Context Prediction in the Social Web using Applied Machine Learning: A Study of Canadian Tweeters

Hamman Samuel
Dept. of Computing Science
University of Alberta
Edmonton, Canada
hwsamuel@ualberta.ca

Benyamin Noori
Dept. of Computing Science
University of Alberta
Edmonton, Canada
bnoori@ualberta.ca

Sara Farazi
Dept. of Computing Science
University of Alberta
Edmonton, Canada
farazi@ualberta.ca

Osmar Zaïane
Dept. of Computing Science
University of Alberta
Edmonton, Canada
zaiane@ualberta.ca

***Abstract*—In this ongoing work, we present the Grebe social data aggregation framework for extracting geo-fenced Twitter data for analysis of user engagement in health and wellness topics. Grebe also provides various visualization tools for analyzing temporal and geographical health trends. Grebe currently has over 18 million indexed public tweets, and is the first of its kind for Canadian researchers. The large dataset is used for analyzing three types of contexts: geographical context via prediction of user location using supervised learning, topical context via determining health-related tweets using various learning approaches, and affective context via sentiment analysis of tweets using rule-based methods. For the first, we define user location as the position from which users are posting a tweet and use standard precision metrics for evaluation with promising results for predicting provinces and cities from tweet text. For the second, we use a broader definition of health using the six dimensions of wellness model and evaluate using manually annotated documents with good results using supervised and semi-supervised machine learning. For the third, we use the indexed tweets to show current trends in emotions and opinions and demonstrate trends in polarity and emotions across various Canadian provinces. The combination of these contexts provides useful insights for digital epidemiology. Ultimately, the vision of Grebe is to provide researchers with Canada-specific social web datasets through an open source platform with an accessible RESTful API, and this paper showcases Grebe's potential and presents our progress towards achieving these goals.**

***Index Terms*—location prediction, health social media, sentiment analysis, big data, Twitter**

I. Introduction

The social web is an ideal source of readily available public conversations on a variety of topics. Various platforms such as Twitter, Facebook, and others provide an avenue for users to publicly express their opinions, advice, and questions on topics such as politics, technology, health, among others. Within the context of health and wellness, this public discourse can provide valuable opportunities for tracking and predicting disease outbreaks, as well as measuring user engagement and opinion towards wellness policies [1]. The keywords used on the social web can enhance understanding about potential health symptoms and risks developing over time.

This research is funded by the Alberta Machine Intelligence Institute (Amii), Edmonton, Alberta, Canada.

Essentially, public conversations on the social web can be leveraged for epidemiology, involving analysis of public health patterns for disease prevention and promotion of wellness [2].

In order for social web posts to be useful for digital epidemiology, we identify three contexts that need to be available for analysis: location, domain, and sentiment. Firstly, the position of the user allows researchers to know where to look for health-related issues and epidemics. As an example, a post that mentions "Ebola" may not be useful unless users' location is known. Secondly, only health-related posts are relevant for epidemiology, and including posts from other domains would lead to noisy data. For instance, analyzing posts discussing politics or sports would not assist in detecting outbreaks or gauging public wellness. Thirdly, a user may mention a health-related topic, but positive or negative emotions expressed in a post could provide a better indication about the overall health context of the user making the posting. For example, a user may mention "Ebola" in a positive sense of learning about the disease at a seminar, instead of referring to their personal well-being or potentially contracting the disease.

One challenge for researchers is to associate social web posts with location so that the geographical context of opinions and health concerns could be studied. For instance, most users on Twitter disable their location when tweeting, while their profile location is free-text and can refer to fictitious places such as "Narnia". The limited coverage of public tweets with specific and verifiable coordinates limits the usefulness of social datasets for digital epidemiology. The limitations on data gathering also compound this problem. For instance, researchers have very limited access to the Twitter public dataset due to Twitter's rate limits and data collection policies. Full access to Twitter's dataset requires paid enterprise accounts via Gnip. To our knowledge, there are rarely any substantive datasets that provide Canada-specific geographical context to digital epidemiology researchers, and most prior studies focus on the United States [3], with some studies on Japan [4].

Another challenge is categorization of text-based postings such as tweets. In order to use tweets for epidemiology, they need to be classified as being health-related. However, this is not a trivial task because there are various aspects of health that need to be considered, including physical, intellectual, occupational, spiritual, emotional, and social wellness.

Within each category are multiple keywords and variants that need to be detected in a tweet for it to be considered as health-related. In addition, a user may mention health keywords in their tweet without necessarily referring to themselves or talking about their own well-being.

Moreover, once health-related posts are identified, their polarity and sentiment can help further analyze the full status of users' health. In other words, we can understand how users are feeling about their own health and personal well-being. For instance, tweets could contain various health-related keywords in a positive context such as feeling healthy and well. On the other hand, tweets could also be referring to health problems, ailments, and symptoms, which would be useful for tracking epidemics and outbreaks.

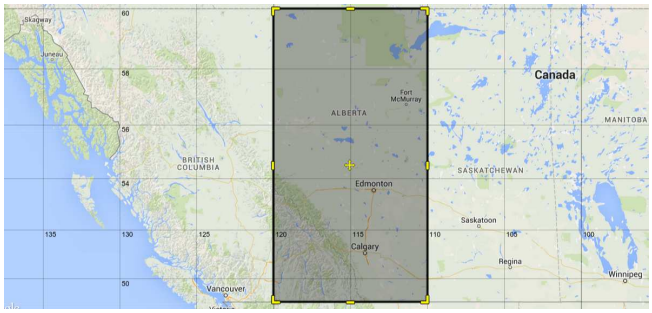
In this ongoing work, we present the Grebe[1] social data aggregation framework for extracting geo-fenced Twitter data, which currently has over 18 million indexed public tweets, and is the first of its kind for Canada-specific social web data for research purposes. Using the large dataset from Grebe, this research work investigates the use of applied machine learning to predict three types of contexts: geographical context via prediction of user location using supervised classification, topical context via determining health-related tweets using various learning approaches, and affective context via sentiment analysis of tweets using rule-based approaches. Moreover, we develop and showcase various visualization tools for analysis of our dataset. The combination of these contexts provides useful insights for digital epidemiology.

II. Methodology

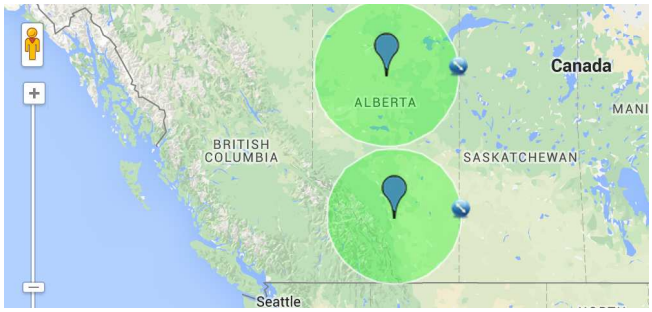
A. Data Gathering

To gather data from Twitter, we use the official and freely available Twitter Streaming and Search RESTful Application Programming Interfaces (APIs) since the inception of the Grebe project in July 2016. The Streaming API[2] provides access to a limited set of randomized realtime tweets, while the standard Search API[3] allows limited searching of historical tweets retroactively.

We ignore retweets, and incoming tweets are filtered by location using API parameters. The streaming API can return tweets made from within a defined bounding box specified using longitude and latitude of a rectangular region[4]. The search API filters tweets within a specified circular geographical region via a radius around a longitude and latitude point[5]. Figure 1 gives a visual illustration of the bounding box and inscribed circle options. Grebe gathers tweets from specific Canadian provinces, hence the bounding boxes and inscribed circles are configured according to the geographical coordinates of the specific regions, with multiple bounding boxes and circles per province where necessary for maximum geographical coverage.



(a) Stream API Bounding Box



(b) Search API Inscribed Circle

Fig. 1. Twitter API Geographical Filtering Options

Grebe is implemented using Python, Flask, and Tweepy[6] on an Infrastructure-as-a-Service (IaaS) cloud platform running Ubuntu, with the Grebe web application and Grebe's RESTful API served via Web Server Gateway Interface (WSGI). The aggregation of tweets is done via cron job while respecting the Twitter rate limits[7]. Grebe is available as an open source Git project via BitBucket for researchers[8].

Ultimately, public tweets retrieved from the Twitter API are indexed and stored in a MariaDB SQL database, with the indexed fields shown in Figure 2. We denormalize the tweets and users entities in order to capture snapshots of user information at the time of tweeting. We also investigated NoSQL databases such as MongoDB, but found the disk space and performance metrics for MariaDB were optimal. MongoDB requires significantly more disk space, while performance for Create, Read, Update, Delete (CRUD) operations is similar between MongoDB and MariaDB.

For optimizing **C**reate operations, a hash of each tweet was stored and checked before new tweets were saved. To optimize **R**ead operations, table indexes for the `tweet`, `tweet_hash`, `created_at`, and `place_name` fields were used. The system does not carry out any **U**pdate or **D**elete operations because saved tweets do not need to be updated or deleted.

[1]Grebe Live Demo http://199.116.235.207/grebe

[2]Twitter Streaming API Developer Documentation https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample

[3]Twitter Standard Search API Developer Documentation https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

[4]BoundingBox tool to configure rectangular geo-fence's edge coordinates and dimensions http://boundingbox.klokantech.com/

[5]FreeMapTools used to configure inscribed circle's centre and radii https://www.freemaptools.com/radius-around-point.htm

[6]Tweepy Library http://www.tweepy.org

[7]Twitter Limits https://developer.twitter.com/en/docs/basics/rate-limiting

[8]Grebe BitBucket Repository https://bitbucket.org/hwsamuel/grebe

TWEETS		
PK	ID	INT(11)
	TWEET	VARCHAR(255)
	TWEET_HASH	VARCHAR(255)
	LONGITUDE	FLOAT
	LATITUDE	FLOAT
	CREATED_AT	DATETIME
	COLLECTED_AT	DATETIME
	LANG	VARCHAR(10)
	PLACE_NAME	VARCHAR(255)
	COUNTRY_CODE	VARCHAR(5)
	USER_ID	VARCHAR(255)
	USER_NAME	VARCHAR(20)
	USER_GEOENABLED	TINYINT(1)
	USER_LANG	VARCHAR(10)
	USER_LOCATION	VARCHAR(255)
	USER_TIMEZONE	VARCHAR(100)
	USER_VERIFIED	TINYINT(1)

Fig. 2. Public Tweet Information Indexed by Grebe

B. Tweet Location Prediction

We define the location of a tweet as the geographical position from which a tweet was made. Grebe also aggregates tweets with missing longitude and latitude information. From our collected dataset of over 18 million tweets, 14% of the tweets contained verifiable location information. This is a general trend, as few users enable geo-tagging when tweeting [5]. In our case, this was also because the Twitter API often returns tweets using an estimated location based on the free-text user profile location. Hence, while the Twitter API categorizes these tweets from a specific region, the tweets' actual longitude and latitude is missing. Nevertheless, these tweets are useful for expanding our sample size. Moreover, this dataset of unmarked tweets, along with tweets with verified location, is useful for investigating whether a tweet's location could be predicted without geo-coordinates.

Other studies have attempted to identify granular tweet location from text using word collocations [6]. For tweet location prediction, we use a supervised classifier to predict the Canadian province from where the tweet was posted. We also investigate city-level predictions. For the features of our classifer, we use tweet *n*-grams. Tweets are tokenized and converted to *n*-grams. All possible combinations of adjacent words of length *n* within a posting are referred to as *n*-grams. For example, a tweet containing words $[w_1, w_2, ..., w_n]$, would generate the list of bigrams as $[w_1w_2, w_1w_3, ..., w_{n-1}w_n]$. We use a combination of uni-, bi- and tri-grams as features.

Our class labels correspond to the city names and postal abbreviations of provinces we have indexed, for example `AB`, `ON`, `SK`, `BC`, `MB`, and `QC`. For evaluation, we use 10-fold cross validation with 80-20% split between training and holdout data respectively at each iteration. We explore four classification approaches: Support Vector Machines (SVM), Multi-Layer Perceptron (MLP) neural network, Latent Dirichlet Allocation (LDA), and Random Forest (RF) [7].

C. Classifying Health-Related Tweets

In order to determine whether a tweet is referring to the tweeter's personal health status, we use the six dimensions of wellness model [8], which broadens the meaning of health as not merely the absence of disease or infirmity. In essence, health can be defined as a combination of physical, intellectual, occupational, spiritual, emotional, and social wellness.

Physical wellness covers physical activity, healthy eating, use of appropriate drugs and supplements to avoid stress, fatigue, and diseases. Intellectual wellness covers lifelong learning, acquisition of skills, and self-education. Occupational wellness covers use of personal talents and skills to perform paid professional work or unpaid volunteering. Spiritual wellness covers the pursuit of peace and harmony through a value system. Emotional wellness covers mental and psychological stability, enablement of positivity, avoidance of negativity, coping with life challenges, and expressing feelings. Finally, social wellness covers personal and communal relationships with friends or strangers, generally anyone we interact with [8].

The problem of classifying tweets as health-related can be formulated as follows. Firstly, how can we classify tweets as being related to health? And secondly, if a tweet is related to health, which dimension of health is it related to? We explore four approaches for a potential optimal solution: keyword search (KS), document search (DS), supervised learning (SUP), and semi-supervised learning (SSUP).

Keyword search is simplistic solution, where a list of keywords is created, each related to one dimension of wellness. If a tweet contains any of the keywords in these lists, we consider it as being related to the dimension the list corresponds to. A tweet can be related to multiple dimensions of health with this approach. If a tweet has more than one of the keywords associated with a dimension, we assume a stronger relationship between the tweet and that dimension of health. For this approach, we manually curated six lists of keywords per wellness dimension by reading documents describing the dimensions. Tweets were then categorized based on keywords.

With the document search method, we approach this problem from the point of view of information retrieval. Firstly, we curate a dataset of ten documents for each dimension of health. Hence, we have a total of sixty grouped documents describing the various dimensions of health. Secondly, every tweet is seen as a query to this database of documents. Based on various measures of similarity, we can make a decision about the label of a tweet.

We query this dataset with each tweet and fetch the most similar documents. Next, we classify each tweet as being related to a health dimension if there is at least one document with a similarity score higher than a pre-defined threshold. The tweet is then classified to the label of its most similar document. We use two measures of similarity: cosine similarity and set containment. In our experiments, we use a threshold of 0.2 to determine whether a document is related to any of the health dimensions.

For cosine similarity, we transform both the document and the query, in our case the tweet, into a vector representation. This vector has *n* items, with *n* being the number of distinct terms in the query and the document. The value of each item in the vector shows the number of times that term has appeared in the query or the document. Using these vectorized formats, we define the similarity of the document and the query as the cosine of the angle between the two vectors. For set containment, we consider both the document and the query as a set of terms, and define similarity based on common keywords as $\frac{|D \cap Q|}{|Q|}$, where D are the document keywords and Q are the tweet keywords. We use manually labeled tweets for evaluation of the results.

In the supervised learning method, we use our manually labeled set of tweets to train a naïve Bayes binary classifier for each health dimension. For every data point, we use tweet *n*-grams as features for training. To evaluate this method, we use 5-fold cross validation. At every iteration, we set aside 20% of the data as holdout, and train our classifier with the remaining 80%. Accuracy is calculated from the holdout set.

Other studies have proposed a semi-supervised binary approach to label a stream of incoming tweets [9]. For this approach, we curate two sets of keywords for the six dimensions of wellness. The first set contains keywords specific to each dimension, while the second set contains general health keywords. If a tweet includes keywords in the first set, they are labeled with the corresponding health dimension. On the other hand, if a tweet has none of the keywords from the first set, but at least one from the second set of general health terms, it is marked as possibly being related to health and set aside. If a tweet has no match in both lists, it is not health-related.

Using these initial labels, we train a naïve Bayes binary classifier for each health dimension. We then use our trained model to label the tweets that were set aside. There is now a larger set of labeled data available for training that reveals new collocated keywords, and the newly labeled tweets can be included in our training dataset. We then use our expanded dataset to train a better, hopefully more accurate classifier. By repeating this process as more tweets are received, the classifier can improve iteratively. Accuracy is calculated every time we complete processing a batch of tweets by evaluation on an evenly split set of labeled samples.

D. Tweet Sentiment Analysis

Sentiment analysis enables evaluation of the emotions expressed in text. There are two potential objectives: continuous metrics or discrete labels. For the former, the polarity of a given text is computed to give an indication about positive, negative, or neutral emotions and the degree or strength of the sentiment. For the latter, emotion-based labels are assigned to the text, such as the eight human emotions of *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust* [10]. For our research, we apply rule-based approaches to determining polarity of the health-related tweets, and also investigate multi-label emotion mining.

To determine the degree of polarity of the health tweets, we use rule-based lexicon approaches using the Liu & Hu[9] lexicon [11] and the VADER (Valence Aware Dictionary and sEntiment Reasoner)[10] lexicon for social media datasets [12]. The rule-based approach was initially applied to customer review datasets, while the VADER tool is specifically targeted towards Twitter datasets. Both lexicons contain positive, negative and neutral words, and the frequency of words in-text is used to compute polarity within the range $[-1, 1]$.

Multi-label emotion mining from text is an interesting area that few studies have explored previously [13]. Since human emotions often tend to co-occur, we apply multi-labels to health tweets using the National Research Council (NRC) Canada's Word-Emotion Association Lexicon (EmoLex) [14], which uses labels from psychology literature, specifically Plutchik's wheel of eight human emotions [10]. EmoLex contains multiple labels associated with words, hence a simplistic rule-based approach is applied. Given a tweet with tokenized words as $[w_1, w_2, ...w_n]$, the associated labels for each word are determined, $L(w_i) = [l_1, l_2, ...l_m]$. Hence, the tweet's emotion multi-labels are all labels with aggregated frequency above a given threshold.

Additionally, a self-reference filter is applied by detecting use of personal pronouns within health tweets. With this filter, the final dataset contains tweets mentioning personal health issues rather than user commentaries on general or public health topics. This simplistic heuristic enables more focused digital epidemiology because users' personal health condition might be more useful for predicting outbreaks.

III. Results

A. Grebe Showcase

Table I provides statistics on the present size of the dataset collected by Grebe from July 2016 till date. It should be noted that aggregation for some provinces was started later. Also, various provinces are not presently being indexed, but there are future plans to expand Grebe's coverage. There are some non-Canadian tweets captured from the neighbouring United States due to the approximate nature of some of the bounding boxes and inscribed circles used in geo-fencing.

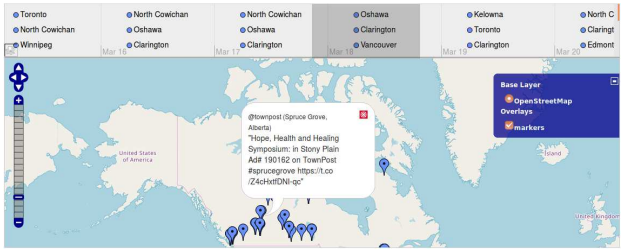
TABLE I
GREBE DATASET STATISTICS

Statistics	Total Amount	Quick Glance
All Tweets	18,250,853	+18M
Tweets with Coordinates	2,555,973	+2M
Tweets from Alberta (AB)	336,228	+300k
Tweets from Ontario (ON)	552,428	+500k
Tweets from Saskatchewan (SK)	56,118	+50k
Tweets from British Columbia (BC)	319,185	+300k
Tweets from Manitoba (MB)	76,026	+70k
Tweets from Quebec (QC)	102,379	+100k

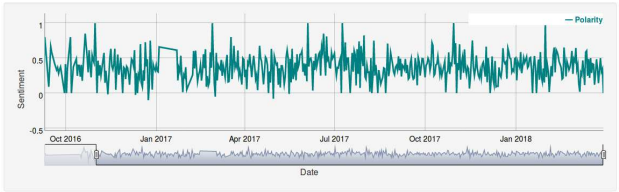
[9]NLTK's Sentiment Analyzer contains Liu & Hu's lexicon http://www.nltk.org/api/nltk.sentiment.html#nltk.sentiment.util.demo_liu_hu_lexicon

[10]VADER sentiment analysis tool is open source and code is available via GitHub https://github.com/cjhutto/vaderSentiment

The vision of Grebe is to provide Canadian and worldwide researchers geo-fenced social data specific to Canada. Grebe consists of three main sub-systems: data aggregator, RESTful API, and tools. Firstly, the data aggregator presently focuses on Twitter but future plans include adding other social web sources. All aggregated data is geo-fenced so that longitude-latitude pairs are partially available. In addition, our preliminary results for location prediction are promising. Secondly, the data stored in Grebe is accessible via a RESTful API[11] as JSON output, with various filtering and querying options. Researchers can use the data on request to develop tools and perform analysis, such as classification of health tweets or sentiment analysis. Thirdly, Grebe provides two visualization tools to assist with analysis: time map and trend graph. The time map enables visualization of data on a map, along with a temporal overview of data variations. The trend graph demonstrates summarization of statistics over time, such as top keywords and hash tags being tweeted, or sentiment polarity. Figure 3 shows the time map and trend graph visualizations.



(a) Time Map



(b) Trend Graph

Fig. 3. Grebe Visualizations Showcase

Grebe currently provides hash tag search that can allow filtering of data being shown on the visualization tools. The most frequently used general hashtags are also recommended for filtering. Another feature under development and testing is a keyword recommender that can suggest health keywords to use for filtering. We use an inverted index of all keywords within tweets, and suggest the top-n medical keywords from the past m months to prevent outdated keywords from appearing, and increase serendipity.

General health keywords are identified by their occurrence in the Systematized Nomenclature of Medicine (SNOMED), a digital collection of medical terms provided by the U.S. National Library of Medicine [15]. In order to properly identify layperson health terms, the Consumer Health Vocabulary (CHV) mapping is also used [16].

[11]Grebe API documentation http://199.116.235.207/static/api_docs.pdf

B. Location Prediction

For province predictions, we balanced the class labels by selecting 50,000 tweets per province based on the minimum number of indexed tweets from Saskatchewan, giving a total of 300,000 tweets used with 80-20% training-holdout split.

For city predictions, out of 3,843 Canadian cities indexed in Grebe, 19 cities with over 10,000 tweets were selected, with class labels balanced at 10,000 tweets per label, giving a total of 190,000 tweets in the training and holdout datasets. The accuracy metrics for province and city predictions for our classifiers are shown in Figure 4.

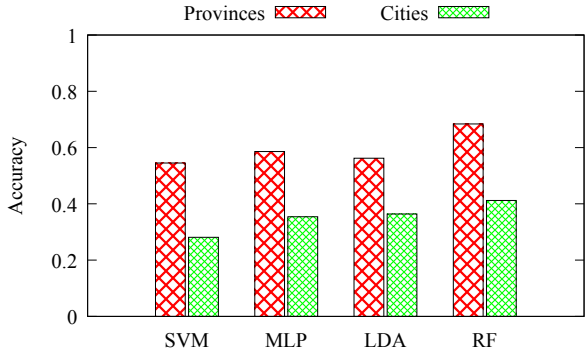


Fig. 4. Prediction of Canadian Provinces and Cities

For our MLP neural network, we used the logistic sigmoid activation function, and Adam solver with adaptive learning. We also explored a deep neural network approach with up to five hidden layers, but the overall performance decreased. For our SVM classifier, we used the sigmoid kernel, while for LDA we obtained the best performance with a learning decay set at 0.65. Overall, the Random Forest classifier recorded the best performance with 68.43% accuracy in predicting provinces and 41.20% prediction accuracy for cities. The low predictability of cities could be attributed to the large size of class labels. We also evaluated accuracy@k, where k represents the top-k labels predicted. For accuracy@3, provinces are predicted with 78.23% accuracy, while cities can be predicted with 53.32% accuracy using the Random Forest classifier.

C. Health Classification

We used 118,363 tweets from the province of Alberta to evaluate our health-related tweet classification strategies. For our evaluation dataset, we manually labeled 100 tweets related to each of the six dimensions of wellness, resulting in 600 tweets. We also labeled 221 tweets as not being related to any of the dimensions of health. Overall, we manually labeled 821 tweets with 221 negative and 600 positive examples.

The process of manual labeling was done by reading through the list of available tweets sequentially, to a point where we had enough labeled samples for physical and emotional dimensions of wellness. Next, we used keyword search to find possibly relevant tweets for other dimensions with sparse tweets. A summary of the accuracy for different classification strategies is shown in Figure 5.

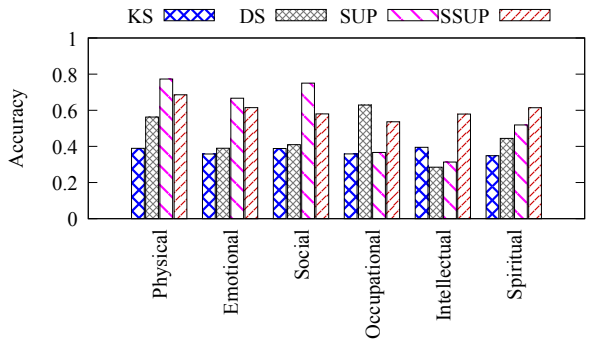


Fig. 5. Identifying Health-Related Tweets by Wellness Dimension

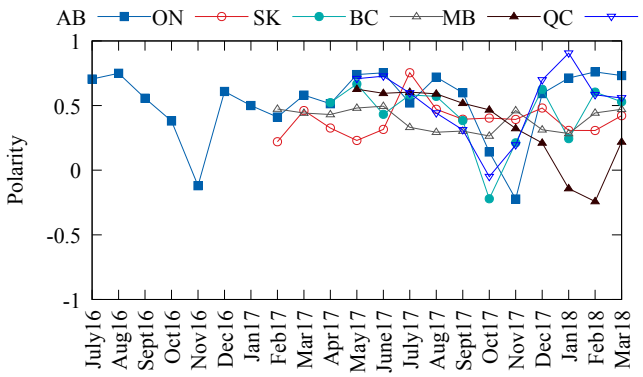


Fig. 6. Average Monthly Polarity of Health Tweets from Canadian Provinces

Supervised learning provided best results for the physical, emotional and social wellness labels, while the semi-supervised learning approach provided best results on average for the occupational, intellectual, and spiritual wellness labels. Generally, the performance of our classifiers was lower for intellectual, spiritual and occupational dimensions of wellness across all methods. This can be attributed to the fact that these dimensions are more difficult to define and to find keywords for. The semi-supervised approach gave better performance for some categories because finding descriptive and well-defined documents for these dimensions proved to be more difficult. Hence, supervised methods could not readily detect patterns and correlations within these documents and required semi-supervised human moderation. Generally, there was less agreement over the label of tweets in these categories compared to other dimensions. For the document search approach, cosine similarity outperformed set containment, while for semi-supervised learning, accuracy increased proportionally with the iteratively increasing size of the training set. The keyword search approach performed poorly in general.

D. Sentiment Analysis

Firstly, we analyzed trends in polarity across provinces by aggregating average polarity scores per month using both the Liu & Hu lexicon and the VADER lexicon for the classified health-related tweets. An overview of polarity of various Canadian provinces over the period of July 2016 to March 2018 is shown in Figure 6. Data collection for some provinces was started later than July 2016. There is some suggestion from the Alberta statistics that users tended to be more negative seasonally around November, but more data is needed to confirm any patterns.

Secondly, we investigated emotion multi-labels of health tweets from the province of Alberta over time using the NRC-Canada's EmoLex lexicon. The frequency of emotion labels is summarized in Figure 7. For Alberta, *joy* was the most frequent emotion expressed, while the *sadness-anger* labels occurred frequently together.

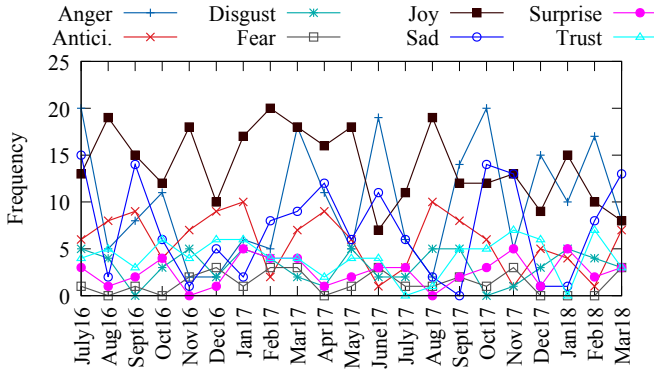


Fig. 7. Emotion Multi-Labels of Health-Related Tweets from Alberta

IV. Discussion

Only publicly available tweets are indexed by Grebe to respect users' right to privacy [17]. Moreover, minimal personal information about users is collected, as the focus is on analysis of textual conversations on the social web rather than user profiling. Technical considerations for Grebe include up-time of the RESTful API, stoppages affecting the cloud server architecture, Twitter rate limits, and cron job errors. The RESTful API is available on-demand as long as the Ubuntu cloud virtual machine does not experience outages from the service provider.

Grebe respects Twitter's rate limits for retrieving tweet streams by suspending our cron jobs when the rate limit is reached, and resuming after the timeout periods recommended by Twitter. There have been instances when the cron jobs do not resume correctly, and this has led to some missing data for certain periods. Over time, with continuous testing, our Python scripts and cron jobs have gotten more stable, leading to a higher volume of data collected.

Limitations of our system include sample size, collection period, and health tweets. Firstly, our sample size, though large, is still a small representative of the overall population. This is a general limitation for social web data because not everyone in the population uses social networks.

Secondly, our data covers just under two years, and has not reached full maturity for identifying patterns and predicting future incidents. Grebe has the potential to provide these services in the near future, as our dataset is continuously expanding. Thirdly, the size of users who talk about their health online is a small representative of the overall population. More research and data is needed to determine whether this is a representative sample.

For future work, we intend to improve accuracy of our predictors, gather more tweets, and add more social web sources. Firstly, we have demonstrated that the contexts of geo-location, health, and sentiment can be predicted and analyzed. We are currently working on further improving the accuracy of these predictors, and also evaluating our sentiment analysis labels. Secondly, we intend to keep expanding our dataset by collecting more tweets, including tweets from timelines for users with verified location. For example, if a user tweets often from Edmonton, Alberta, there is a high likelihood that this user's future tweets will also be from this location, even if they are not geo-tagged. Thirdly, we plan on exploring social web sources such as Facebook, Snapchat, Instagram, and others. Ultimately, Grebe is useful for providing new data sources for existing digital epidemiology tools such as ARTSSN [18], where Grebe's datasets can provide additional insights to public health via correlations with emergency ward data sources typically used in epidemiology tools.

V. Literature Review

There have been several projects in the past on monitoring, classification and analysis of tweets with health-related information. In [19], a method to detect influenza epidemics through Twitter data is proposed using an SVM classifier to detect influenza-related tweets. However, their study focused on datasets from the Japanese Infection Disease Surveillance Center (IDSC).

The ChatterGrabber system has been proposed for publicly available health social web surveillance [3], which is able to collect and categorize a high volume tweets for syndromic surveillance. ChatterGrabber uses supervised learning, which requires manual annotation of relevant tweets. Our semi-supervised learning approach does not require manual curation of health-related tweets. ChatterGrabber is also presently not available online and does not provide access to its datasets. Moreover, ChatterGrabber only collects queried hits specific to Gastrointestinal Illness (GI), and prior knowledge of keywords to use is required. However, for new disease outbreaks and trends, keywords regarding symptoms may not be known ahead of time, and a broader collection of data is required, which Grebe provides.

Other research works have looked at visualizations for disease tracking, including usage of Google Trends for disease tracking [20]. Part of our interface is based on Google Trends, and is applicable to disease tracking. However, our workflow uses publicly available social web data, while Google Trends statistics are based on Google's internal web search query logs.

We are also working on extending search features available in typical trend graphs by recommending synonyms and related words to users based on their query. In addition, the interface shows recommendations for keywords to search based on frequency statistics.

While other visualizations have also incorporated stand-alone maps, such as [1] and [18], having both a map and timeline is a relatively recent application of the time map interface metaphor regarding digital epidemiology. This dual visualization available in Grebe could be very useful for detecting geographical trends over time and has not been explored in other epidemiology-based studies.

VI. Conclusion

In this ongoing work, we showcased the Grebe social data aggregator that has been used for extracting geo-fenced Twitter data and for preliminary analysis of user engagement in health and wellness topics via visualization and prediction tools. Grebe is an open source, on-request system that has over 18 million indexed public tweets from Canadian provinces, and its vision is to provide social web data for Canadian and worldwide researchers in computing science, machine learning, social sciences, among others. We provide an on-demand RESTful API for access to our datasets, and also have developed various visualization tools to help sift through the data. In this research work, we used our dataset and applied state-of-the-art machine learning to analyze three tweet contexts: geo-location, health-relation, and sentiment. We demonstrated promising results for predicting a tweet's province and city using supervised learning, even when it is not geo-tagged. We also showed that health-related tweets can be identified and used for further analysis such as determining tweet polarity and emotion labels. The combination of these contexts has the potential to be useful for digital epidemiology. For future work, we intend to keep expanding our dataset by including additional social web data sources such as known user profiles from Canada-specific locations and querying of tweets using geo-specific keywords/hashtags. We also plan on improving our visualization tools and enhancing our predictive machine learning methods by development of gold and silver standard datasets for supervised and semi-supervised machine learning and evaluation.

Acknowledgment

We wish to thank the Alberta Machine Intelligence Institute (Amii), formerly known as the Alberta Innovates Centre for Machine Learning (AICML), for funding and supporting this research. Amii is a research lab at the University of Alberta working to enhance understanding and innovation in a number of subfields of machine intelligence. We would also like to thank Cybera for providing Grebe's cloud hosting platform. Cybera is a not-for-profit technical agency that is helping Alberta advance its IT frontiers.

References

[1] Brennan, S., Sadilek, A., Kautz, H.: Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), AAAI Press (2013) 2783–2789

[2] Salathé, M.: Digital Epidemiology: What Is It, And Where Is It Going? Life Sciences, Society and Policy **14**(1) (2018) 1

[3] Schlitt, J.T., Lewis, B., Eubank, S.: ChatterGrabber: A Lightweight Easy to Use Social Media Surveillance Toolkit. Online Journal of Public Health Informatics (OJPHI) **7**(1) (2015)

[4] Zaraket, H., Saito, R.: Japanese Surveillance Systems and Treatment for Influenza. Current Treatment Options in Infectious Diseases **8**(4) (Dec 2016) 311–328

[5] Sloan, L., Morgan, J.: Who Tweets with their Location? Understanding the Relationship Between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. PloS One **10**(11) (2015)

[6] Han, B., Cook, P., Baldwin, T.: Text-Based Twitter User Geolocation Prediction. Journal of Artificial Intelligence Research **49** (2014) 451–500

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: SciKit-Learn: Machine Learning in Python. Journal of Machine Learning Research **12**(Oct) (2011) 2825–2830

[8] Strout, K.A., Howard, E.P.: The Six Dimensions of Wellness and Cognition in Aging Adults. Journal of Holistic Nursing **30**(3) (2012) 195–204

[9] Ali, A., Magdy, W., Vogel, S.: A Tool for Monitoring and Analyzing Healthcare Tweets. In: Proceedings of the ACM SIGIR Workshop on Health Search & Discovery. Volume 28. (2013) 23

[10] Plutchik, R.: The Nature Of Emotions: Human Emotions Have Deep Evolutionary Roots, A Fact That May Explain Their Complexity And Provide Tools For Clinical Practice. American scientist **89**(4) (2001) 344–350

[11] Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2004) 168–177

[12] Gilbert, C.H.E.: VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: Eighth International Conference on Weblogs and Social Media. (2014)

[13] Yadollahi, A., Shahraki, A.G., Zaïane, O.R.: Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Computing Surveys **50**(2) (2017) 25

[14] Mohammad, S.M., Kiritchenko, S., Zhu, X.: NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In: Proceedings of the 7th International Workshop on Semantic Evaluation Exercises. (June 2013)

[15] Cornet, R., de Keizer, N.: Forty Years of SNOMED: A Literature Review. BMC Medical Informatics and Decision Making **8**(1) (2008) S2

[16] Keselman, A., Smith, C.A., Divita, G., Kim, H., Browne, A.C., Leroy, G., Zeng-Treitler, Q.: Consumer Health Concepts That Do Not Map To The UMLS: Where Do They Fit? Journal of the American Medical Informatics Association **15**(4) (2008) 496–505

[17] Humphreys, L., Gill, P., Krishnamurthy, B.: How Much Is Too Much? Privacy Issues On Twitter. In: Conference of International Communication Association. (2010)

[18] Smetanin, P., Biel, R.K., Stiff, D., McNeil, D., Svenson, L., Usman, H.R., Meurer, D.P., Huang, J., Nardelli, V., Sikora, C., et al.: An Early Warning Influenza Model using Alberta Real-Time Syndromic Data (ARTSSN). Online Journal of Public Health Informatics **7**(1) (2015)

[19] Aramaki, E., Maskawa, S., Morita, M.: Twitter Catches the Flu: Detecting Influenza Epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 1568–1576

[20] Carneiro, H.A., Mylonakis, E.: Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. Clinical infectious diseases **49**(10) (2009) 1557–1564