evaluated [7]. In fact, a topic generated with topic modeling approaches is often representative of mixed content, while seldom corresponding to a specific concept (e.g., an ailment) [26].

As a step toward overcoming this limitation, Paul and Dredze proposed the Ailment Topic Aspect Model (ATAM) in [26], which could isolate various ailments within a tweet corpus. Although ATAM derives from LDA, it can output much more coherent ailments, such as obesity, respiratory, and dental. Similar to LDA, ATAM contains parameters that require tuning. Their tuning relied on a specially focused corpus they developed. Although the method described herein differs vastly from the ATAM approach, we use their corpus in our evaluation.

Different from the approaches focusing on a specific health condition, our approach efficiently identifies general emerging health conditions, rendering it of practical interest.

III. TWITTER CORPUS

ATAM, as described previously, involved a corpus of 1.6 million health-related tweets culled from a much larger corpus of 2 billion tweets. The larger corpus of tweets was collected by [24] and contains tweets from May 2009 to October 2010.

The work presented here uses this same health-specific corpus. This narrowly focused corpus [26] was created by

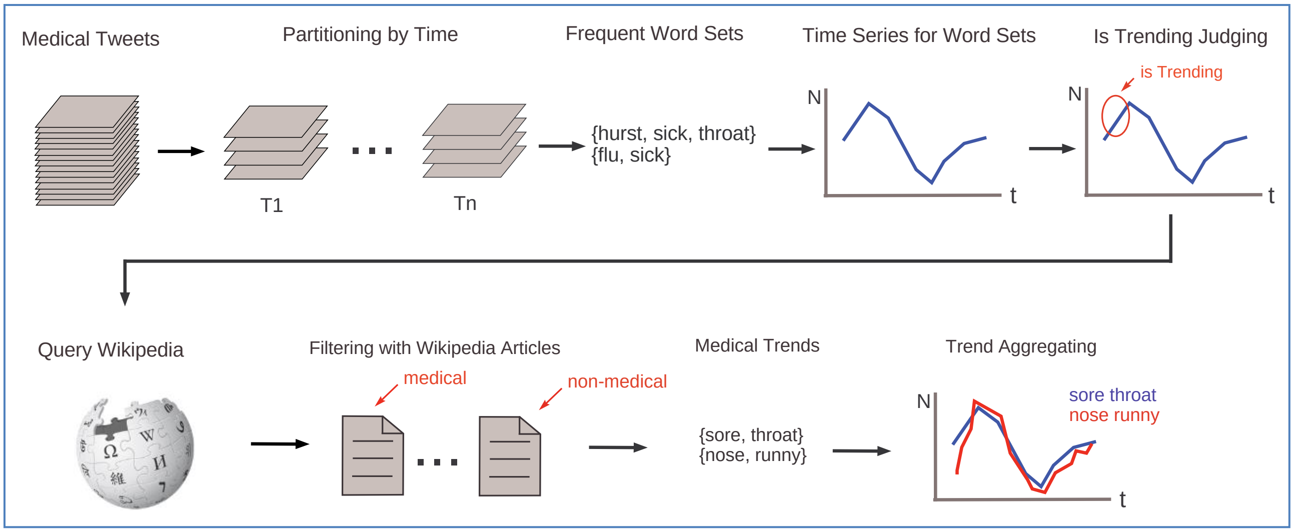


Figure 1: A High-Level View of the Core Algorithm

removing 99.92% of the content from the larger twitter corpus through a multiple pass filter. The first filtering pass removed tweets that did not contain at least one of 20,000 key phrases related to illnesses/diseases, symptoms, and treatments scraped from wrongdiagnosis.com/lists/{symptoms,condsaz,treats}.htm and mtworld.com/tools_resources/commondrugs.php. The second pass removed re-tweets and tweets containing URLs. The last, and arguably most important, filtering operation applied a custom built SVM classifier. The SVM classifier was trained using data collected from Mechanical Turk and was designed to favor high precision over high recall.

IV. THE FRAMEWORK

The goal of our framework is to automatically detect emerging public health concerns using Twitter. We want to do this without designating a priori which public health concern(s) is (are) most important. In other words, we want to interact with our system to discover emerging public health concerns (e.g., “Question: What illnesses seem to be occurring more frequently lately? Answer: Flu”) rather than providing feedback on a user-specified health concern (e.g., “Question: Is flu occurring more frequently lately? Answer: Yes”).

Our framework is based on a core assumption that people will describe the chief complaint (i.e., primary symptoms) of an illness on Twitter. Our framework is designed to find illnesses that are sometimes associated with chief complaints that are commonly discussed on Twitter.

To provide the desired capability our framework leverages three mature open-source resources: Mahout, Lucene, and Wikipedia. The parallel FP-Growth [21] implementation in Mahout is used to find frequent word sets. Wikipedia searches are performed to associate frequent word sets with Wikipedia articles. These searches are performed programmatically using a Lucene index containing the complete database of Wikipedia articles. Finally, the Wikipedia articles returned from a search are filtered so only medically relevant articles are highlighted. A high-level view of the algorithm is shown in Figure 1 with a detailed description of each section described in the following subsections, respectively.

A. *Partitioning the Corpus by Time*

The first step towards implementing our framework is to partition the Twitter corpus into multiple month-corpuses based on the month in which each tweet was authored. Partitioning the corpus in this way enables us to mimic the flow of incoming monthly data dumps. We opted to partition the Twitter corpus by month, as opposed to weeks, to reduce the temporal variability in which word sets are considered “frequent” word sets.

B. *Finding Frequent Word Sets*

Before frequent word sets can be found, the tweets within a month-corpus must be standardized. The raw text of each tweet is standardized using the following operations:

- Punctuation characters are replaced with spaces
- All text is converted to lowercase
- The text is tokenized
- Stop words are removed
- Duplicate tokens are removed

After standardization, each tweet is treated as a set of words that can be analyzed using off-the-shelf association rule mining techniques [5]. In particular, we use the parallel FP-Growth implementation within Apache’s data mining library Mahout to find the frequent word sets within each month-corpus. We vary the minimum support used when mining each month-corpus to ensure that the conceptual definition of “frequent” remains constant from month-corpus to month-corpus. The minimum support is always set to the smallest integer *n* such that *n* is at least 0.1% of the tweets within that particular month-corpus. Consequently, any set of words that does not reach this threshold will not be detected using the current parameter settings.

C. *Creating Time Series for Word Sets*

After mining a month-corpus we have a collection of frequent word sets like {{flu, sick}, {headache, feel}, {hurts, sick, throat}, {feeling, stomach}…}. For each frequent word set, we build a time series that shows how prevalent that particular word set is over time. An example is shown in Figure 2. These time-series are used to determine which word sets have recently seen a significant increase in prevalence, that is, which word sets are trending.

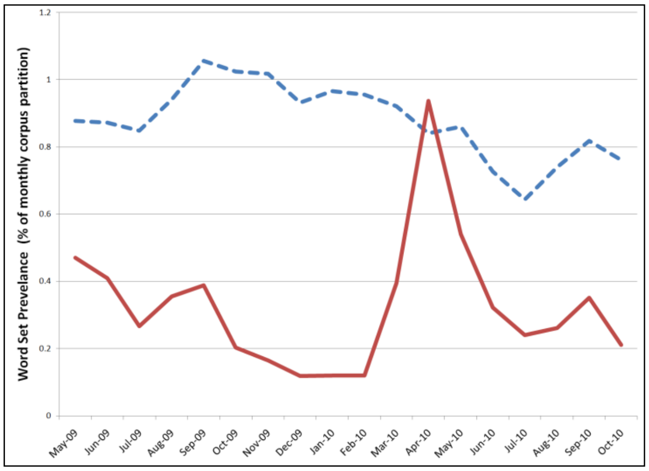


Figure 2: Prevalence of Two Frequent Word Sets by Month: Solid Line = "allergies feel", Dashed Line = "feel sick"

D. *Make "Is Trending" Decisions*

Clearly, we cannot detect potentially interesting trends in Twitter data merely by observing that some word sets are common. In fact, the word set {feel, sick} is the most prevalent word set in every monthly partition (see IV.A) of our dataset (keep in mind, stop words have been removed and the SVM filter was designed to find a specific flavor of tweet). When deciding whether or not a frequent word set "is trending" we use the rule:

$$isTrending(t) = \left(isFrequent(t)\ AND \left(\frac{prevalence(t)}{prevalence(t-1)} > GROWTH\_RATE \right) \right)$$

The purpose of the *isFrequent*(time t) clause is to discourage false positives by ensuring that the prevalence at time *t* is large enough to make the local derivative less subject to noise (see Section IV.B for more). The GROWTH_RATE parameter was set to 1.5.

We realize that making an "is trending" decision about a particular word set can be considered its own research topic. One interesting problem in this domain is whether the "is trending" decision should reflect absolute counts or a proportional count. This section makes no claim about how to best make the "is trending" decision; it merely reports our methodology.

E. *Query Wikipedia*

We use Wikipedia to associate trending word sets with the topics found in Wikipedia. Wikipedia was selected because of its wide coverage and the fact that it is written in layman's English (closely resembling the tweets considered). Later we filter out topics that are not pertinent to public health. We built a searchable index of Wikipedia using the high quality, mature open-source Lucene package. The complete Wikipedia compressed archive we indexed was found at http://en.wikipedia.org/wiki/Wikipedia:Database_download).

Before each Wikipedia article is indexed, we parse it and store the article introduction and any info boxes if they exist. We explicitly store these two fields because they are used to determine which Wikipedia articles may be relevant to public health. Figure 3 shows a typical health-related article that contains an info box mentioning International Statistical Classification of Diseases and Related Health Problems (ICD) codes.

The index is built using the StandardAnalyzer from Lucene version 3.5. Once the fully built index is in hand we push every frequent word set through our Lucene Wikipedia search system.

F. *Filtering Wikipedia Results*

Many frequent word sets have no obvious connection to public health concerns. For example, the frequent word set {big, time} does not look nearly as likely to generate health related topics as the word sets {sore, throat}, and {allergies, hate}. As expected, none of the top 50 Wikipedia articles returned from a "big time" query relate to health while many of the articles returned from "sore throat" or "allergies hate" queries have a health angle to them.

Because it is difficult to programmatically determine a priori which word sets will generate health related topics, we convert every frequent word set to a query and filter the Wikipedia articles Lucene returns. Two filtering methods were considered. The first method only returns Wikipedia articles that contain ICD codes. The second method returns Wikipedia articles that contain ICD codes and articles with introductions that contain a large proportion of medically related words.

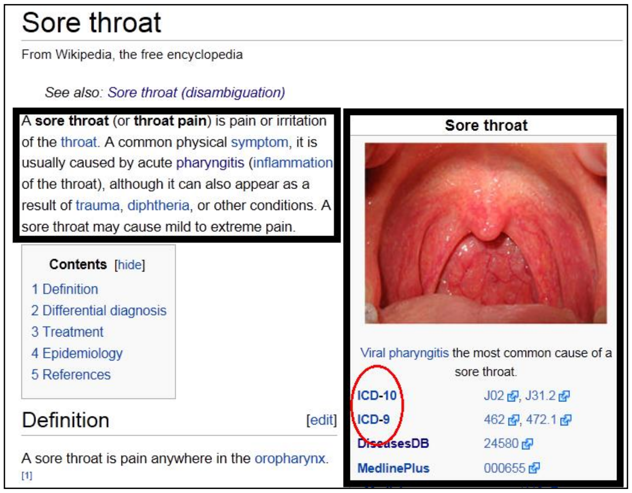
1) Precision Filter:

The first filtering method used to differentiate health-related Wikipedia articles from non-health-related articles is based on the presence (or lack thereof) of an ICD code within the article. The ICD coding system is an international standard classification system that has been used extensively to encourage inter-operability of medical and insurance computer systems. The 10th revision of ICD, ICD-10, contains over 14,440 different codes distributed across different sub-classes like diseases and medical procedures. Figure 3 shows a typical Wikipedia article that has an info box containing an ICD code. Finding an ICD code within an info box is a strong indicator that the article is medically relevant. The strength of this required indicator ensures that the set of articles that pass this filter will have a significant health aspect to them.

2) Recall Filter:

The second filtering method we consider is more inclusive and so its recall is higher than the prior filter. This second filtering method accepts every article that the precision filter accepts as well as articles containing "medically relevant" introductions.

When we used the term "introduction" we must be careful because Wikipedia articles do not have an officially labeled Introduction section. However, Wikipedia articles generally do have labeled sections. The "Sore Throat" article, a portion of which is shown in Figure 3, has the following 5 sections: Definition, Differential Diagnosis, Treatment, Epidemiology, and References. We classify any text that comes before the first labeled section as the introduction of that article. We do not include info boxes as part of the introduction even though the text that defines them appears before the first labeled section.



Sore throat

From Wikipedia, the free encyclopedia

See also: Sore throat (disambiguation)

A **sore throat** (or **throat pain**) is pain or irritation of the throat. A common physical symptom, it is usually caused by acute pharyngitis (inflammation of the throat), although it can also appear as a result of trauma, diphtheria, or other conditions. A sore throat may cause mild to extreme pain.

Contents [hide]

1 Definition
2 Differential diagnosis
3 Treatment
4 Epidemiology
5 References

Definition [edit]

A sore throat is pain anywhere in the oropharynx.[1]

Sore throat

Viral pharyngitis the most common cause of a sore throat.

ICD-10	J02, J31.2
ICD-9	462, 472.1
DiseasesDB	24580
MedlinePlus	000655

Figure 3: A Snippet from a Typical Wikipedia Article: The introduction and info box are enclosed in rectangles. The ICD codes are circled.

Once an article's introduction is isolated, we analyze the introduction to determine if it is discussing a "medically relevant" topic. To make this determination we:

- Tokenize the introduction
- Remove stop words
- Count the tokens
- Count the medical tokens
- If: the overall token count ≤ 10
  Then: Return "is not medical"
- If: numMedicalTokens / numTokens ≥ .75
  Then: Return "is medical"
  Else: Return "is not medical"

The steps shown above require the ability to determine if an individual token is medical. We make this determination by searching for the token in Stedman's Medical Dictionary available online at http://www.medilexicon.com/medicaldictionary.php.

G. Aggregating Trends

It is possible – and indeed likely – that multiple word sets will be associated with the same Wikipedia article. For example, {sore, throat}, {nose, runny}, and {cough, nose} will all contain the "Common Cold" article within their respective query results. These three word sets can also be designated as trending word sets at different times. For each Wikipedia article, we aggregate the trending times generated by all the word sets that highlight that particular article. This aggregation helps us differentiate between spurious illness detection and illness detection that has been confirmed using multiple word sets.

V. RESULTS

Our results confirm that seasonal increases in common health conditions are indeed detectable without using search strategies customized to detect those specific health conditions. In particular, we detect (among other things) allergy season, flu season, and even summertime ice-cream headaches (i.e. "brain freeze") using one general purpose algorithm. Our results also illustrate that our methodology is likely to highlight multiple medical conditions with similar symptoms as opposed to highlighting just one or two conditions that could be considered the "best response" for a particular trending word set. For example, several different types of headaches are simultaneously detected as are multiple respiratory ailments like influenza, the common cold, cough, and acute bronchitis.

A. A Sample Detection: Influenza

The curve in Figure 4 shows the number of times the "Influenza" Wikipedia article is associated with a trending word set. By comparing the system results in Figure 4 with true influenza incidence shown in Figure 5, we can see that our detection framework produces the weakest signal (i.e., the smallest values) when the slope of the true incidence is negative. Our detection scheme also produces its strongest signal when the slope of the true incidence curve is strongly positive (in Sept and Oct of 2009). The beginning of the mild 2010 flu season also coincides with an uptick in Figure 4.

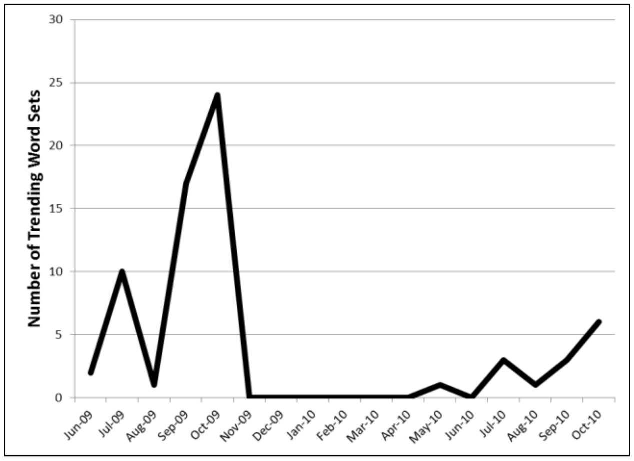


Figure 4: The Number of Trending Word Sets Associated with the "Influenza" Wikipedia Article

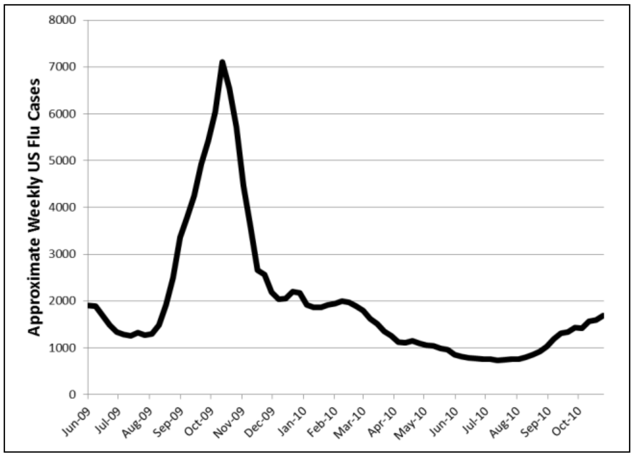


Figure 5: Approximate Weekly Flu Cases in the United States From June 09 – October 2010 [4]

In an ideal world, any non-zero entry in Figure 4's curve would indicate real world influenza cases were indeed growing in number. However, this is not the case. The moderately strong detection signal seen in July of 2009 (when "Influenza" was associated with 10 trending word sets) does not correspond to a simultaneous increase in US flu cases. We attribute this data point to the notable increase in flu interest that occurred after the WHO raised the worldwide pandemic alert level to Phase 6 on June 11th of 2009. It is possible that much of the lag from June 11th to July can be accounted for by the reporting delay for official CDC flu incidence numbers which typically required one-to-two weeks to gather, tabulate, and publish.

It should also be noted that the comparison between the curves in Figures 4 and 5 is subject to one small caveat. Our corpus of 1.6 million tweets was not explicitly filtered to contain only US based tweets. However, we do not believe this is a significant problem in practice because tweets published with geographic data are extremely likely to have originated from within the US.

B. Precision Filter vs. Recall Filter

In section 4 we mention that two different filters are used to separate health-related Wikipedia articles from non-health-related articles. We believe the precision focused filter that requires an ICD code to be within the article, is preferable to the recall focused filter which accepts either an ICD code mention or medically related terms in the introduction. The recall focused filter allows a few obviously non-health related articles through but the majority of the additional articles merely define a body part or system (e.g., Mucous, Nasal cartilages, Cough reflex). Although the identification of a body part or system does provide additional information, it fails to further identify a general trended condition. Since, we aim to identify a general health diagnosis we prefer the precision focused filter over the recall focused filter.

C. Confounding by Symptoms and Syntax

Our methodology highlights 11 different articles having to do with one respiratory ailment or another. It also highlights 12 different Wikipedia articles that pertain to headaches and migraines. The interesting difference between these two groups is that the existence of each "family" is driven by markedly different phenomenon. The group of respiratory results is created by tweets the describe symptoms. For example, "runny nose" and "sore throat" both highlight multiple respiratory conditions when those word sets are trending. The batch of headache results is driven by the two different meanings of the word headache: physical pain (e.g., "I bumped my head and now I have a headache") and annoyance (e.g., "My computer crashed – what a headache"). As a result of these disparate drivers the signals associated with the family of respiratory results have a much better cohesion than the signals associated with the family of headache results.

Although the batch of headache results is confounded by the colloquial use of the word "*headache*", some promising news within that collection exists. The detection curve for the article "Ice-cream headache" shown in Figure 6 has significantly smaller values than almost all of the other headache related articles like "Vascular headache" and "Tension headache" (the migraine articles also show these reduced absolute values). The reason for this is that many word sets containing the word headache do not flag the "Ice-cream headache" article. This is good because the signals associated with the word sets {eating, headache}, {headache, ice}, and {cream, headache, ice} (among others) are not drowned out by the multitude of signals emitted by the colloquial use of the word headache.

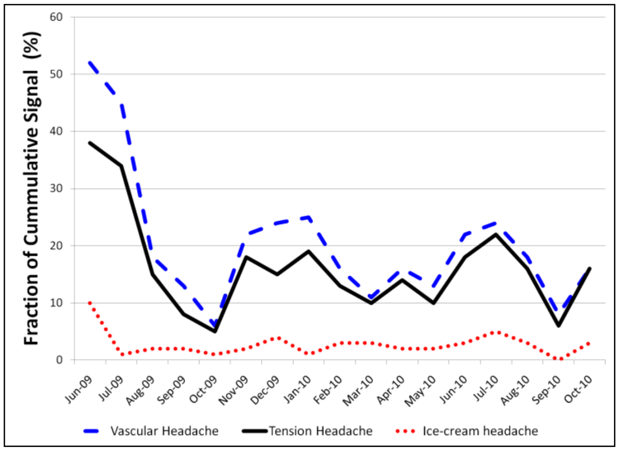


Figure 6: Relative Detection Signal Strength of Different Types of Headaches

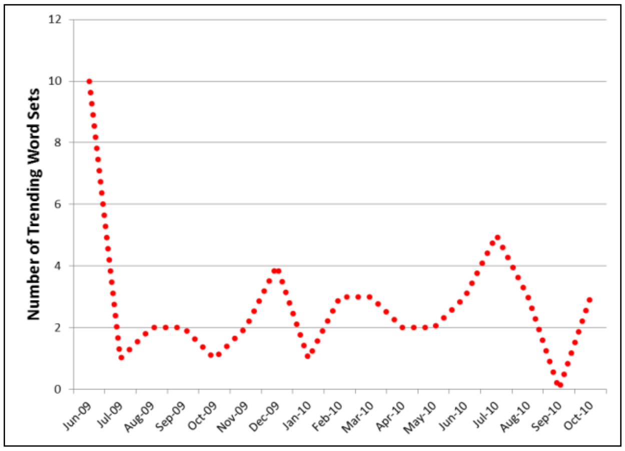


Figure 7: The Number of Trending Word Sets Associated with the "Ice-cream Headache" Wikipedia Article

This raises the confidence in the pleasant result that "Ice-cream headaches" are flagged as trending in June of 2009 and July of 2010 (which, at the time, was the hottest month on record in many places throughout the US). The peaks in the dotted curve of Figure 6 in June of 2009 and July of 2010 are significantly more noticeable when the dotted curve is plotted by itself as shown in Figure 7.

One curious result comes from the Wikipedia articles getting highlighted due to word sets like: {allergies, lol}, {allergies, asthma}, and {allergies, eyes, itchy}. These 3 word sets (and many similar word sets) all trend during the early spring. From the word sets themselves and the time those word sets trend it is clear the underlying condition is the pollen related allergies that are prevalent during the spring. On a positive note, we detect multiple seasonal allergy related Wikipedia articles – 2 of which are shown in Figure 8. The problem is that multiple food allergies are also highlighted as trending medical conditions. It is possible that a medical synonyms set as in [34] may prove useful when addressing the problems that common symptoms present.

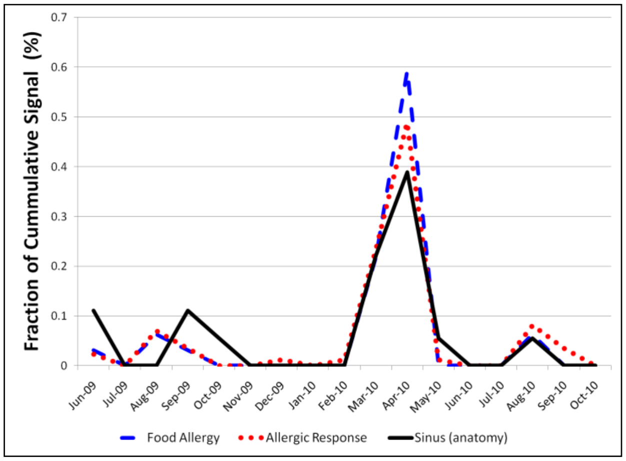


Figure 8: Allergy Related Results

D. Duplicate Detection is Preferred

It seems reasonable to assume that when real world medical problems are trending – and those problems are discussed on Twitter with a somewhat unique vocabulary – then we might expect word sets containing one or more of those unique terms to also trend. Notice, we use the plural "word sets" because we do expect multiple sets to trend. This expectation is driven strongly by combinatorics. For instance, if 5 words are highly likely to be used when a person is writing about the flu on twitter, we can expect several, if not most, combinations of these 5 words to trend at the same time. We can also expect many of these word combinations to trend when paired with additional words e.g., {flu, hate}.

This observation is helpful for two reasons. First, it enables us to be better prepared to deal with medical conditions that get flagged as "trending" by only a small number of word sets in any given month. The second reason duplicate detection is helpful is that it enables better accuracy just like ensemble method in data mining or increasing the sample size in statistics.

E. Results Discussion

The results shown above are promising. Taken together they form a good proof of principle. The framework detects the well-known seasonal medical ailments of influenza and springtime allergies without any ailment specific customization. These results were obtained while a minimum support of 0.01% word set prevalence was required (discussed in section 4.2). We do not think the min support must be this high for the trend detection methodology to work. In other words, we do not believe this methodology is only good for detecting common conditions. In fact, we believe reducing the minimum support and searching for seasonal sports related injuries would be a useful exercise. It would be promising if concussions were flagged as a trending health condition when the high school football season started because concussions should happen rarely in the general population. Thus, if they are detectable then we would have good reason to believe that other somewhat rare health conditions could also be detected.

Our framework does produce some false positives. For example, "Food Allergy" is flagged as a trending condition in March and April of 2010 because the "Food Allergy" article contains many of the words people use to discuss pollen allergies on Twitter. For now, we choose to err on the side of having better recall even at the expense of precision. After all, any positive result will need to be vetted and verified by a health professional before any significant action can be taken.

Recall, this work was performed using a corpus of health-related tweets that was culled from a larger corpus using three filters. It is unclear if using the filtered dataset would generate better results. Due to the absence of a strong intuition about which corpus would be best we opted to use the more manageable corpus of 1.6 million tweets over the .5 TB corpus of 2 billion tweets.

VI. CONCLUSION AND FUTURE WORK

We demonstrated a single framework for detecting a multitude of public health trends which clearly detected the seasonal afflictions of influenza, allergies, and summertime ice-cream headache. The framework is simple to implement and operates efficiently because it is built on top of the already mature resources Lucene, Mahout, and Wikipedia.

We detect public health trends because we use the filtered corpus and Wikipedia/ICD codes to filter results. We could conceivably detect other types of trends by changing the filters to suit the new topic of interest.

We have two main future development goals: (1) We would like to run the framework on a larger scale to comfortably enable increasing the temporal resolution from months to weeks and possibly even days and (2) We would like to investigate using a resource besides Wikipedia and ICD to filter out non-medically related trending topics. Using Wikipedia and ICD makes detecting previously known (and possibly common) ailments easy; but it may also prevent the detection of novel ailments.

ACKNOWLEDGMENT

This work was partially supported by: The Models of Infectious Disease Agent Study (MIDAS), under Award Number U01GM070708 from the NIGMS, The Johns Hopkins Medical School DHS Center on Preparedness and Catastrophic Event Response (PACER), under Award Number N00014-06-1-0991 from the Office of Naval Research, and Joshua M. Epstein's NIH Director's Pioneer Award, Number DP1OD003874 from the Office of the Director, National Institutes of Health.

REFERENCES

[1] Twenty six percent of online adults discuss health information online; privacy cited as the biggest barrier to entry.http://www.businesswire.com/news/home/20121120005872/en

[2] Twitter blogs: measuring tweets. http://blog. twitter.com/2010/02/measuring-tweets.html.

[3] Twitter statistics. http://www.statisticbrain.com/twitter-statistics/

[4] http://www.google.org/flutrends/us/#US

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of 20th International Conference on Very Large Data Bases, VLDB, pages 487–499, 1994.

[6] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, pages 1568–1576, 2011.

[7] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[8] S.T. Brown, J.H. Tai, R.R. Bailey, P.C. Cooley, W.D. Wheaton, M.A. Potter, R.E. Voorhees, M. Lejeune, J.J. Grefenstette, D.S. Burke, S.M. McGlone, B.Y. Lee. Would school closure for the 2009 H1N1 influenza epidemic have been worth the cost? : a computational simulation of Pennsylvania. BMC Public Health. 2011, May 20;11(1):353.

[9] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, NIPS, pages 288–296, 2009.

[10] W. Chou, Y. Hunt, E. Beckjord, R. Moser, and B. Hesse. Social media use in the United States: implications for health communication. Journal of medical Internet research, 11(4), 2009.

[11] A. Cohen. Optimizing feature representation for automated systematic review work prioritization. In Proceedings of AMIA Annual Symposium, volume 2008, page 121, 2008.

[12] A. Cohen, K. Ambert, and M. McDonagh. Cross-topic learning for work prioritization in systematic review creation and update. Journal of the American Medical Informatics Association, 16(5):690–704, 2009.

[13] C. Corley, A. Mikler, K. Singh, and D. Cook. Monitoring influenza trends through mining social media. In Proceedings of the International Conference on Bioinformatics Computational Biology, ICBCB, pages 340–346, 2009.

[14] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In Proceedings of the 1st Workshop on Social Media Analytics, pages 115–122, 2010.

[15] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Towards personalized learning to rank for epidemic intelligence based on social media streams. In Proceedings of the 21st international conference companion on World Wide Web, WWW, pages 495–496, 2012.

[16] J.M. Epstein, D.M. Goedecke, F. Yu, R.J. Morris, D.K. Wagener, et al. (2007) Controlling Pandemic Flu: The Value of International Air Travel Restrictions. PLoS ONE 2(5): e401. doi:10.1371/journal.pone.0000401

[17] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. Nature, 457(7232):1012–1014, 2008.

[18] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S. Lawson. "i can't get no sleep": discussing #insomnia on twitter. In Proceedings of the ACM annual conference on Human Factors in Computing Systems, CHI, pages 1501-1510, 2012.

[19] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11): 2169-2188, 2009

[20] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. ACM Transactions on Intelligent Systems and Technology, 3(4):72, 2012.

[21] H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Chang. PFP: Parallel FP-growth for query recommendation. In Proceedings of the ACM Conference on Recommender Systems, pages 107-114, 2008

[22] E. Mykhalovskiy, L. Weir, et al. The global public health intelligence network and early warning outbreak detection: a Canadian contribution to global public health. Canadian journal of public health, 97(1):42, 2006.

[23] A. Nakhasi, R. Passarella, S. Bell, M. Paul, M. Dredze, and P. Pronovost. Malpractice and malcontent: Analyzing medical complaints in twitter. In AAAI Fall Symposium Series, 2012.

[24] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the 4th International Conference on Weblogs and Social Media, ICWSM, 2010.

[25] J. Parker and J.M. Epstein. A Distributed Platform for Global-Scale Agent-Based Models of Disease Transmission. ACM Trans. Model. Comput. Simul. 22, 1, Article 2 (December 2011), 25 pages.

[26] M. Paul and M. Dredze. A model for mining public health topics from twitter. HEALTH, 11:16–6, 2012.

[27] M. J. Paul and R. Girju. A two-dimensional topic-spect model for discovering multi-faceted topics. In Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010.

[28] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, WWW, pages 851–860, 2010.

[29] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR, pages 841–842, 2010.

[30] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the 4th International Conference on Weblogs and Social Media, ICWSM, 2010.

[31] I. Uysal and W. B. Croft. User oriented tweet ranking: a filtering approach to microblogs. In Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM, pages 2261–2264, 2011.

[32] B. Wenerstrom, M. Kantardzic, E. Arabmakki, and M. Hindi. Multi-tweet summarization for flu outbreak detection. In AAAI Fall Symposium Series, 2012

[33] R.W. White RW, N.P. Tatonetti, N.H. Shah, R.B. Altman, E. Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc. 2013 May 1;20(3):404-8. doi: 10.1136/amiajnl-2012-001482. Epub 2013 Mar 6.

[34] A. Yates and N. Goharian, "ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites", In Proceedings of the 35th European Conference on Information Retrieval (ECIR 2013), 2013

[35] Y. Zhu and N. Goharian, "To Follow or Not to Follow: A Feature Evaluation", Proceedings of the 22nd international conference on World Wide Web (WWW'13), 2013.