Social Media Sensing Framework for Population Health

Alvaro Esperanca
Electrical and Computer Eng. Dept.
Indiana Univ. Purdue Univ. Indianapolis
Indianapolis, IN 46202
aesperan@iu.edu

Zina Ben Miled
Electrical and Computer Eng. Dept.
Indiana Univ. Purdue Univ. Indianapolis
Indianapolis, IN 46202
zmiled@iu.edu

Malika Mahoui
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, IN 46285
mahoui_malika@lilly.com

***Abstract*— Conducting large health population studies is expensive. For instance, collecting field information about the efficacy of health campaigns or the impact of a disease may require the involvement of many health providers over an extended period of time and sometimes may not reach the target population. In fact, due to the aforementioned difficulties, health-related population statistics may be unavailable or lag by several years. Recently, social media networks have emerged as a source of sensory data for various aspects of social behavior. This source of information is used to drive marketing campaigns, conduct threat analysis and profile groups of individuals among numerous other applications. However, these applications are usually limited to specific case studies and do not provide a systematic approach to translating social media data into knowledge. In this paper, we propose a framework that can extract knowledge from social media networks in support of large scale health studies. The framework consists of an automated workflow designed to collect data from social media platforms, filter the data based on geographical criteria, and extract information relevant to a target hypothesis. The framework is demonstrated in the case of mortality and incidence of three chronic diseases, namely asthma, cancer, and diabetes. However, the utility of the proposed framework extends to other areas in the health sector and can help automate data-driven hypothesis validation for social media studies.**

***Index Terms*—SVM, Social Media, Data Mining, Classification**

I. Introduction

Social media data and the volume of information it embodies can help improve the way large scale population studies are conducted in various fields. This communication media is in fact increasingly becoming a reflection of societal behavior at scale and a sensor for trends and daily activities for individuals as well as for population groups. The field of marketing is already taking advantage of this data to, for example, customize campaigns and efficiently collect sentiments about various products. While there are clear benefits to social media sensing, obtaining statistically reliable information from this source is challenging. Indeed, social media offers a large repository of data. However, this data is noisy and unstructured. Therefore, previous studies that rely on social media data often require substantial domain expertise, are limited in scope, and often necessitate a substantial amount of manual processing. Our aim is to develop a structured workflow that can provide the answer to health-related questions using social media data. The capabilities of the proposed framework are demonstrated through a case study that aims at analyzing the mortality and incidence counts for three chronic diseases (i.e., asthma, cancer, and diabetes) in the US. The social network of choice in this case study is Twitter. Mortality and incidence trends for the selected diseases are investigated at the national level and the state level. Moreover, for validation purposes, the results are compared to published incidence and mortality counts by the Center of Disease Control (CDC).

The remainder of the paper is organized as follows. Section 2 includes a review of related work. Section 3 describes the framework. Section 4 illustrates the use of the framework for the case study and Section 5 summarizes our findings and outlines direction for future work.

II. Related Work

According to [1], more than 8 in 10 online users in the United States use social media and 25% of online adults use Twitter. This elevated level of penetration, the ease of access coupled with the extended reach, made social media networks an attractive source of information for social behavior studies at scale. As a result, increasing research efforts are focusing on using social media to detect trends and sentiments in various sectors [2].

In the health sector, twitter was used, for example, to monitor and predict the spread of influenza [3]–[5]. It was also used to monitor the adverse effect of medications in [6] and [7], track medication adherence in [8] and for the understanding of the well-being of military populations in [9].

These studies demonstrate the value of social media data in public health but also highlight the difficulties associated with social media sensing including:

- developing the query lexicon for the target hypothesis and
- reducing the level of noise in the extracted data.

Despite the potential benefits, we believe that the above difficulties explain the limited duration or scope of previous social media sensing studies in public health. For instance, [3] and [5] track influenza over a fleeting period of one and two months, respectively. Medication adverse effect is studied in [6] over a period of 6 months. With respect to scope, in [3] a limited number of cities are covered and in [5] the spread

This is the author's manuscript of the article published in final edited form as:

Esperanca, A., Miled, Z. B., & Mahoui, M. (2019). Social Media Sensing Framework for Population Health. 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 0298–0304. https://doi.org/10.1109/CCWC.2019.8666534

of influenza is investigated at the national level rather than at the state or county levels.

This paper highlights the investigative potential afforded by an automated framework that can extract health data from social media over an extended period for regions with varying geographical boundaries. This potential is demonstrated by investigating the use of social media mentions as a sensor for actual incidence or mortality rates for three chronic diseases. Previous work [11] show that relying on social media for health data sensing can lead to prediction errors [12]. Therefore, the ability to rapidly extract and classify data from social media can help support a stronger validation for social media sensing in the health sector.

As mentioned above, the first difficulty associated with social media sensing is the substantial effort needed to develop an adequate lexicon [3], [5] in order to extract data relevant to a given hypothesis. Domain expertise has been the standard method for building the query lexicon for health related studies. In this paper an iterative approach is used. It starts with generic terms and incrementally refines the lexicon.

The second difficulty relates to reducing the level of noise inherent to social media data. In the remainder of the paper we equate noise to semantically irrelevant tweets with respect to the hypothesis. Detecting and subsequently eliminating the irrelevant tweets can be achieved by using a classifier. Several classifiers were used in [8] for the study of medication adherence using Twitter. These classifiers include Bayesian networks, random forests, logistic regression, and support vector machines (SVM). The SVM classifier was found to have higher accuracy compared to the other three classifiers and therefore it was adopted in this paper.

An automated framework based on an SVM classifier was succesfully used to predict movie box-office success from multiple social media sources [13]. This paper is inspired by this previous work but targets a different sector, namely, large scale population health.

III. Social Media Sensing Framework

The proposed framework is a structured workflow that can turn social media data into an extended and dynamic population health sensor. The stages of the workflow are shown in (Figure 1) and consist of:

- Query lexicon: The lexicon contains the seed query keywords that are used to initiate data collection from the social media network.
- Data extraction: During this stage of the workflow, records related to the keywords in the lexicon are extracted. These records are mapped and tagged according to a geographical area of interest.
- Classification: This stage reduces the level of noise in the extracted data. A classifier is used to organize the extracted data according to its relevance to the hypothesis.
- Validation: During this stage of the framework, the hypothesis is assessed based on the data resulting from the previous stages and available benchmark data from other sources.

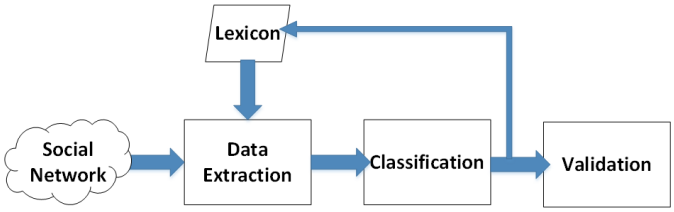

Fig. 1. Social media sensing workflow.

These above stages of the framework are described in the following subsections.

A. Query Lexicon

The seed query lexicon for each hypothesis is derived from generic terms that refer to the target diseases. The lexicon in the proposed framework starts with a limited set of seed keywords and is enriched through an iterative procedure. At each iteration, a subset of the records that are false negative are examined in order to extract high frequency keywords. These keywords can then be used to augment the lexicon. Depending on the target hypothesis, only few iterations may be needed to obtain a comprehensive lexicon. The case study that is presented in the next section of the paper shows that a limited number of keywords is also sufficient. For example, when investigating trends related to cancer, the initial lexicon only included the keyword cancer. After one iteration of the workflow, the keyword cancerous was added to the lexicon. However, the extended lexicon did not yield a significant improvement in the query search space in this case.

B. Data Extraction

Data is collected from Twitter by querying the Twitter's advanced search page. Only tweets that were made publicly available are collected. These tweets include the text, the handle, the date, the permalink and the geolocation. The text is the message a user posts and the handle is a unique identifier for each user. The permalink uniquely identifies a tweet and the geolocation consists of the latitude and longitude of either the tweet or the user.

A search query is initiated for each keyword in the lexicon. The result of the query is a set of records that contain the keyword as part of the handle or the text. For instance, if the handle of a user includes the keyword cancer (e.g., @_CANCERLOVE), then all the tweets posted by this user will be retrieved regardless of their relevance to cancer.

Each record is then mapped to a specific geographical area according to a set of geographical filters. These filters are customized to each state and are derived using a mapping tool [14]. For example, Figure 2 depicts the filters for the state of Virginia. These filters are required because Twitter only allows geolocation filtering by circular regions defined by a center and a radius. The center can be defined in terms of an address, a point of interest, or GPS coordinates (i.e., latitude and longitude). In the proposed framework GPS coordinates

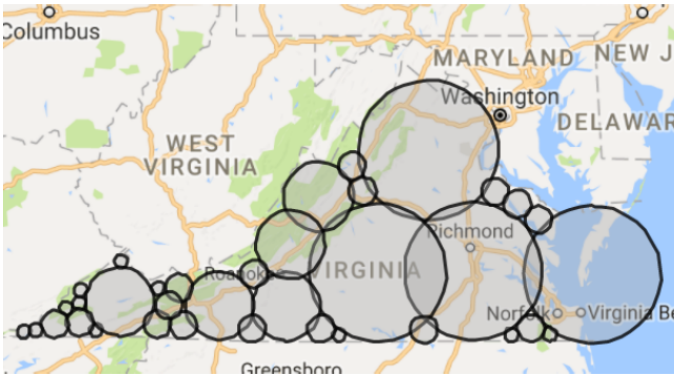

Fig. 2. Geolocation filters for the state of Virginia.

are used and a unique query is constructed for each keyword in the lexicon and each filter.

As shown in Figure 2, full coverage of a given state requires that some of the filters overlap. Therefore, a post processing step is needed to remove records with duplicate permalinks. Moreover, in Twitter the GPS coordinates of a tweet provide the most accurate geographical location. However, most users do not use this feature. In fact, out of the total 845,832 tweets collected in this study, only 12 tweets include a GPS coordinate. Therefore, the GPS coordinates of the handle are used to geotag the extracted tweets.

C. Classification

The records collected in the previous stage include a substantial amount of noisy data. The level of noise in the data can vary depending on the lexicon and the target hypothesis. However, in most cases, it is significant and has to be addressed. For instance, the tweet "If there's someone you know who's a Cancer" is extracted because it contains the keyword cancer despite its irrelevance to any health study focused on the disease cancer.

A classifier is used to reduce the level of noise in the dataset. SVM was selected as the classifier of choice because it was previously shown to be efficient in classifying natural language [15]. In general, SVM identifies the optimal decision boundary that separates the dataset into different classes after projecting that same data onto a different space using a kernel. The choice of an appropriate kernel depends heavily on the type of data to be classified [16]. If the data is linearly separable, a linear kernel is sufficient [15]. Otherwise a different type of kernel has to be used. In developing the proposed framework, three different kernels were explored: linear, third degree polynomial and RBF. These kernels are given by equations 1, 2 and 3, respectively.

$$K(\mathbf{X}, \mathbf{Y}) = \mathbf{X} \cdot \mathbf{Y} \tag{1}$$

$$K(\mathbf{X}, \mathbf{Y}) = (1 + \mathbf{X} \cdot \mathbf{Y})^3 \tag{2}$$

$$K(\mathbf{X}, \mathbf{Y}) = exp\Big(- \frac{\|\mathbf{X} - \mathbf{Y}\|^2}{2\sigma^2} \Big) \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ refer to two records in the dataset.

The cost parameter (C) for the SVM classifier was also varied. This parameter defines a soft margin in the decision boundary and impacts the number of mis-classifications allowed by the classifier. The higher the value of C, the lower the number of mis-classifications.

A classifier is developed for each hypothesis and has to be trained on the corpus relevant to the hypothesis. The records retrieved in the data collection stage are pre-processed prior to being presented to the classifier. For each record, the text is tokenized, and the stop words are removed. Moreover, a training set consisting of 1,000 records is randomly sampled from the entire dataset. These records are manually inspected and labeled with -1 for irrelevant, and 1 for relevant. For example, the tweet

> *@CancerFollowers: #Cancer may be stubborn and hard headed*

is labeled as irrelevant (i.e., given a label of -1) because it refers to the zodiac sign cancer whereas the tweet

> *Great grandma has liver cancer smh*

is labeled as relevant (i.e., given a label of 1) due to its reference to the disease.

The last step in the pre-processing consists of deriving a numerical representation of the training set. This numerical representation is a matrix where each column corresponds to a token and each row represents a tweet. The entries in the matrix correspond to the number of occurrences of a token in a tweet. Once the pre-processing phase is complete, the model can be trained using the resulting matrix.

After the training phase, the classifier is presented with the entire pre-processed dataset. The pre-processing at this stage is identical to the pre-processing of the training data except for the labeling of relevant and irrelevant tweets. Each tweet in the dataset is labeled by the classifier as relevant or irrelevant. In order to assess the accuracy of the classifier, a testing set consisting of 1,000 tweets is randomly sampled and manually inspected using an approach identical to that of the training set. This testing set is needed because reviewing all the records in the collection is impractical. The manual labels given to each record in the testing set are then compared to those generated by the classifier. The comparison results in a confusion matrix consisting of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

If the number of mis-classifications is high, the classifier is iteratively improved as mentioned above whereby the FNs in the test set undergo a frequency analysis and the high frequency tokens are added to the lexicon. Alternatively, another set of tweets can be randomly sampled, labeled, added to the training set and used to retrain the model. In both cases, manual processing is limited to the labeling of the subset of training and testing datasets used for each iteration.

D. Validation

Once the classification is complete, the extracted data is ready to be used for hypothesis validation. This step is important in order to evaluate the resulting model as a sensor for

health indicators before deploying it in production. Validation is usually performed by comparing the predictive results of the model against a well established benchmark. For the case study being considered in this paper, the validation is based on the correlation between the volume of twitter activity for the target chronic disease and either the corresponding incidence or mortality rate as published by CDC. Other statistical validation approaches can also be considered.

IV. Case Study of Three Chronic Diseases

In order to demonstrate the potential use of the proposed framework, it was used to investigate the following hypothesis: *Is the volume of Twitter activity related to the chronic diseases asthma, cancer, and diabetes correlated with the respective incidence or mortality rates of these diseases?* A positive answer to this question will support the use of Twitter data as a cost-effective mechanism for tracking incidence or mortality rates for these chronic diseases for targeted regions in the US. A negative answer will highlight the limitations of the use of social media data as a sensor for these chronic diseases. The choice of the target diseases was motivated by their high incidence and mortality rates in the US.

A. Methodology

Towards the above objective and using the proposed framework, three collections of tweets were extracted, one for each target disease, namely, asthma, cancer, and diabetes.

TABLE I
Query lexicon.

Asthma	**Cancer**	**Diabetes**
Original query lexicon		
asthma	cancer	diabetes
asthmatic	cancerous	diabetic
Extended query lexicon		
attack	breast	type
allergy	awareness	sugar

The seed query words for each disease are shown in Table I. These keywords make-up the original query lexicon that is used in the first iteration of the framework to retrieve tweets for each disease. Table II shows the number of tweets retrieved using the lexicon in Table I for the period of 2010 to 2015. This table also shows the additional number of tweets obtained by using an extended lexicon which was derived from the results of the first iteration of the framework. The added high frequency keywords include "attack" and "allergy" for asthma, "breast" and "awareness" for "cancer", and "type" and "sugar" for Diabetes.

The time interval of 2010 to 2015 was selected because: a) Twitter activity started increasing significantly in 2010 [17] and b) CDC only provides incidence and mortality data up to 2015.

Sets of 1,000 records were randomly sampled from each of the three collections of tweets and manually labeled. The resulting training matrices consisted of 1,000 rows (i.e., tweets) and 2,982, 4,306, and 3,828 columns (i.e., features) for asthma, cancer and diabetes, respectively. These matrices were then used to train the classifiers in order to identity relevant and irrelevant tweets for each disease. Testing sets consisting of 1,000 records per collection were also randomly selected and manually labeled.

TABLE II
Number of tweets extracted from 2010 to 2015 by using the original and the extended lexicons.

	Asthma	Cancer	Diabetes
Number of tweets extracted by using the original query lexicon	69,549	731,874	102,914
Additional number of tweets extracted by using the extended query lexicon	6,677	16,715	29,268

When developing the classifiers, the linear (Equation 1), polynomial (Equation 2) and RBF (Equation 3) kernels were evaluated. RBF was selected as the kernel of choice for the three classifiers because it yielded the best results compared to the other two kernels. Table III shows the performance of the resulting classifiers for the target diseases. These classifiers achieved high accuracy, precision and recall.

TABLE III
Accuracy, precision and recall for the classifier for each chronic disease.

	Asthma	Cancer	Diabetes
Accuracy	89%	75.7%	81%
Precision	89.57%	75.68%	81.31%
Recall	98.85%	100%	99.38%

B. Results

Table IV shows the total number of tweets retrieved for each year of the study period and for each of the target diseases. We observe a decline in the number of tweets in the year 2015. Possible causes for such decline are: a) interest levels for the diseases decreasing due to more prevalent events occurring that year; b) a significant number of Twitter users privatizing their accounts thus making their tweets publicly inaccessible. Despite this decline, the number of tweets that are publicly available may still be a representative sample of the target population.

Table V shows the total number of tweets classified as relevant for each target disease by the classifier. During the validation stage, the Pearson's correlation coefficient between tweet count and incidence count as well as tweet count and mortality count is calculated for each disease. This is done at the national level and at the state level. The incidence and mortality counts are obtained from CDC [18]. In order to evaluate the statistical significance of our findings, the p-value was calculated with an alpha-level of 0.05.

C. Analysis and Discussion

At the national level, the results (Table VI) show a moderate strength correlation between the tweet count and the incidence

TABLE IV
NUMBER OF TWEETS FOR EACH YEAR DURING THE PERIOD 2010 TO 2015.

Disease	**Asthma**	**Cancer**	**Diabetes**
2010	2,253	31,482	5,268
2011	6,585	73,103	11,712
2012	11,710	126,464	21,049
2013	20,265	181,174	33,942
2014	25,279	236,945	46,023
2015	8,497	99,421	14,188

TABLE V
NUMBER OF TWEETS CLASSIFIED AS RELEVANT FROM 2010 TO 2015 FOR EACH DISEASE.

Disease	Total Tweets	Relevant Tweets	Percent Relevant
Asthma	74,589	72,562	97.28%
Cancer	748,589	746,994	99.78%
Diabetes	132,182	130,655	98.85%

count as well as between the tweet count and the mortality count for each disease. It is interesting to note that the incidence count for asthma and diabetes are higher than that of cancer. However, cancer has a higher tweet count. This may be due to the fact that the rate of cancer mortality is higher than that of asthma and diabetes. That is, the deadlier the disease, the more likely it is to be mentioned in social media.

Running the model on a per-state basis shows inconsistent results with those at the national level. This supports the need for the proposed framework in order to easily validate data-driven social media model for population health. Tables VII, VIII, and IX show the states that exhibited significant correlation between the tweet count and the incident count as well as between the tweet count and the mortality count for asthma, cancer and diabetes, respectively. States that were omitted from the above tables either showed weak or statistically insignificant correlation. For instance, Alaska and Wyoming had limited number of tweets. Moreover CDC incidence and mortality rates for these states were missing for few of the years during the study period.

Table VII shows that tweet counts are highly correlated with the incidence count for the states of Hawaii, North Carolina, South Carolina, Tennessee and Virginia for asthma. In the case of Louisiana, Nevada and New Jersey, the tweet count is highly correlated with the mortality count of the disease. This indicates the potential for the first and second group of states to use Twitter in order to track trends in incidence and mortality rates for asthma, respectively. Similar observation can be made for cancer and diabetes from tables VIII and IX.

We also found that a state-to-state comparison offers some insightful information about trends in population health. For example, Texas and California have the largest populations and the highest volume of twitter activity. Both show a strong correlation for cancer incidence count. However, only California shows a strong correlation for diabetes incidence count. It should also be noted that neither of these states show a strong correlation for asthma.

TABLE VI
THE NUMBER OF TWEETS, THE INCIDENCE AND MORTALITY COUNTS FOR ASTHMA, CANCER, AND DIABETES AT THE NATIONAL LEVEL FOR EACH YEAR OF THE STUDY. THE LAST ROW IN EACH SECTION OF THE TABLE SHOWS THE CORRELATION BETWEEN THE NUMBER OF TWEETS AND THE INCIDENCE COUNT AND THE CORRELATION BETWEEN THE NUMBER OF TWEETS AND THE MORTALITY COUNT FOR EACH DISEASE.

Year	Tweets	Incidence	Mortality
Asthma			
2010	2,158	25,069,374	3,107
2011	6,423	22,605,965	3,061
2012	11,486	25,954,769	3,145
2013	19,584	26,227,467	3,295
2014	24,395	26,955,183	3,272
2015	8,408	25,839,239	3,255
Correlation		0.67	0.77
Cancer			
2010	31,425	1,608,786	574,738
2011	72,768	1,622,948	576,685
2012	125,737	1,609,724	582,607
2013	178,971	1,616,286	584,872
2014	234,178	1,653,798	591,686
2015	98,973	1,633,390	595,919
Correlation		0.64	0.60
Diabetes			
2010	5,254	21,326,457	69,071
2011	11,645	22,440,004	73,831
2012	20,828	24,095,092	73,932
2013	33,259	24,801,332	75,578
2014	45,107	25,709,248	76,488
2015	14,075	25,866,273	79,535
Correlation		0.70	0.45

TABLE VII
STATES WITH STATISTICALLY SIGNIFICANT CORRELATION BETWEEN TWEET COUNT AND INCIDENCE COUNT AS WELL AS BETWEEN TWEET COUNT AND MORTALITY COUNT FOR ASTHMA.

State	Incidence	Mortality
Hawaii	0.87	0.46
North Carolina	0.71	0.57
South Carolina	0.71	0.37
Tennessee	0.82	-0.52
Virginia	0.82	0.43
Louisiana	0.19	0.79
Nevada	0.23	0.78
New Jersey	0.30	0.70

Moreover, New York does not generate as much Twitter activity as one would expect. The total number of tweets for New York accounts for about 5% of the total number of tweets for each disease whereas these percentage for Texas and California are 12% and 10%, respectively.

V. CONCLUSION

This paper introduces an iterative workflow for social data sensing that requires limited manual intervention or domain expertise. During the first stage of the framework, social media data is extracted using an initial query lexicon composed of generic terms pertaining to a target disease. The extracted records are then processed by using an SVM classifier that determines the relevance of each record in the collection to the target disease. This classifier is based on an RBF kernel and it is trained and tested using a limited subset of 1,000

TABLE VIII
STATES WITH STATISTICALLY SIGNIFICANT CORRELATION BETWEEN TWEET COUNT AND INCIDENCE COUNT AS WELL AS BETWEEN TWEET COUNT AND MORTALITY COUNT FOR CANCER.

State	Incidence	Mortality
California	0.89	0.48
Colorado	0.92	0.53
Georgia	0.71	0.65
Hawaii	0.93	0.73
Kentucky	0.89	0.60
Louisiana	0.78	0.82
Maine	0.78	-0.43
Maryland	0.47	0.81
Mississippi	0.47	0.93
Missouri	0.68	0.76
Texas	0.90	0.77
Virginia	0.91	0.37

TABLE IX
STATES WITH STATISTICALLY SIGNIFICANT CORRELATION BETWEEN TWEET COUNT AND INCIDENCE COUNT AS WELL AS BETWEEN TWEET COUNT AND MORTALITY COUNT FOR DIABETES.

State	Incidence	Mortality
California	0.80	0.60
Delaware	0.75	0.51
Georgia	0.67	0.73
Hawaii	0.72	0.51
Illinois	0.73	0.47
Iowa	0.79	0.33
Maryland	0.36	0.85
Massachussetts	0.73	0.08
Michigan	0.25	0.71
Nebraska	0.75	-0.14
New Hampshire	0.78	0.71
New Jersey	0.86	-0.21
New Mexico	0.57	0.77
North Carolina	0.71	0.56
North Dakota	0.86	-0.44
Pennsylvania	0.67	0.80
Tennessee	0.84	0.29
Texas	0.63	0.60
Utah	0.75	0.37
Washington	0.70	0.22

randomly sampled records. It should be noted that the only records that require manual labeling are those included in the above training and testing subsets. Despite the limited number of records used for training (i.e., 1,000 records), the classifier had a high accuracy of 89%, 75.7%, and 81% for asthma, cancer, and diabetes, respectively. Moreover, only a limited number of iterations of the framework may be needed in order to achieve this level of accuracy. That said, both the number of records and the number of iterations required for training are hypothesis dependent.

The proposed framework was used to derive models for the target chronic diseases at the national level as well as at the individual state level over an extended time period of 5 years. These models would have been extremely tedious to develop without the proposed framework. In order to exemplify the potential of the framework to support health population studies, the hypothesis of whether or not social media can be used to track trends in incidence and mortality rates for the three selected target diseases was investigated. The results show that social media sensing for population health is possible. However, the approach has be applied with care as trends that may be applicable at a given scale may not translate to a different scale (e.g., from national to state level). It is recommended that the resulting models be validated against well established benchmarks and that the geographical target areas of the studies be varied in order to gain insightful understanding of the predictive capabilities of the models.

Future work will consider integrating demographics (e.g., age, gender) as well as including multiple social media sources. Moreover, the models presented in this paper have shown high classification accuracy with limited lexicons. We believe that this due to the simplicity of the language used in social media. Nonetheless, we would like to use topic modeling in order to refine the classification of the records, for example, into relevant records for incidence as opposed to mortality.

REFERENCES

[1] S. Greenwood, r. Perrin, and M. Duggan, "Social Media Update 2016," Nov. 2016. [Online]. Available: http://www.pewinternet.org/2016/11/11/social-media-update-2016/

[2] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[3] K. Byrd, A. Mansurov, and O. Baysal, "Mining twitter data for influenza detection and surveillance," in *Proceedings of the International Workshop on Software Engineering in Healthcare Systems*. ACM, 2016, pp. 43–49.

[4] S. Song and Z. B. Miled, "Digital immunization surveillance: Monitoring flu vaccination rates using online social networks," in *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2017, pp. 560–564.

[5] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 115–122.

[6] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta, "Digital drug safety surveillance: monitoring pharmaceutical products in twitter," *Drug safety*, vol. 37, no. 5, pp. 343–350, 2014.

[7] E. Tutubalina and S. Nikolenko, "Exploring convolutional neural networks and topic models for user profiling from drug reviews," *Multimedia Tools and Applications*, pp. 1–19, 2017.

[8] A. Klein, A. Sarker, M. Rouhizadeh, K. O'Connor, and G. Gonzalez, "Detecting personal medication intake in twitter: An annotated corpus and baseline classification system," *BioNLP 2017*, pp. 136–142, 2017.

[9] U. Pavalanathan, V. V. Datla, S. Volkova, L. Charles-Smith, M. Pirrung, J. J. Harrison, A. Chappell, and C. D. Corley, "Discourse, health and well-being of military populations through the social media lens." in *AAAI Workshop: WWW and Population Health Intelligence*, 2016.

[10] E. Santoro, G. Castelnuovo, I. Zoppis, G. Mauri, and F. Sicurello, "Social media and mobile applications in chronic disease prevention and management," *Frontiers in psychology*, vol. 6, 2015.

[11] Google, "Google Flu Trends." [Online]. Available: https://www.google.org/flutrends/about/ https://www.google.org/flutrends/

[12] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, mar 2014. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1248506

[13] Y. Lu, R. Krüger, D. Thom, F. Wang, S. Koch, T. Ertl, and R. Maciejewski, "Integrating predictive analytics and social media," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 2014, pp. 193–202.

[14] Draw a circle with a radius on a map. [Online]. Available: https://www.mapdevelopers.com/draw-circle-tool.php

[15] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, Sep 1999.

[16] D. Elizondo, "The linear separability problem: some testing methods," *IEEE Transactions on Neural Networks*, vol. 17, no. 2, pp. 330–344, Mar. 2006.
[17] "Twitter MAU in the United States 2017 | Statistic." [Online]. Available: https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/
[18] CDC, "CDC Works 24/7," Apr. 2017. [Online]. Available: https://www.cdc.gov/index.htm