Mining Social Media Streams to Improve Public Health Allergy Surveillance

Kathy Lee, Ankit Agrawal, Alok Choudhary
EECS Department
Northwestern University
Evanston, IL USA
Email: {kml649, ankitag, choudhar}@eecs.northwestern.edu

***Abstract*—Allergies are one of the most common chronic diseases worldwide. One in five Americans suffer from either allergy or asthma symptoms. With the prevalence of social media, people sharing experiences and opinions on personal health symptoms and concerns on social media are increasing. Mining those publicly available health related data potentially provides valuable healthcare insights. In this paper, we propose a real-time allergy surveillance system that first classifies tweets to identify those that mention actual allergy incidents using bag-of-words model and NaiveBayesMultinomial classifier and applies in-depth text and spatiotemporal analysis. Our experimental results show that the proposed system can detect predominant allergy types with high precision and that allergy-related tweet volume is highly correlated to the weather data (daily maximum temperature). We believe that this is the first study that examines a large-scale social media stream for in-depth analysis of allergy activities.**

***Keywords*-Social Media; Twitter; Allergy; Public Health;**

I. Introduction

Allergy is the fifth most common chronic diseases in the United States[1]. The complexity and severity of allergic diseases are increasing worldwide [24]. One in five Americans have either allergy or asthma symptoms. In 2012, 7.5% of adults (17.6 million adults) and 9% of children (6.6 million children) were diagnosed with hay fever [5, 4]. Continuous use of allergy medication can worsen patients' health conditions and lead to side effects and other serious medical complications. Furthermore, increasing number of allergy patients gives rise to allergy-related health care cost and leads to reduced work productivity. $7.9 million is annually spent on allergy-related health care systems and business. 4 million workdays are lost due to hay fever each year. Therefore, accurate allergy surveillance and forecast is important to minimize the healthcare cost and maximize work productivity lost due to allergy symptoms.

Twitter[2], one of the largest social networking website, allows users to post short text messages called tweets that can be up to 140 characters in length. Twitter has over 645 million active registered users. Twitter has been used as a valuable real-time information resource for various applications. For instance, twitter data have been used to detect earthquakes in Japan [25], predict the stock market [6] and for an in-depth study of 2011 Egyptian Revolution [10].

On twitter, people not only make general chatters but also share photos, news, opinions, emotions, and even health conditions including symptoms and medications they are taking for their diseases. In recent years, many researchers have investigated using twitter for disease surveillance, especially for influenza epidemic detection and prediction [23, 8, 1, 28, 3, 7, 16, 26, 19].

In this paper, we mine a large scale twitter data collected over 28 months to monitor allergy levels. More specifically,

1) A bag-of-words supervised learning approach is employed to distinguish tweets that mention actual incidents of allergy from those that talk about news or general awareness about allergy.
2) Text-mining techniques such as n-gram extraction and part-of speech tagging are applied to extract predominant allergy types.
3) A spatiotemporal mining is applied to track allergy levels over time and space.

We believe that our work is the first framework towards real-time allergy surveillance using fine-grained spatiotemporal analysis on a large-scale social media data. The data analysis results reveal that Twitter is a good source of detecting allergy prevalence. Our proposed system can help see the past and current trend of allergy levels detected in social media stream. The real-time analysis results are updated on our allergy project website[3].

II. Our Approach

A. Datasets

1) Twitter dataset: We collected allergy-related tweets from public tweet stream using twitter's streaming API[4]. We have collected over 6.3 million tweets that mention 'allergy' or 'allergies' created by over 3.1 million unique users over 28 months from January 2013 to April 2015. Some talk about their allergy symptoms (e.g., *Walked out of my house*

[1]http://www.webmd.com/allergies/allergy-statistics
[2]https://twitter.com
[3]http://pulse.eecs.northwestern.edu/~kml649/allergy/
[4]https://dev.twitter.com/docs/streaming-apis

ASONAM '15, August 25-28, 2015, Paris, France
 ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2808896

confused as to why my eyes felt like they were on fire and then I realized it's allergy season.) while others talk about allergy types (e.g., *I sneezed like eight times in a row. This pollen allergy is killing me.*) or allergy treatments/medication they take (e.g., *sitting in doctor's office just to get an allergy shot.*).

2) Ground Truth Data:

Pollen dataset: We collected monthly average pollen levels and 90 day historic pollen levels for US major cities from pollen.com[5]. The pollen level is a number between 0 and 12 and divided into five categories: 0-2.4 (low), 2.5-4.8 (low-med), 4.9-7.2 (medium), 7.3-9.6 (med-high), 9.7-12.0 (high).

Climate dataset: Climate Data Online (CDO)[6] provides free access to National Climatic Data Center (NCDC)'s archive of global historical weather and climate data. We collected daily and monthly temperature and precipitation data generated since January 2013 (because the earliest allergy-related twitter data we have was generated in January 2013) for US major cities and states. More than half of the climate data collecting stations did not report daily temperatures at all, and many, among those that did report temperature, had missing values.

Allergy patients' dataset: We use the data from the first Quest Diagnostics Health Trends allergy report, Allergies Across America[7]. This report is the largest analysis of allergy testing of patients in the United States under evaluation for medical symptoms associated with allergies. We collected the ranked list of most prevalent food allergies grouped by patients' age. We also collected ranked list of the worst U.S. cities for different allergy types.

B. Methodology

1) Data Preprocessing: As we are interested in messages that mention actual allergy incidents, we removed all retweets (20.51% of our initial dataset) and tweets that are not written in English (2.9% of our initial data set). Special HTML characters are replaced with human-readable characters (e.g., replace < with < (i.e., less-than sign), replace > with > (i.e., greater-than sign)) and all hyperlinks are replaced with string 'URL'.

2) Data Classification: While some tweets talk about a person having allergy symptoms, other tweets talk about news, questions, general awareness of allergy season or information/advertisement regarding allergy medicine/treatments. It is important to distinguish tweets that mention actual allergy incidents to infer precise allergy levels. Hence, we classified tweets into two classes. First, we manually labeled 2000 randomly selected tweets into *positive* and *negative*. A tweet is labeled as *positive* if it talks about the author or someone around the author having allergy symptoms. A tweet is labeled as *negative* if it talks about news, advertisement, or general awareness of allergies. Table I shows examples of positive and negative labels. The text in parenthesis indicates the reason for the positive or negative annotation. We used a bag-of-words text classification where n-grams in documents are used as features. We removed common stop words except the following pronouns: *I, me, my, you, your* as we found that these pronouns were important features in classifying tweets into positive and negative examples of actual incident of allergy. To create features, we applied Weka [15]'s StringToWordVector filter. All unigrams, bigrams, and trigrams are used, if they appeared at least twice in the training data, to construct the feature vector. Then the filter converted words into their stems, applied TF-IDF weighting scheme, and kept 500 most frequently used n-grams in the final feature vector. We then explored four different machine learning algorithms (NaiveBayes (NB), NaiveBayes Multinomial (NBM), Random Forest (RF), Support Vector Machine (SVM)) that are commonly used for text classification. In our classification

Table I: Tweets with positive and negative labels. A tweet is positive if it talks about the author or someone around the author having allergy. A tweet is negative if it talks about news, question, general awareness or information about allergies.

Positive(+1) / Negative(-1)	Tweet
+1	My allergies are going insane today. (Author has allergy)
+1	Stupid allergies not letting me sleep. (Author has allergy)
+1	Recently my lovely allergy to cats has led to my throat closing up n barely being able to breathe. (Author has allergy)
+1	I never been able to enjoy spring cause my allergies. I hate having itchy eyes and running nose. (Author has allergy)
+1	@user1 @user2 and @user3 are all dying because of their allergies.. and Im just sitting here.. #popapill (People around author have allergies)
-1	In the United States, around 15 million people have food allergies, according to Food Allergy Research and Education. (News)
-1	Does anyone know good food near Happy Hollow that has vegetarian options and is easy for seafood allergies? (General question)
-1	Notice the increase in allergy ads on TV? Yep, spring is around the corner. (Awareness about spring season)
-1	RT @CureAllergies: What You Should Do To Manage Your Allergies - URL. (Information for allergy management)

Table II: Classification performance of various classifiers using 10-fold cross validation. The best classification performance (F-measure of 0.811 and ROC area of 0.905) was obtained using NaiveBayesMultinomial (NBM).

Classifier	Precision	Recall	F-measure	ROC Area
NBM	0.811	0.811	0.811	0.905
NB	0.799	0.793	0.793	0.864
Random Forest	0.812	0.800	0.799	0.888
SVM	0.818	0.810	0.809	0.814

[5]http://www.pollen.com/

[6]http://www.ncdc.noaa.gov/cdo-web/

[7]https://www.questdiagnostics.com/dms/Documents/Other/2011_QD_AllergyReport.pdf

task, both precision and recall are equally important. Thus, F-measure and ROC area are used to compare performance of classification algorithms.

As shown in Table II, the best classification performance (F-measure of 0.811 and ROC area of 0.905) was obtained using NBM and 10-fold cross validation on labeled data. We built a model using NBM on our training set, and classified all remaining tweets (after removing retweets and tweets in non-english) into positive or negative. We used NBM because it had best performance on our training data, and several prior works have shown NBM to outperform other classication algorithms. For example, McCallum and Nigamcite [22] found NBM to outperform simple NB, especially at larger vocabulary sizes, and Lee et. al. [18] showed performance of NBM to be better than that of NB or SVM in 18-class tweet text classification. In our entire allergy corpus, 63% of tweets were classified as *positive* and 37% were classified as *negative*. Only tweets in positive class (i.e., tweets classified as mentions of actual allergy incidents) are used for our analysis.

TF-IDF (term frequency–inverse document frequency) [21]. The tf-idf measure allows us to evaluate the importance of a word to a document. The importance is proportional to the number of times a word appears in the document but is offset by the frequency of the word in the document. Thus tf-idf is used to filter out common words.

NaiveBayes Multinomial (NBM) [22]. A document in NB would model as the presence and absence of particular words. A variation of NB is Naive Bayes Multinomial (NBM), which considers the frequency of words and can be denoted as:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c), \tag{1}$$

where $P(c|d)$ is the probability of a document d being in class c, $P(c)$ is the prior probability of a document occurring in class c, and $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c.

3) Text Mining: We wanted to investigate whether we could automatically discover the most predominant allergy types that people suffer from or talk about on social media by examining the texts in twitter posts. From our allergy-related tweet corpus, we extracted most frequently occurring 2-grams where the second word is 'allergy'. N-gram is a contiguous sequence of n words in a sequence of text. N-gram models are widely used in statistical natural language processing.

Part-Of-Speech (POS) tagging is a process of tagging a word with a part-of-speech (lexical category) such as noun, pronoun, verb, adjective, etc. We applied POS tagging to each 2-gram. For example, the POS tag for string 'natural allergy' is 'adjective noun' and the POS tag for string 'peanut allergy' is 'noun noun'. Table III shows the list of 15 most frequently used 2-grams and corresponding POS tags in the descending order of frequency of use.

Table III: A list of most frequently used 2-grams where the second word is *allergy*, ranked by frequency of use in entire allergy corpus. It includes many actual allergy types that are in 'noun noun' POS tag.

Rank	Most Frequently Used 2-grams	POS-tag
1.	food allergy	noun noun
2.	peanut allergy	noun noun
3.	gluten allergy	noun noun
4.	nut allergy	noun noun
5.	natural allergy	adjective noun
6.	hate allergy	verb noun
7.	skin allergy	noun noun
8.	lower allergy	comparative-adjective noun
9.	cat allergy	noun noun
10.	milk allergy	noun noun
11.	issues allergy	verb noun
12.	worst allergy	superlative-adjective noun
13.	dog allergy	noun noun
14.	severe allergy	adjective noun
15.	pollen allergy	noun noun

Our assumption is that the POS tag of all allergy types (e.g., food allergy, nut allergy, pollen allergy, dust allergy, egg allergy) should be in the form of 'noun noun' and, therefore, we can obtain a list of allergy types by removing all 2-grams that are not in 'noun noun' form. In other words, we need to remove all 2-grams that contain non-nouns (e.g., natural allergy (adjective noun), worst allergy (superlative-adjective noun)) to get the final list of allergy types. All 2-grams that contain twitter screen name (e.g., @username), stop words or non-english words are also removed.

4) Spatio-temporal Mining: Every tweet comes tagged with a timestamp that indicates the time when the tweet is posted. For example, the timestamp 'Sun Mar 02 05:55:02 +0000 2014' indicates that the tweet is created on Sunday, March 2, 2014 at 5:55am GMT (Greenwich Mean Time). Since we are interested in tracking allergy levels over time, we use the timestamps to count the volume of tweets posted each day that mention allergy or a specific allergy type, symptom, or treatment.

There are two types of tweet location, a sensor-based geolocation and a text-based user profile location. A geolocation provides the exact location where the tweet was posted with latitude and longitude values. This data is available to others only if the twitter user selects it to be publicly available. Twitter users can identify home location in his/her twitter user profile. We examined user profile locations and extracted state information. Examples of users' home locations that have state information are 'Riverside, CA', 'somewhere in NY' and 'Gainesville, Florida'. Examples of home locations that lack state information are 'Home Sweet Home', 'Somewhere over the rainbow' and 'Traveling'. We tag each tweet with a 2-character state code (e.g., CA for California) if we are successful extracting state information from twitter user profile.

Some tweets have both geolocation and user profile location, some have one or the other, and the rest do not have any location information. Geolocations are first translated into human-readable addresses using reverse geocoding API[8] and then the state name is extracted from the address. For tweets that do not have geolocation, we obtain state name from the user profile. Those that do not have any of the two locations are not used in the spatial analysis.

Table IV: 30 most frequently mentioned allergy types automatically extracted by our algorithm. Numbers indicate the rank of frequency the 2-gram appears in the allergy corpus and +/- signs indicates whether it is an actual allergy type(+) or not(-). 26 out of 30 were true positives achieving a very high precision of 86.7%.

Rank	Allergy Types (positive(+) / negative(-))		
1.	food allergy (+)	16.	shellfish allergy(+)
2.	peanut allergy (+)	17.	claritin allergy(-)
3.	gluten allergy (+)	18.	drug allergy(+)
4.	nut allergy (+)	19.	eye allergy (+)
5.	skin allergy (+)	20.	asthma allergy (-)
6.	cat allergy (+)	21.	sun allergy (+)
7.	milk allergy (+)	22.	mucinex allergy (-)
8.	dog allergy (+)	23.	prescription allergy (-)
9.	pollen allergy(+)	24.	nickel allergy (+)
10.	spring allergy(+)	25.	meat allergy (+)
11.	latex allergy (+)	26.	bee allergy (+)
12.	dairy allergy (+)	27.	alcohol allergy (+)
13.	dust allergy (+)	28.	seafood allergy (+)
14.	egg allergy (+)	29.	mite allergy (+)
15.	wheat allergy(+)	30.	penicillin allergy (+)

Table V: Most prevalent food allergies. The rank of the most prevalent food allergies extracted from twitter data is very similar to that obtained from actual allergy patients' data.

	Ground Truth		Twiiter Data
Rank	Most prevalent food allergies (Age>10)	Rank	Most mentioned food allergies
1.	peanut allergy	1.	food allergy
2.	wheat allergy (gluten allergy)	2.	peanut allergy
3.	soybean allergy	3.	gluten allergy
4.	milk allergy	4.	nut allergy
5.	egg allergy	5.	milk allergy
		6.	dairy allergy
		7.	egg allergy
		8.	wheat allergy

III. Experimental Results

A. Text Analysis

Allergy Types. Instead of using a pre-defined keyword list, we automatically identified allergy types mentioned in our dataset by using natural language processing methods. For the ground truth data, we created a list of allergy types by combining data from multiple online resources[9]. Table IV lists top 30 most frequently mentioned allergy types extracted from our allergy corpus by applying methods described in section II-B3. The numbers indicate the rank of frequency (1 means the highest frequency, 30 means the lowest frequency). The signs in the parenthesis indicate whether the extracted allergy type is positive (an actual allergy type) or negative (not an actual allergy type). Out of top 30 allergy types, 26 were true positives and only 4 were false positives, leading to precision of 86.7%. Two of the four false positive cases (claritin, mucinex) were allergy medicines, and the other two cases were allergy-related disease (asthma) and term (prescription). The traditional method that uses a pre-defined keyword list often fails to identify new types of diseases, and new keywords (i.e., new disease types) have to be manually added. However, with our proposed method that automatically identifies disease types, we would not need the step where new disease types are manually added.

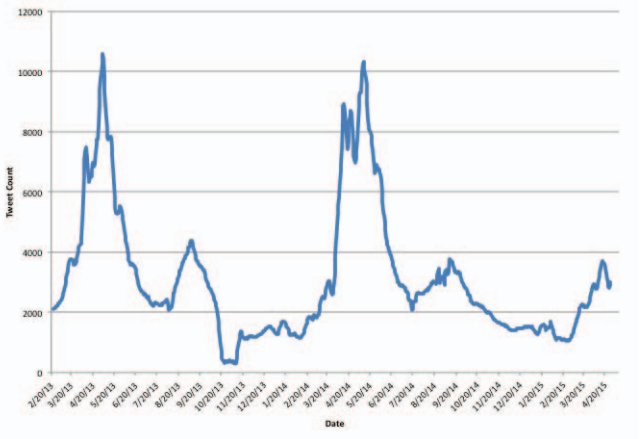


Figure 1: Time-series graph of daily allergy levels detected in tweets (February 2013 - April 2015). Only those allergy-related tweets labeled as positive are used to create the graph. The graph illustrates the general allergy level trend over time. The allergy level is the highest in mid–May, goes down in June and July, starts rising again in August, and reaches it local maximum point in mid–September. Similar seasonal pattern is observed in both 2013 and 2014.

Most Prevalent Food Allergies. We further evaluate our twitter data analysis results by comparing it to the real-world allergy patients' data. Table V shows the ground truth value of the most prevalent food allergies in allergy patients in the first column and the list of most mentioned food-related allergy types from table IV. We use the data for patients older than age ten because most twitter users fall into this age group. The allergy types in two columns show that they are in a very similar order of ranking. Note that gluten and wheat allergy can be considered to be the same and milk and diary allergy can also considered to be the same. This proves not only how the extracted allergy types are precise in identifying actual allergy types, but also the rank of prevalent allergy types have a very strong relationship to the real-world allergy patients data.

[8]https://developers.google.com/maps/documentation/geocoding/

[9]http://www.foodallergy.org/allergens, http://www.webmd.com/allergies/guide/allergy-symptoms-types, http://acaai.org/allergies/types, http://www.healthline.com/health/allergies/alcohol

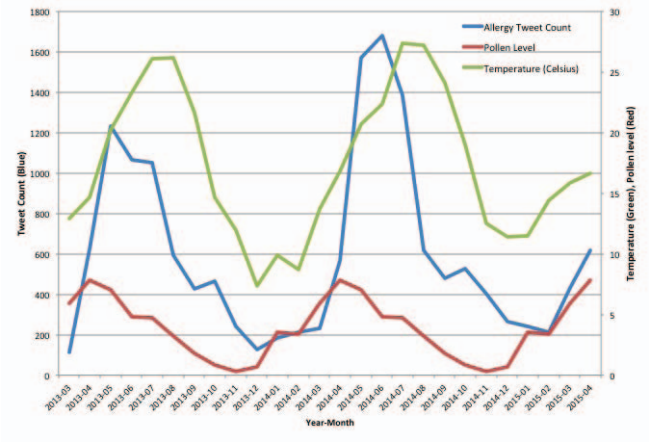


Figure 2: Monthly average data for allergy tweet count (blue), daily highest temperature (green), and pollen level (red) for Washington state (March 2013 – April 2015). Pollen level is highly correlated with Δtemperature (correlation of 0.776) and Δtweet count (correlation of 0.706). Tweet count is very strongly correlated with temperature (correlation of 0.668).

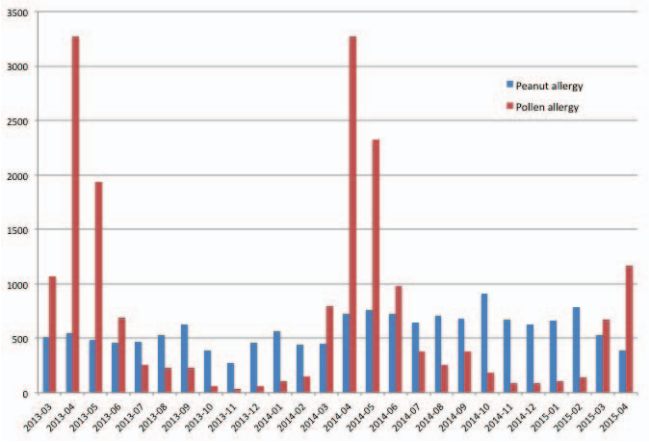


Figure 3: Monthly distribution of mentions of peanut and pollen allergies (March 2013–April 2015). A huge seasonal variation is observed in monthly pollen allergy (a seasonal allergy) level compared to that of peanut allergy (a food allergy).

B. Spatio-Temporal Analysis

In temporal model, we track activities of allergy, various allergy types, symptoms and medications over time using tweet timestamps. Figure 1 shows the allergy related tweet volume changes over two-years period from February 2013 through April 2015. The allergy level reaches its annual global maximum in mid-May and a local maximum in mid-September and this seasonal pattern is observed in both 2013 and 2014. The increased number of people chatting about their allergies in May and in September indicates that a very large population suffers from spring allergies such as tree pollen allergies and there is also a quite large population that has allergy symptoms in the fall.

To validate our experimental results, we compare our twitter data against the actual pollen levels and the weather data. Because pollen levels and temperatures vary depending on location, we partitioned allergy-related twitter data into finer space granularity to a US state level. Figure 2 compares three trend-lines: allergy tweet timeline (blue), monthly average pollen level (red), and monthly mean max temperature (green) for Washington state. We show Washington state data not just because a large volume of allergy-related tweets were generated in WA but also because the ground truth temperature data for WA was available for all dates from March 2013 through April 2015. It is clear from the graph that all three trend lines illustrate seasonality. An interesting pattern is that there is an order in time of three trend lines reaching their maximum and minimum points. The pollen level starts rising first and reaches its peak, followed by tweet counts and temperature. The trend lines also decrease in the same order.

Our analysis shows that the pollen level is highly correlated with rate of temperature change (correlation of 0.776) as well as with the rate of tweet count change (correlation of 0.706). In other words, pollen level reaches its peak point when the temperature sharply increases in spring, and, at the same time, allergy-related tweet volume also sharply increases. Also, tweet count has a strong correlation with daily temperature (correlation of 0.688), meaning allergy tweet count increases as the temperature increases. The high correlation values show how well the social media data reflects the real-world allergy activities and can be a good source of health data information.

In Figure 3, we track how the trend of mentions of two different allergy types differ over time. The tweet volume mentioning ‘pollen allergy’ (a seasonal allergy) rises very high during the spring and the fall and remains very low in the summer. However, unlike pollen allergy, the tweet volume mentioning ‘peanut allergy’ (a food allergy) stays relatively constant throughout the year. Note that we also carried out the same experiment in US state level and observed similar patterns in each state. This observation implies that the seasonality observed in overall allergy dataset in figure 2 and figure 3 comes from tweets mentioning various seasonal-allergy-related terms such as spring, tree pollen, or hay fever, rather than terms related to non-seasonal allergies such as dog, cat, milk or egg.

Figure 4 is a time-series graph showing tweet volume changes for different allergy symptoms. Sneezing (blue) is the most common allergy symptom throughout the year, followed by cough (green), runny nose (sky blue), watery eyes (red), and itchy throat (turquoise). It is very interesting that the rank for different allergy symptoms on each day is consistent throughout the year. Note that the percentage of

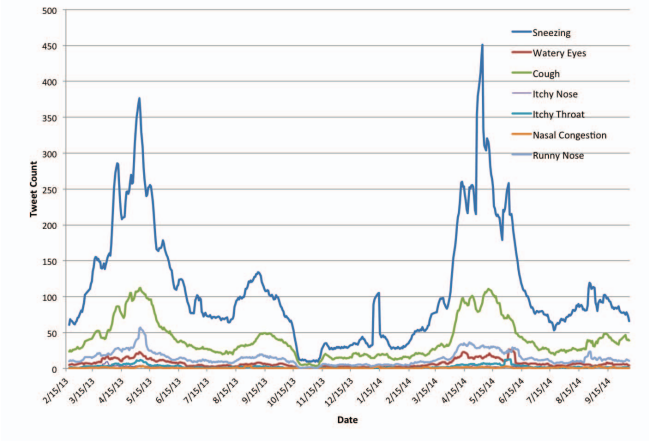


Figure 4: Time-series graph of tweet count for various allergy symptoms (Feb 2013–Sep 2014). The most common allergy symptom is sneezing (blue line) throughout the year, followed by cough (green) and runny nose (sky blue).

twitter users who enable their location publicly available has been steadily increasing since we started collecting our data.

For 20% of the tweets in our allergy data set, we were able to identify US state names. 11.4% of those had actual geolocation (longitude and latitude) values. For the remaining 88.6%, state names were extracted from the user profile locations.

Figure 5 shows monthly snapshot of tweets with geolocations that helps us visualize allergy levels across the U.S. Due to lack of space, we show one map per season for 2013. Each red dot on the map represents a tweet that was posted from the location. This map shows a general spatiotemporal trend of allergy activities. The allergy level starts increasing in early spring and gets extremely severe in May. It remains high throughout the summer, and goes down in the fall. Interestingly, most allergy-related tweets come from eastern part of the country although there are some from the west coast.

Next, using the US state information we obtained from geolocation and user profile locations, we visualize the distribution of tweets that mention different allergy types. Figure 6 compares levels of peanut allergy (blue bar) and gluten allergy (red bar) detected by social media sensors for each US state. The tweet count is normalized by state population and scaled to range between 0 and 100. The data normalization step solves the problem of higher number of tweets being generated from bigger states that have larger population. Kansas has the highest level of peanut allergy (94.51). South Dakota has the lowest level of both allergy types (3.85 for peanut allergy and 0 for gluten allergy). Most states have higher allergy level of peanut than gluten with a few exceptions. For example, unlike most other states, Oregon (OR), Delaware (DE), and Montana (MT) has higher gluten allergy levels.

IV. Related Works

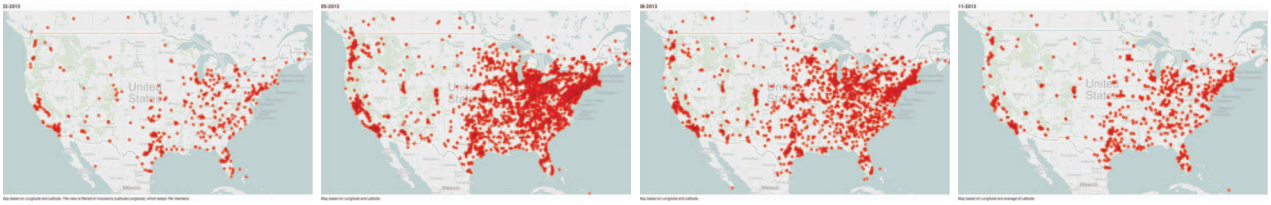
Before the Internet was widely used, over-the-counter pharmaceutical sales data [20] and telephone triage data [13] were among the methods that were used for surveillance of diseases.

Disease Surveillance using online data: In the past decade, with the dramatic increase of internet use, online data has been extensively used to retrieve health information and to detect disease activities. Web search queries data have been studied to track influenza activities. Ginsberg et al. [14] used flu-related google search queries data to estimate current flu activity near real time, 1-2 weeks in advance of the records by the traditional flu surveillance system[10]. Recent research on public health and disease surveillance using online data have mostly focused on monitoring and predicting influenza levels. Researchers have used twitter data to monitor influenza outbreak and to predict flu activities. Lee et al. [17] built a real-time disease surveillance system that uses twitter data to track flu activity. Signorini et al. [27] attempted estimating current influenza activity by tracking public sentiment and applying support vector machine algorithm on Twitter data generated during the Influenza A H1N1 pandemic. Chew et al. [9] analyzed content and sentiment of tweets generated during the 2009 H1N1 outbreak and showed the potential and feasibility of using social media to conduct infodemiology studies for public health. There are many others who have used Twitter data for flu outbreak detection [23, 8, 1, 28, 3, 7, 16, 26, 19]. Unlike earlier researchers who used twitter for flu activity detection and prediction, to the best of our knowledge, our work is the first attemp examining allergy activities using a large scale twitter stream.

Tweet Classification: Lee et al. [18] classifies trending topics into 18 general cagetories using text-based and network-based models. Aramaki et al. [2] proposed a Twitterbased influenza epidemics detection method that used Natural Language Processing (NLP) to filter out negative influenza tweets. Tuarob et al. [29] used ensemble machine learning techniques to identify health-related messages in a heterogenous pool of social media data. In our work, we use bag-of-words model and explore using four different machine learning algorithms to find the best model to classify tweets into those that mention actual allergy incidents and those that mention general awareness or information about allergy season.

Studying relationship between weather, pollen, and allergy: Many researches have studied the relationship between weather and pollen levels, and how it affects severity of allergy symptoms in patients [11, 30, 12]. In our work, the allergy levels are extracted from social media data instead of from allergy patients, and we study relationship between

[10]http://www.cdc.gov/flu/



(a) Feb 2013 (b) May 2013 (c) Aug 2013 (a) Nov 2013

Figure 5: Distribution of allergy tweets with geolocations. The seasonal pattern of allergy levels across U.S. is clearly visible. Allergy level is the highest in spring and the lowest in winter.

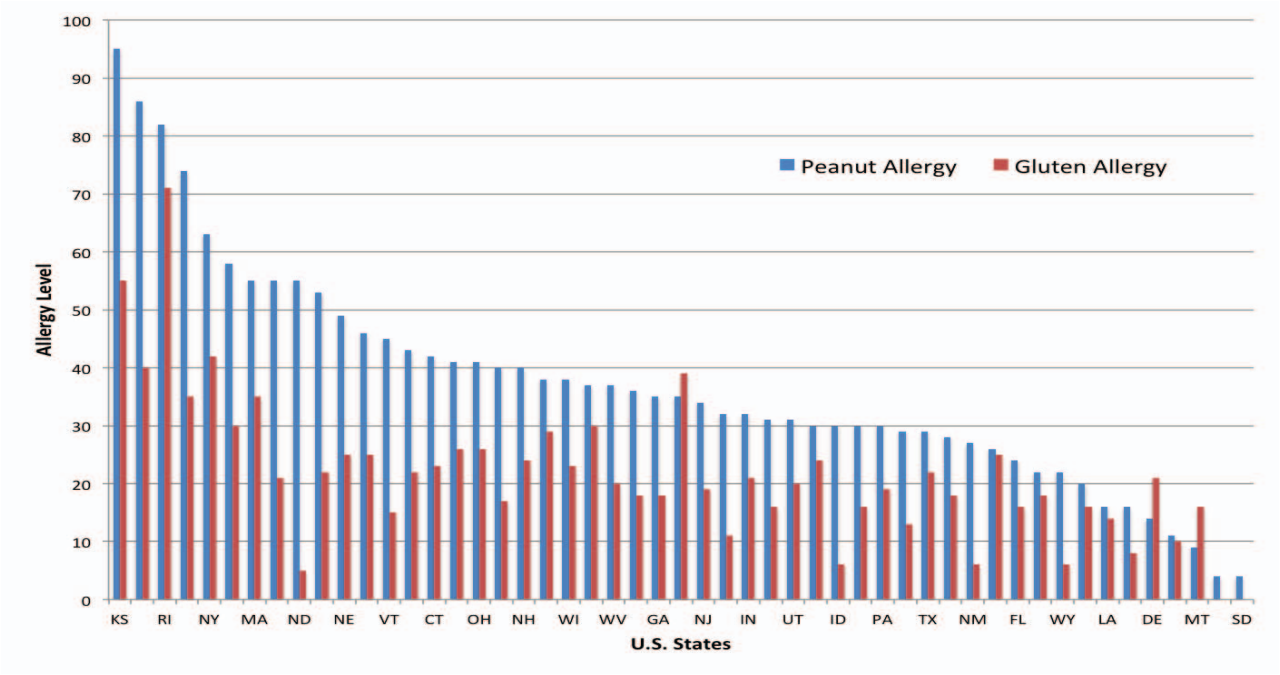


Figure 6: Bar chart comparing monthly social-media-sensed peanut and gluten allergy levels for each US state. The tweet count is normalized by state census population and scaled to range between 0 and 100. In most US states, peanut allergy level is higher than gluten allergy level.

the trend of allergy-related tweets with actual pollen levels and temperature at US state level.

In this paper, we focus on examining only allergy activity using a large Twitter stream collected over two years and show in-depth spatiotemporal analysis results. We also apply natural language processing techniques to automatically identify prevalent allergy types from Twitter contents.

V. Conclusion and Future Works

In this paper, we propose a system that monitors allergy levels near real-time by analyzing streaming twitter data. We first classify tweets to identify those that mention actual allergy incidents using bag-of-words model and Naive-BayesMultinomial classifier and then use those tweets with positive labels for text and spatiotemporal analysis.

We use text-mining techniques to automatically detect predominant allergy types. The top thirty allergy types extracted by our algorithm had precision of 86.7%. The experimental results further showed that the rank of the most prevalent food allergy types detected from tweet stream is highly correlated to the ground truth value, the ranked list of prevalent allergies obtained from real-world allergy patients' data.

We demonstrated that tweet time-series graph mentioning seasonal allergy related terms (e.g., pollen) show clear seasonal patterns (large volume of tweets during spring and low volume of tweets during winter) whereas those mentioning non-seasonal allergy related terms (e.g., peanut) remain relatively constant throughout the year. By studying relationship between allergy tweets with pollen and weather data, we showed all three data have similar seasonal patterns and allergy tweet data has a very strong relationship with the daily maximum temperature (correlation of 0.688).

We believe that our work is the first study that examines a

large-scale social media data for in-depth analysis of allergy activities. Although our work has specifically focused on studying allergy activities, the model can be generalized to track activities of other diseases.

In our future work, we are interested in investigating whether we could predict allergy level using social media data in conjunction with other physical data such as pollen counts and weather conditions. We will also apply more advanced NLP methods for automatic detection of prevalent allergy types to ensure better quality of extracted 2-grams.

ACKNOWLEDGMENT

This work is supported in part by the following grants: NSF awards CCF-1029166, IIS-1343639, and CCF-1409601.

REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 2011.

[2] 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 '11, pages 1568–1576, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[4] D. L. Blackwell, J. W. Lucas, and T. C. Clarke. Summary health statistics for u.s. adults: National health interview survey, 2012. http://www.cdc.gov/nchs/data/series/sr_10/sr10_260.pdf, 2013.

[5] B. Bloom, L. I. Jones, and G. Freeman. Summary health statistics for u.s. children: National health interview survey, 2012. http://www.cdc.gov/nchs/data/series/sr_10/sr10_258.pdf, 2012.

[6] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[7] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, and J. C. and. Forecasting a moving target: Ensemble models for ili case count predictions. In *SDM*, 2014.

[8] L. Chen, H. Achrekar, B. Liu, and R. Lazarus. Vision: Towards real time epidemic vigilance through online social networks: Introducing sneft – social network enabled flu trends. In *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*, MCS '10, pages 4:1–4:5, New York, NY, USA, 2010. ACM.

[9] C. Chew and G. Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):e14118, 11 2010.

[10] A. Choudhary, W. Hendrix, K. Lee, D. Palsetia, and W.-K. Liao. Social media evolution of the egyptian revolution. *Commun. ACM*, 55(5):74–80, May 2012.

[11] L. de Weger, T. Beerthuizen, P. Hiemstra, and J. Sont. Development and validation of a 5-day-ahead hay fever forecast for patients with grass-pollen-induced allergic rhinitis. *International Journal of Biometeorology*, 58(6):1047–1055, 2014.

[12] J. Emberlin, J. Mullins, J. Corden, W. Millington, M. Brooke, M. Savage, and S. Jones. The trend to earlier birch pollen seasons in the uk: a biotic response to changes in weather conditions? *Grana*, 36(1):29–33, 1997.

[13] J. U. Espino, W. R. Hogan, and M. M. Wagner. Telephone triage: a timely data source for surveillance of influenza-like diseases. In *AMIA Annual Symposium Proceedings*, volume 2003, page 215. American Medical Informatics Association, 2003.

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

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[16] V. Lampos, T. De Bie, and N. Cristianini. Flu detector-tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. Springer, 2010.

[17] K. Lee, A. Agrawal, and A. Choudhary. Real-time disease surveillance using twitter data: Demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1474–1477, New York, NY, USA, 2013. ACM.

[18] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 251–258. IEEE, 2011.

[19] J. Li and C. Cardie. Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*, 2013.

[20] S. Magruder. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL technical digest*, 24(4):349–53, 2003.

[21] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[22] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*, pages 41–48. AAAI Press, 1998.

[23] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*. The AAAI Press, 2011.

[24] R. Pawankar, G. W. Canonica, S. T. Holgate, and R. F. Lockey. Wao white book on allergy. http://www.worldallergy.org/UserFiles/file/WAO-White-Book-on-Allergy_web.pdf, 2011.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *In Proceedings of the Nineteenth International WWW Conference (WWW2010). ACM*, 2010.

[26] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4, Dec. 2013.

[27] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5):e19467, 05 2011.

[28] M. Sofean and M. Smith. A real-time architecture for detection of diseases using social networks: Design, implementation and evaluation. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 309–310, New York, NY, USA, 2012. ACM.

[29] S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram. Discovering health-related knowledge in social media using ensembles of heterogeneous features. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pages 1685–1690, New York, NY, USA, 2013. ACM.

[30] M. Wilson, S. Villalba, H. Avila, J. Hahn, and A. Cepeda. Correlation between atmospheric tree pollen levels with three weather variables during 2002-2004 in a tropical urban area. *Journal of Allergy and Clinical Immunology*, 127(2):AB170, 2011.