

A Framework for Public Health Surveillance

Andrew Yates¹, Jon Parker^{1,2}, Nazli Goharian¹, Ophir Frieder¹

¹Information Retrieval Lab, Department of Computer Science, Georgetown University

²Department of Emergency Medicine, Johns Hopkins University

{andrew,jon,nazli,ophir}@ir.cs.georgetown.edu

Abstract

With the rapid growth of social media, there is increasing potential to augment traditional public health surveillance methods with data from social media. We describe a framework for performing public health surveillance on Twitter data. Our framework, which is publicly available, consists of three components that work together to detect health-related trends in social media: a concept extraction component for identifying health-related concepts, a concept aggregation component for identifying how the extracted health-related concepts relate to each other, and a trend detection component for determining when the aggregated health-related concepts are trending. We describe the architecture of the framework and several components that have been implemented in the framework, identify other components that could be used with the framework, and evaluate our framework on approximately 1.5 years of tweets. While it is difficult to determine how accurately a Twitter trend reflects a trend in the real world, we discuss the differences in trends detected by several different methods and compare flu trends detected by our framework to data from Google Flu Trends.

Keywords: LR Infrastructures and Architectures, Social Media, Trend Detection

1. Introduction

With over 400 million tweets posted per day¹, Twitter contains a wealth of information that can be used to learn about public health. While some have explored using language recognition methods to identify the spread of influenza using Twitter (Aramaki et al., 2011; Culotta, 2010; Lamos and Cristianini, 2012), few have studied the task of performing general public health surveillance using Twitter.

We describe an open source framework² for performing public health surveillance on social media, such as Twitter, that expands and generalizes Parker et al. (2013)'s effort. Our framework generalizes this previous effort by defining interchangeable language recognition components that operate independently of each other (i.e., components to perform concept extraction, concept aggregation, and trend detection), allowing for experimentation with and evaluation of individual components. Our framework improves upon other efforts to monitor the spread of influenza on Twitter by not requiring a public health topic to be specified a priori. That is, our framework is designed to detect any public health topic present on Twitter rather than being designed to detect instances of a specified health topic (e.g., the seasonal flu). While we focus on detecting topics related to public health, our framework's architecture is not specific to health and could be used to detect other types of trends (e.g., increased mentions of a celebrity or geographic location). Furthermore, our framework can be used to detect both trends within a geographic area (when supplied with geotagged tweets) and general trends (i.e., trends that are not limited to a geographic area).

2. Related Work

Most previous work on using social media for public health surveillance has focused on correlating Influenza-

like-illness (ILI) rates obtained using social media with ILI rates from a traditional source. This differs from our goal of detecting health-related trends without specifying a condition of interest (e.g., ILI) a priori. Culotta (2010) experiment with different feature selection methods and regression models to predict ILI rates on Twitter. Aramaki et al. (2011) classify tweets as flu related or unrelated using a SVM. Similarly, Lamos and Cristianini (2012) investigate using LASSO to predict ILI rates and the amount of rainfall in an area. Corley et al. (2009) find that blog trends can also indicate ILI rates; they correlate ILI-related trends with ILI incidence data from the US Center for Disease Control and Prevention (CDC).

Parker et al. (2013) propose a framework for identifying health-related trends using frequent term sets that does not require topics to be specified a priori. It differs from this work in the scope of the proposed framework; it focuses on detecting trends using frequent term sets generated by an association rule mining algorithm (Li et al., 2008) and using Wikipedia to map term sets to health concepts, allowing the specifics of how frequent term sets are chosen and how term sets are associated with Wikipedia articles to vary. Rather than focusing on a particular trend detection strategy, this work generalizes Parker et al. (2013) by defining three types of components from which a trend detection system can be built. In this framework, frequent term set generation is an example of a *concept extraction* method, and mapping term sets to Wikipedia articles is an example of a *concept aggregation* method. We include these methods, along with others, as components of our framework.

3. Framework

Our framework is designed to facilitate experimentation with different methods for detecting health-related trends on Twitter. To that end, the framework defines interchangeable components that work together to achieve this goal. Social media documents and (optionally) a health dictionary or thesaurus are the inputs to the framework; the

¹<http://www.theguardian.com/technology/2013/sep/12/twitter-ipo-stock-market-launch>

²Our reference implementation is available at <http://ir.cs.georgetown.edu/data/lrec14>

framework outputs health-related trends supported by the input documents. The framework consists of three primary components:

- A concept extraction method responsible for identifying any health-related concepts (e.g., “sneezing”) expressed in a social media document (e.g., a tweet). Concept extraction methods may use a health dictionary to determine what concepts should be detected.
- A concept aggregation method responsible for inferring the possible presence of trend-level concepts based on the concepts identified by the concept extraction method (e.g., if the “sneezing” and “itchy eyes” concepts are both present in a document, they might be combined to produce the “allergies” concept). Concept aggregation methods may use a health thesaurus to determine when two concepts are equivalent (e.g., “alopecia” and “hair loss”).
- A trend detection method responsible to determine when, if ever, the trend-level concepts returned by the concept aggregation method are trending.

The framework’s architecture is illustrated in Figure 1. The concept extraction method (a) identifies health-related concepts (e.g., “itchy eyes” or “allergies”) expressed in the tweets provided as input (b). Concepts are then associated with potential trend-level concepts (e.g., the “itchy eyes” concept may be a symptom of the “allergy” condition) by the concept aggregation method (c). The concept aggregation method may choose to return some of its input concepts unaltered when they cannot or should not be mapped to a trend-level concept (e.g., if there is no trend-level concept to map a concept to, as with “influenza”). These trend-level concepts are then analyzed by the trend detection component (d) to determine if any trend-level concept is occurring more often than usual within a time period (e.g., if there is a high incidence of seasonal allergies). The trend detection component outputs trending health topics and time periods associated with them (e).

Each component relies only on the previous component’s output and method-specific inputs (e.g., a thesaurus or Wikipedia), allowing the components to easily be interchanged; each component’s standard inputs and outputs are shown in Table 1. In the following sections, we describe each component in greater detail and identify methods that can potentially be used to perform each component’s task. Subject to any applicable licensing restrictions, components will be continuously added, updated, and made available on our Website.

3.1. Concept Extraction

3.1.1. Overview

The concept extraction method is responsible for identifying the health-related concepts expressed in each tweet. An example tweet and its concepts are shown in Figure 2.

Concept extraction differs from traditional named entity recognition (NER) in that concepts do not need to consist of sequential terms as named entities traditionally do. Furthermore, we are concerned with identifying when concepts are

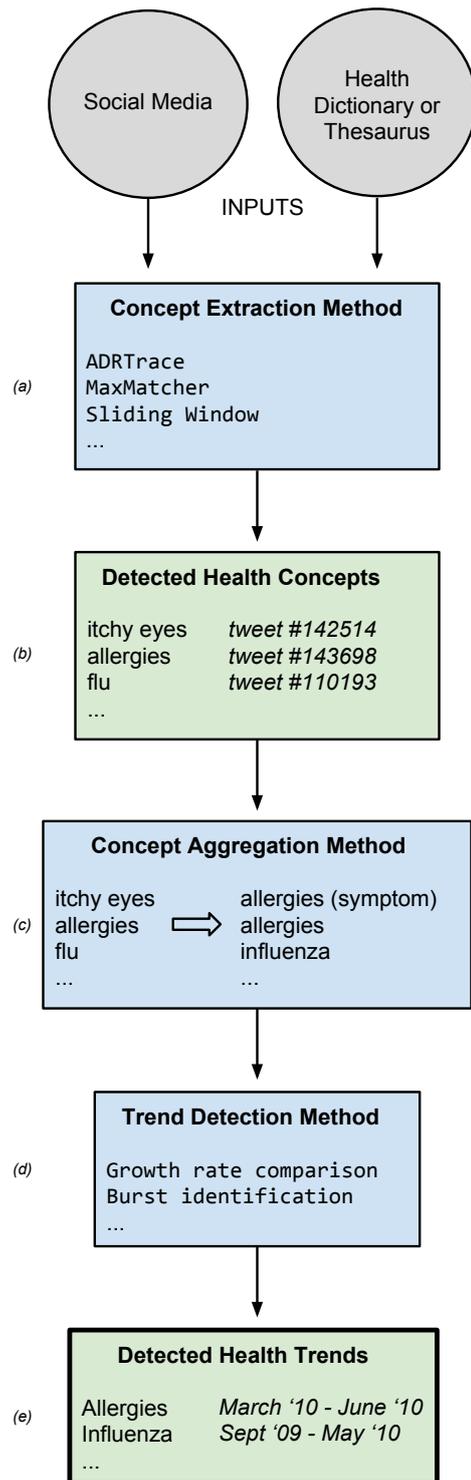


Figure 1: System Architecture

expressed rather than with categorizing named entities in text (e.g., as a person or place). Figure 2 shows an example tweet where the terms “nose” and “and” separate “itchy” and “throat” in the “itchy throat” concept; the concept extraction method’s task is to extract this “itchy nose” and “itchy throat” concepts from this tweet, not to categorize “itchy nose and throat” as a symptom.

The concept extraction method may optionally use a

<i>Component</i>	<i>Input(s)</i>	<i>Output</i>
Concept Extraction	Corpus file specifying document and metadata files (Optional) Dictionary or thesaurus	A (concept, document id) pair for each concept detected
Concept Aggregation	Concept pairs (from concept extraction) (Optional) Thesaurus	A (concept, document id) pair for each trend-level concept created from the input concepts
Trend Detection	Corpus file specifying document and metadata files Trend-level concept pairs (from concept aggregation)	The time periods each trend-level concept is trending and the strength of each trend

Table 1: Each component’s standard inputs and outputs, as of version 2 of our framework’s API. A full description of the latest version of our framework’s API is available on our Website. Each input and output is a file in JSON format. Components may take additional method-specific inputs as needed.

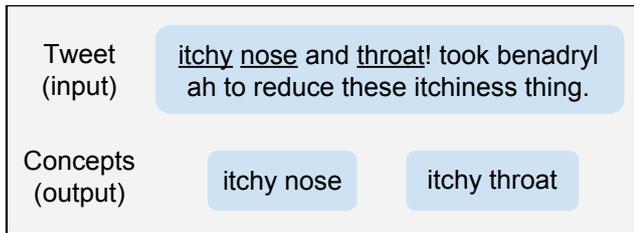


Figure 2: Concept extraction example

domain-specific dictionary or thesaurus to aide in the identification of relevant terms. In Figure 2, for example, a concept extraction method might identify the “itchy throat” concept by considering the dependency relations between health-related terms appearing in a dictionary (i.e., “itchy,” “nose,” and “throat”).

3.1.2. Methods

Many concept extraction methods have been proposed that could be used with our framework. We focus on those methods designed to perform concept extraction in a health-related domain. ADRTrace (Yates and Goharian, 2013) extracts text matching lexical patterns (e.g., “pain in my leg” matches ‘<X>in my <Y>’) and associates them with concepts in a health thesaurus (e.g., “leg pain). Parker et al. (2013) describe a system for detecting health-related Twitter trends that identifies frequent term sets that correspond with concepts. Though it is not labeled as such, the system’s frequent term generation component performs concept extraction. The system avoids the need for a thesaurus by discarding tweets that are unrelated to health. The concept extraction component from this system is available in our framework as `component.parker.concept_extraction`. As in Parker et al. (2013), we require term sets to have a minimum support of 0.001, we generate frequent term sets over one month’s worth of tweets at a time, and we discard those term sets from each month that did not occur more frequently than in the previous month.

Leaman et al. (2010) use a bag-of-words sliding window to identify concepts present in a dictionary. Rather than matching terms against a dictionary directly, Leaman et al. (2010) compute the Jaro–Winkler

(Winkler, 1990) distance between terms in the window and in the dictionary. A bag-of-words sliding window method is available in our framework as `component.thesaurus.sliding_extraction`; this method differs from Leaman et al. (2010) in that it performs exact matching between terms in the window and terms in the dictionary.

MaxMatcher (Zhou et al., 2006) identifies match candidates and then performs weighted partial matching between the match candidates and a dictionary to identify expressed concepts. Match candidates are identified using rules designed for medical literature (e.g., a match candidate must “begin with a noun, number, or an adjective”) that may not be appropriate for social media.

Traditional NER methods such as the Stanford Named Entity Recognizer (Finkel et al., 2005) could likewise be used as concept aggregation methods, though the types of entities these methods detect are different from the concepts that are of interest for performing public health surveillance. NER could be used to detect other types of trends, however, such as a trending celebrity (i.e., an entity with the “PERSON” label).

3.2. Concept Aggregation

3.2.1. Overview

After a document’s concepts have been extracted, the concept aggregation method attempts to map each concept with one or more trend-level concepts that it may be related to. The trend detection component’s task is to identify trends in trend-level concepts, so the concept aggregation method should output concepts at the level of granularity desired for trends. The term trend-level concept is intentionally vague, as the type of concept being mapped to is specific to the use case. For public health surveillance, the concept aggregation method might map symptoms (e.g., “itchy eyes” and “cough”) to one or more health conditions associated with the symptoms (“allergies”).

The concept aggregation method’s output may differentiate between mentions of a health condition and mentions of a concept that may be a symptom of a health condition. This allows the trend detection method to weight mentions of a health condition (e.g., “allergies”) more highly than mentions of a symptom (e.g., “itchy eyes,” which may be a symptom of “allergies” and, thus, could be labeled “allergies symptom” by the concept aggregation method).

In some domains the concept aggregation component might simply map synonymous concepts to a canonical concept (e.g., mapping different forms of a celebrity’s name to the same term).

3.2.2. Methods

Many concept aggregation methods have been proposed in various domains. As part of their system for detecting health-related trends, Parker et al. (2013) use information retrieval methods to match concepts (e.g., “itchy eyes” and “sneezing”) against health-related Wikipedia articles describing the condition they may be symptoms of (e.g., “allergies”). This method is available in our framework as `component.parker.concept_aggregation`.

Concept aggregation can be viewed as a disambiguation task in which concepts are disambiguated by being associated with Wikipedia pages (i.e., “Wikification”). Many Wikification methods have been proposed. Han and Zhao (2009) considers a semantic network derived from Wikipedia when calculating similarity. Ratinov et al. (2011) propose a system that considers both local and global disambiguation factors when performing Wikification. Cheng and Roth (2013) improve Wikification’s performance by considering the relations between concepts. Finally, Spitzkovsky and Chang (2012) use the conditional probabilities of terms in Wikipedia pages’ anchor text to associate text with Wikipedia articles.

Alternatively, a domain-specific ontology such as the Unified Medical Language System’s semantic network (Bodenreider, 2004) may be used in place of Wikipedia-based methods. A thesaurus look-up method is available in our framework as `component.thesaurus.concept_aggregation`; the method takes a thesaurus as input and maps each concept to one or more matching entries in the thesaurus.

3.3. Trend Detection

3.3.1. Overview

After each tweet is associated with zero or more trend-level concepts, the trend detection method is responsible for identifying health topics that are occurring more often than usual (i.e., trending) within a time period. For example, the system shown in Figure 1 detects an increased incidence of allergies between March and June 2010, and an increased incidence of the flu between September 2009 and May 2010. These dates correspond with seasonal allergies and flu season, respectively. The trend detection method may take a unit of time to consider as a parameter (i.e., daily or weekly trends) or the unit of time may be determined by the method.

The trend detection method outputs a list of time periods each topic is trending and the trend strength associated with each time period. The trend detection method is responsible for choosing granularity of the time periods. Similarly, the trends’ level of granularity is specific to the method used; continuous strengths are the most informative, but a method could output a boolean strength if desired (i.e., mark each time period as “trending” or “not trending”).

3.3.2. Methods

Many trend detection methods exist. (Parker et al., 2013) determine whether a concept is trending by comparing the concept’s frequency between two time periods. This method is available in our framework as `component.parker.trend_detection`. Fisichella et al. (2010) propose a formula for identifying “bursty periods” (i.e., periods during which a concept is trending). Ihler et al. (2006) use a Poisson process model to detect events; we provide a similar method as `component.trend_detection.poisson_process`. We provide three methods for detecting trends at the daily, weekly, or monthly level in `component.trend_detection.simple_{daily,weekly,monthly}`, which always mark a concept as trending and combine concept occurrences over a time period.

4. Results

To illustrate the utility and flexibility of our framework, we compare health trends detected using a variety of different components on a Twitter corpus (described in section 4.2.). The methods we use in our experiments and their parameters are shown in Table 2. The health-related thesaurus that we use with the components that require a dictionary or thesaurus (i.e., `thesaurus.sliding_extraction` and `thesaurus.concept_aggregation`) is described in section 4.1.. It is often difficult to determine how accurate a detected trend is because of the lack of detailed data about the trending concept’s actual incidence. For this reason we focus on two trends that are easier to validate: seasonal allergies and the flu. We investigate the differences in trends detected for these concepts when the concept extraction, concept aggregation, and trend detection components are varied.

4.1. Thesaurus

We use the MedSyn thesaurus (Yates and Goharian, 2013) with the components that require a dictionary or thesaurus. MedSyn is derived from UMLS, the Unified Medical Language System (Bodenreider, 2004); it eliminates many concepts in UMLS that are only tangentially health-related. MedSyn consists of both lay terms (e.g., “joint pain”) and expert medical terms (e.g., “arthralgia”) related to medical symptoms, common conditions, and adverse drug reactions (e.g., “sore throat,” “flu,” “hair loss”). While we cannot distribute MedSyn directly due to the terms of our UMLS license, instructions on re-creating it from a copy of UMLS are available in Yates and Goharian (2013); UMLS licenses may be obtained free of charge from the US National Library of Medicine in many countries. Additionally, domain-specific thesaurus construction and synonym discovery techniques (Alfonseca et al., 2005; McCrae and Collier, 2008; Pantel et al., 2009; Yates et al., 2014) could be used to extend MedSyn or to create a new thesaurus specialized for a particular domain.

4.2. Corpus

Our Twitter corpus consists of health-related tweets extracted from tweets collected from the public Twitter API

feed between June 2009 and November 2010. We use only health-related tweets because one of our concept extraction methods, Parker et al. (2013), does not perform any filtering; it requires health-related tweets as input. The `thesaurus.sliding` extraction concept extraction method does not require its input to be health-related because it uses a dictionary to identify health-related concepts. The tweets in our corpus were filtered using the filtering methodology and SVM described in Paul and Dredze (2010), which reduced the corpus from 2 billion tweets to 1.6 million health-related tweets. While Twitter’s terms of service prevent us from directly distributing our corpus, the data needed to retrieve the tweets in our corpus and reproduce our results are available on our Website³, including:

- The list of all tweet ids in the 2 billion tweet collection. This can be used to retrieve the tweets using Twitter’s API, though Twitter’s rate limits admittedly make this a time consuming process.
- The list of all tweet ids contained in the 1.6 million health-related tweet subset.
- The output of our concept aggregation methods, which can be used as input for a trend detection method to detect trends in the health-related tweets. This allows other researchers to experiment with different trend detection methods without acquiring any tweets or a thesaurus.

4.3. Allergies

We first investigate trends related to the “Allergy” concept, which should be less prevalent during the winter when less pollen is in the air. The trends detected by two method configurations on a monthly basis are shown in Figure 3. The y-axis shows the strength of the detected trend normalized by the maximum strength returned by each configuration. The concept extraction and concept aggregation method configurations that are not shown (i.e., *Term Sets + Wiki* and *Term Sets + Thesaurus*) behaved similarly to *Sliding Window + Thesaurus*. All four configurations show a decline in winter 2009 before peaking in May 2010; *Sliding Window + Wiki* declines less in the winter, however, and stays at roughly the same level for much of the time between September 2009 and March 2010. Furthermore, all four configurations detect a similar increase in October 2010, but show different declines in October 2009. The trends detected on a weekly basis are shown in Figure 4. While both method configurations exhibited similar behavior over summer 2010 when viewed at the month level, they differ in June 2010 when viewed at the week level. It is impossible to determine which configuration is more accurate without more detailed information on the actual incidence of seasonal allergies in 2009-2010, but this result illustrates our framework’s utility as a tool for hypothesis generation.

4.4. Flu

The seasonal flu is a useful benchmark for public health surveillance methods because flu incidence data are available from several sources, such as from Google Flu Trends

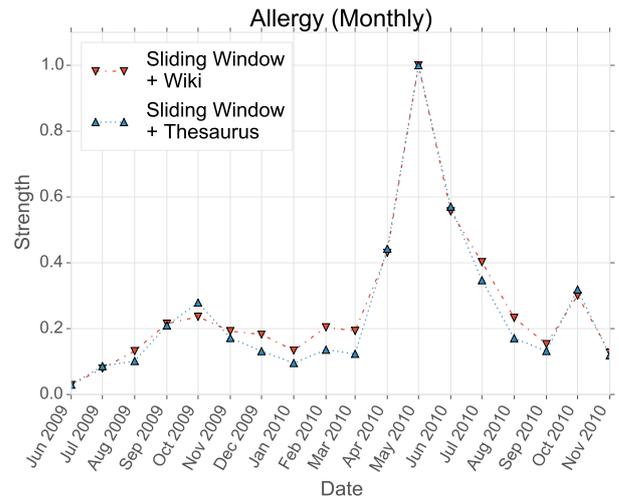


Figure 3: Trends related to the “Allergy” concept identified using monthly trend detection. Both method configurations exhibit similar behavior in summer 2010, but show different declines in winter 2009.

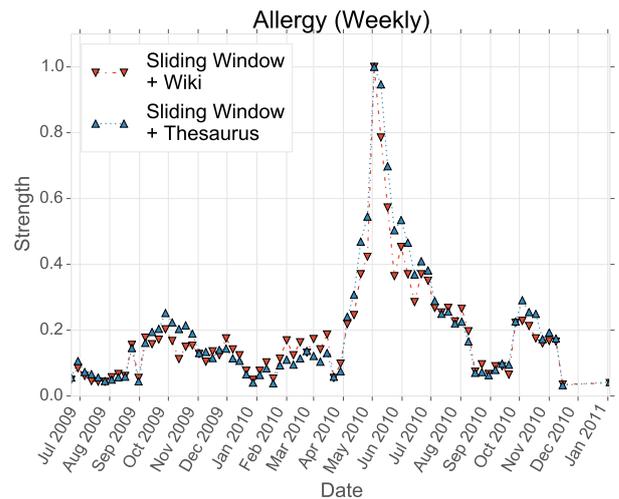


Figure 4: Trends related to the “Allergy” concept identified using weekly trend detection. While both configurations exhibited similar monthly trends over summer 2010, the weekly trends differ.

(Ginsberg et al., 2009) and the US Center for Disease Control and Prevention (CDC). Existing public health surveillance systems such as HealthMap (Brownstein et al., 2008), a system that monitors reports of disease outbreaks, also often have flu incidence data. To investigate how closely the different method configurations in Table 2 match Google Flu Trends, we detect trends related to the “Influenza” concept and compare them to Google Flu Trends. The trends detected by our method configurations are shown in Figure 5; the trends detected by Google Flu Trends are shown in Figure 6. The *Sliding Window + Wiki* configuration differs substantially from both the three other configurations and Google Flu Trends, indicating that this method config-

³<http://ir.cs.georgetown.edu/data/lrec14/>

Component	Method	Abbreviation	Parameters
Concept Extraction	parker.concept_extraction	Term Sets	Minimum growth rate: 1.5 Minimum support: 0.001
Concept Extraction	thesaurus.sliding_extraction	Sliding Window	Window size: 5 terms Dictionary: terms in MedSyn
Concept Aggregation	parker.concept_aggregation	Wiki	Max. trend-level concepts mapped to each concept: 10 Max. WP pages returned: 50
Concept Aggregation	thesaurus.concept_aggregation	Thesaurus	Matching algorithm: <i>Attempt to match entire concept term against thesaurus. If nothing matches, match each term within concept.</i> Thesaurus: MedSyn
Trend Detection	trend_detection.simple_weekly	Weekly	-
Trend Detection	trend_detection.simple_monthly	Monthly	-

Table 2: Components and methods

uration is a poor indicator of the seasonal flu. Both Google Flu Trends and the three similar configurations peak around October 2009, which suggests these three configurations are correctly identifying seasonal flu trends. *Term Sets + Wiki* and *Term Sets + Thesaurus* behave almost identically, illustrating that the *Wiki* and *Thesaurus* concept aggregation methods are almost equivalent in this case. These two configurations differ slightly from *Sliding Window + Thesaurus*, but exhibit the same general trends.

While three of our configurations detected similar trends to Google Flu Trends, our configurations appear to detect a spike in November 2009, whereas Google Flu Trends' spike appears to be in late October 2009. We identified flu trends at a weekly level to investigate this discrepancy (shown in Figure 7). When our configurations' trends are viewed at the weekly level, the three similar configurations spike in late November 2009; this differs from the trend identified by Google Flu Trends, which begins to decline in early November 2009.

5. Conclusions

We have described an open source framework for detecting public health topics on social media, such as Twitter, that generalizes previous health trend detection efforts. Our framework consists of three independent components that perform concept extraction, concept aggregation, and trend detection. Our results in section 3 illustrate that, while a component can perform its task using many different methods, it is often difficult to compare a method's performance against real-world health trends. Furthermore, each different method may generate different hypotheses, and the methods used by the framework can be easily changed. Given this and the difficulty of obtaining real world trends, we envision our framework to be used primarily as a hypothesis generation tool.

6. Acknowledgments

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

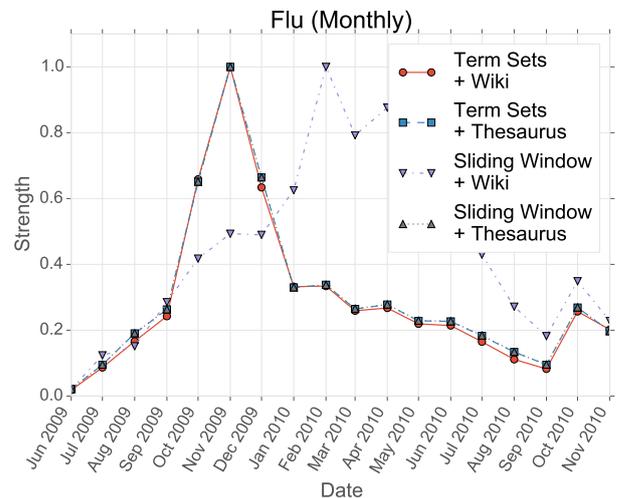


Figure 5: Trends related to the “Influenza” concept identified using monthly trend detection. Three of the four configurations exhibit similar behavior, but all vary slightly.

7. References

- Alfonseca, E., Castells, P., and Ruiz, M. (2005). Using context-window overlapping in synonym discovery and ontology extension. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP '05)*, number Lyons 61.
- Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 1568–1576, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–70, January.

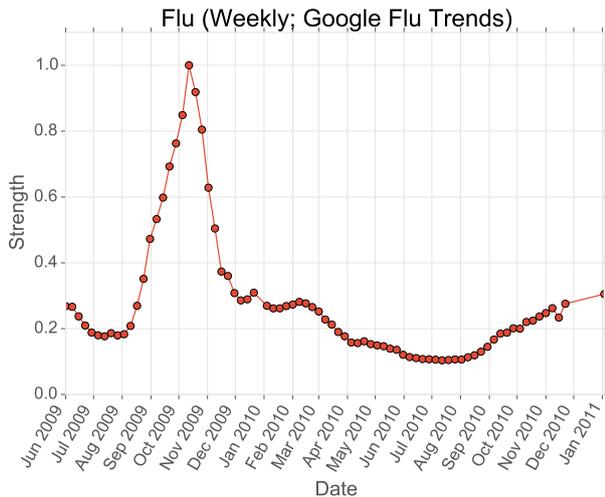


Figure 6: Estimated weekly flu cases for the United States in June 2009 – November 2010. *Data Source: Google Flu Trends* (<http://www.google.org/flutrends>)

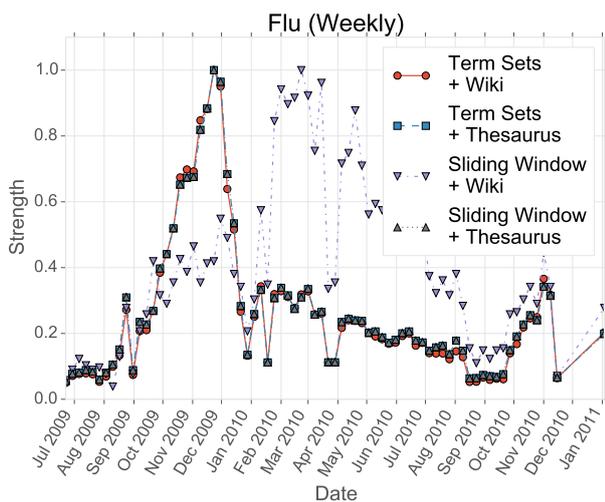


Figure 7: Trends related to the “Influenza” concept identified using weekly trend detection. Three of the method configurations spike in late November 2009, whereas Google Flu Trends spikes in late October 2009.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Med*, 5(7):e151, 07.

Cheng, X. and Roth, D. (2013). Relational Inference for Wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (RANLP '13)*.

Corley, C., Mikler, A., Singh, K., and Cook, D. (2009). Monitoring Influenza Trends through Mining Social Media. In *Proceedings of the International Conference on Bioinformatics Computational Biology (ICBCB '09)*.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the*

First Workshop on Social Media Analytics (SOMA '10), pages 115–122.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*, number 1995.

Fisichella, M., Stewart, A., Denecke, K., and Nejd, W. (2010). Unsupervised public health event detection for epidemic intelligence. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 1881, October.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014. doi:10.1038/nature07634.

Han, X. and Zhao, J. (2009). Named entity disambiguation by leveraging wikipedia semantic knowledge. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09*, pages 215–224, New York, NY, USA. ACM.

Ihler, A., Hutchins, J., and Smyth, P. (2006). Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, pages 207–216, New York, NY, USA. ACM.

Lamos, V. and Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, 3(4):72:1–72:22, September.

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., and Gonzalez, G. (2010). Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP'10)*, number July, pages 117–125.

Li, H., Wang, Y., Zhang, D., Zhang, M., and Chang, E. Y. (2008). Pfp: Parallel fp-growth for query recommendation. In *Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys '08*, pages 107–114, New York, NY, USA. ACM.

McCrae, J. and Collier, N. (2008). Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC bioinformatics*, 9, January.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-m., and Vyas, V. (2009). Web-Scale Distributional Similarity and Entity Set Expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2 (EMNLP '09)*, number August.

Parker, J., Wei, Y., Yates, A., Frieder, O., and Goharian, N. (2013). A Framework for Detecting Public Health Trends with Twitter. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM'13)*.

Paul, M. J. and Dredze, M. (2010). A Model for Mining Public Health Topics from Twitter. Technical Report May 2009.

Ratinov, L., Roth, D., Downey, D., and Anderson, M.

- (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11*, pages 1375–1384, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Spitkovsky, V. I. and Chang, A. X. (2012). A Cross-Lingual Dictionary for English Wikipedia Concepts. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.
- Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.
- Yates, A. and Goharian, N. (2013). ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In *Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13)*.
- Yates, A., Goharian, N., and Frieder, O. (2014). Relevance-Ranked Domain-Specific Synonym Discovery. In *Proceedings of the 36th European Conference on Information Retrieval (ECIR '14)*.
- Zhou, X., Zhang, X., and Hu, X. (2006). MaxMatcher: Biological concept extraction using approximate dictionary lookup. *PRICAI 2006: Trends in Artificial Intelligence*, 4099:1145–1149.