This is your Twitter on drugs. Any questions?

Cody Buntain
Dept. of Computer Science
University of Maryland
College Park, MD 20742
cbuntain@cs.umd.edu

Jennifer Golbeck
College of Information Studies
University of Maryland
College Park, MD 20742
jgolbeck@umd.edu

ABSTRACT

Twitter can be a rich source of information when one wants to monitor trends related to a given topic. In this paper, we look at how tweets can augment a public health program that studies emerging patterns of illicit drug use. We describe the architecture necessary to collect vast numbers of tweets over time based on a large number of search terms and the challenges that come with finding relevant information in the collected tweets. We then show several examples of early analysis we have done on this data, examining temporal and geospatial trends.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; K.4.0 [**Computers and Society**]: General

Keywords

social media, social networks, Twitter, drug use, drug trends

1. INTRODUCTION

Because it is large, free, open, and largely public, Twitter has become a popular social media source to study information about nearly any event happening in society. Used by governments and terrorist groups, celebrities and small town teens, it is a place where literally anyone can say literally anything (as long as it's under 140 characters). As such, it has become an excellent platform to detect and analyze trends. Whether over time, location, or a combination of the two, new research is showing data from Twitter can enhance our ability to understand everything from earthquakes to epidemics to music preferences.

In this project, we are leveraging Twitter to detect patterns in illicit drug use. Given the significant delay in publishing large-scale drug usage surveys (on the order of years) and especially with new designer and synthetic drugs, it is important to gather real-time information on how they emerge in communities, how awareness of them spreads, and how widely they are used. To this end, we have developed an architecture that collects data from Twitter on a large set of drug names and slang terms. The data is then filtered and fed into tools that analyze frequency, temporal, and geospatial patterns in the data.

In this paper, we present an overview of our system in which we describe the architecture, the implementation, and the challenges that come with trying to build a comprehensive picture of drug use when there is often great ambiguity surrounding the most common slang terms. We present examples of the insights we have been able to gather in this initial phase of the research and conclude with challenges for future work in the space of public health and Twitter analysis.

MSM'15, WWW 2015 May 18–22, 2015, Florence, Italy.

2. RELATED WORK

Tracking patterns of activity on Twitter, by time or location, is a growing and diverse area of research. New techniques, algorithms, and impactful applications are emerging every year. Notably, social network sites like Twitter provide an invaluable vehicle for research into social contagion (the spread of ideas or information through social networks) and large-scale user behavior [10]. Researchers out of Indiana have shown correlations between sentiment and mood of the Twittersphere and fluctuations in major stock markets [3]. These results have led the financial industry to leverage social networking sites like Twitter in support of algorithms for buying and selling stocks and other assets. Similarly, in 2010, Sakaki et al. demonstrated how Twitter could be used as an early warning system for earthquakes in Japan as Twitter users shared news of an earthquake on Twitter almost instantaneously, and this information propagates faster through radio waves and wires than earthquakes propagate through the Earth's crust [15]. These researchers were even able to calculate the epicenter of an earthquake with decent accuracy by analyzing locations shared in users' tweets. Building on this work, Crooks et al. used Twitter as a sensor system for expanding earthquake detection[6]. The authors argue that user responses to earthquake events can help localize and identify them. These investigations into user behavior and response in Twitter demonstrate the fundamental feasibility for tracking real-world events through social networks on which our work relies.

Integrating geospatial information with these behavioral patterns has yielded a wide range of application areas as well. Researchers in [5] used geospatial information from tweets to better understand the Occupy Wall Street Movement, including where attention to tweets was coming from

(e.g., within the areas protests were happening or outside). Other work leveraged Twitter to understand users' musical preferences [8]. The geospatial information from their tweets was combined with the information extracted from the tweet content to create a browsable "world of music". Geolocation was similarly used to create a map of world languages in [14]. After automatically detecting the language of tweets, the geocoded information was used to map location in which each language was used. Researchers in [13] showed that geotagged tweets could be used to find breaking news in near-real time for specific locations.

Perhaps most related to our research though is the study by Lampos and Cristianini from the University of Bristol on using Twitter to track the spread of flu in the United Kingdom (UK)[9]. Lampos and Cristianini showed that, by tracking geolocated tweets containing specific markers related to influenza-like illnesses, they were able to calculate "flu-scores" for five regions of the UK. The authors then showed how these flu-scores exhibit high correlation with independently gathered health data from the UK's Health Protection Agency. Given the similarities between the notions of viral epidemics and epidemics of drug usage, this research provides significant support for the feasibility of our research.

3. MOTIVATIONS

Despite the controversies around illicit drug use and whether it constitutes a matter of public health or a matter of criminality, widespread agreement exists on the importance of tracking and understanding this drug use. Procedures and technologies for measuring drug use are currently far behind other public health reporting mechanisms. While communities are incentivized to report patterns of infectious disease promptly, the significant stigma surrounding drug use and the socioeconomic complexities of how drug use spreads have contributed to large gaps between when data is collected and when it is reported.

A prime example of this lag can be seen in the most recent version of the "National Survey on Drug Use and Health" (NSDUH), a document published jointly by the US Department of Health and Human Services and Substance Abuse and Mental Health Services Administration (SAMHSA) [1]. This survey was last released in 2014 and covers use of tobacco, alcohol, illicit drugs, and mental health during 2012-2013. At the same time, the NSDUH does not differentiate between many different types of drugs and instead collapse many types into large classes. While this approach is reasonable given the variety of drugs available, it provides little insight into how specific types of drugs or new designer drugs are spreading. As the primary source of information on drug use in the United States, researchers or policy makers are forced to use information that is at least a year old, and given the high demand and monetary value of the illegal drug trade, changes in this landscape likely outpace this survey data rapidly. While the US government is taking steps to address these deficiencies with rapid reporting programs like the Community Drug Early Warning System (CDEWS) and its successor, the National Drug Early Warning System (NDEWS), these frameworks are in their infancy [19].

The work detailed in this paper sought to augment these existing resources by leveraging the massive volumes of data regular people routinely share via social media. Since existing work already validates the idea that changes in behavior can be tracked across geographic regions by processing data from sites like Twitter, a natural extension is to apply this same idea to track changes in use or popularity of various drugs on Twitter. Such an approach would yield results much more rapidly than procedures like the NSDUH while also augmenting such resources with additional information.

To this end, we present a first step toward establishing a functioning architecture for the collection, filtering, and analysis of drug-related messages, or tweets, on Twitter. To bootstrap this project, a group of subject matter experts provided a list of 21 different drugs and 300 terms for each. Some are technical (e.g., buprenorphine, an opioid pain killer), but most are slang (e.g., "bud", "herb", and "yerba" among the many terms for marijuana).

We searched for each of these terms (described further in the next section) as a first step toward understanding when, how, and where each drug is used.

4. DATA COLLECTION AND ANALYSIS

The shear volume and velocity of social networking sites like Twitter and Facebook comprise their most powerful assets but simultaneously pose a significant technical challenge. Twitter sees nearly half a million tweets per minute, and while a tweet is limited to 140 characters, metadata with each tweet (e.g., tweet location, links to other media, or user mentions) represents a sizable chunk of data, with the tweet's content accounting for only about 4% of the total data. This volume works out to an average of just over one gigabyte of data generated *per minute*.

Realizing the commercial value of this data, Twitter has monetized access to its full stream of tweets through data resellers like Gnip [1], with customers using this data for applications like real-time financial trading and business intelligence. Since access through these resellers can be prohibitively expensive for many nascent research projects, Twitter also provides a publicly available stream containing a 1% sample of the full stream of tweets. While 1% may seem overly restrictive, this down-sampled stream still yields an average of 3,600 tweets per minute. Along with this public sample stream, Twitter also provides a mechanism for tracking tweets containing specific keywords, authored by specific users, written in a specific language, or published from specific locations as part of their public interface. This filtered stream is designed to return all tweets matching the given specifications up to a maximum of 1% of the total stream (that is, at most 1% of the stream can be obtained through this filtering). For the experiments outlined herein, we leverage both of these data sources: the 1% sampled stream for undirected data gathering, and the filtered stream with prespecified, drug-related keywords for targeted collection.

Collecting tweets from these sources is relatively straightforward given the plethora of tools available. The first step, however, is to acquire an access token through Twitter's developer site[2]. With that token in hand, we then leverage Twitter4j[3], a popular Java library that wraps Twitter's developer API, that provides both sample and filter stream access. For the sample stream, no configuration is needed as it merely collects tweets sent from Twitter's servers. For the filtered stream, on the other hand, one must provide a

[1] `http://gnip.com`
[2] `http://dev.twitter.com`
[3] `http://twitter4j.org`

list of keywords, locations, languages, or authors to track. Since our interest is in drug-related tweets, we obtained a keyword list for eleven drug types, twenty-one drugs, and approximately 300 slang terms, which we used to record any tweet containing at least one of these keywords [19]. All acquired tweets were then stored in a cluster-based file system for distributed processing later.

Our collection of tweets from the 1% sample stream covered April through July of 2014 for a total of 645,761,955 tweets (247GB compressed). Tweets from the filtered stream covered October 30 to November 26, 2014, containing 176,742,962 records (81GB compressed).

As can be seen from the sizes of our data collections, analyzing Twitter streams, even after down-sampling to 1% of the total volume, falls squarely in the realm of "Big Data" with data sizes beyond the analysis capabilities of a typical personal computer. Fortunately, the analytics industry has tackled this Big Data problem head-on with the development of several distributed processing tools based on the popular MapReduce paradigm. Specifically, we make use of the Apache projects Hadoop, Pig, and Spark, which leverage the Hadoop file system (HDFS) for distributed processing. In addition, Twitter maintains a set of libraries, called ElephantBird, for processing tweets on these platforms. Using these platforms, we are able to extract tweets conforming to specific criteria (e.g., selecting tweets for a certain keyword or filtering out tweets with no geolocation data).

Ambiguity is a major challenge of this process. Many of the terms we used, especially slang terms, have other popular meanings. Ecstasy slang includes terms like "candy", "eve", "malcom", "molly", "peace", "skittles", and "smartees", all words commonly used in non-drug contexts. It is not just one drug that has this problem. Other slang for different drugs include the terms "boxes", "house", "horse", "grass", "glass", and "pink". In some cases, the term ends up being used primarily in the drug context (e.g., "weed"). In others (e.g., "skittles"), it was almost impossible to find drug-related tweets in a hand analysis of the data.

However, for many terms, there is a mix of relevant and irrelevant content for this application, and some automated techniques can be useful to sort tweets by relevance. One tool with which we have had success in our early analysis is with the topic modeling package Mallet [12]. For example, the term "oxy" has some ambiguity online. While it is slang for oxycontin (which is itself an ambiguous term since it can be used as a brand name or another term for oxycodone), it is also the name of a relatively popular YouTube gaming channel, Team OXY, and the stock symbol for Occidental Petroleum. We found that Mallet was able to easily distinguish drug-related topics from others, even with a small number of topics. Below is a list of terms for several topics. Note that for topic 1, there are almost no drug terms and many related to the gaming content posted on Team OXY's YouTube channel. The other two topics, especially the third, are clearly drug related.

1. oxy youtube video ice redman prince mac tah trailer oxygen oxymoron health make months gotas tonight gabapertina sweet lumina
2. i'm pain morons real people reel credits supply system pills man treatment kinda feeling oxycontin strange amino ubos feelin
3. amp bars it's xanax drugs oxy state day oxycodone prescription speed morphine buy percs percocets ecstasy adderal purplethizzle relaxers muscle

Clearly, much more work is necessary to understand how topic modeling – and other techniques – could be integrated into the workflow of our system. However, preliminary results like this show that there is potential for such tools to deal with part of the major ambiguity issues that arise in projects like ours.

5. CASE STUDIES

This section describes a series of initial experiments we conducted to investigate the dynamics of drug references in Twitter. The first of these experiments explored the frequency of tweets mentioning specific drugs to determine drug popularity and how that changed over time. Following this popularity analysis, we then made use of Twitter's geolocation information to investigate whether patterns of drug references differed between regions of the United States and between different drugs.

5.1 Trends in Popularity

As mentioned, previous research studies have shown patterns in search engine queries and social network posts can provide high-quality predictors of real-world events like the spread of epidemics (especially the flu and other influenza-like illnesses) [7, 9]. We adapted these techniques to our drug domain by tracking specific drug-related keywords present within the 1% sample stream from April–July of 2014 and plotting the number of tweets per day. Figure 1 shows these per-day frequencies for three popular drugs (cocaine, heroin, and meth) and one new, designer drug (called "yaba"[4]). To smooth out the high variation in this data, we applied a 3-day moving average across all four drugs as well (Figure 2).

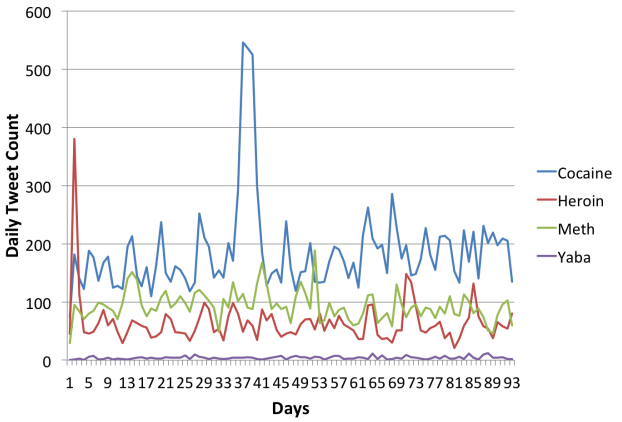

Figure 1: Daily Tweet Frequencies, April–July, 2014

One can see from this data that cocaine seems to be the more popular drug of the four despite its significant peak around day 38, an aberration discussed below. Yaba, on the other hand, is the least popular, which is consistent with

[4] http://www.justice.gov/archive/ndic/pubs5/5048/index.htm

its relatively recent introduction to the market. Furthermore, this ordering is generally consistent with the findings of the 2014 Global Drug Survey, which found cocaine, amphetamines, and opioids (in that order) to be three of the top twenty most used drugs [18].

Beyond overall popularity, we also examined this data to determine popularity *trends*; that is, whether a particular drug is experiencing an increase or decrease in its mentions over time. We postulate that such trends can be revealed by applying linear regression to this data and inspecting the slope of the resulting line (higher slopes indicate increasing popularity). Table 1 presents these slopes and their squared Pearson coefficients (R^2) to show goodness of fit. From this table, both heroin and meth have negative trends, but their regression lines exhibit poor fit, so we cannot draw definitive conclusions about their popularity. Yaba, on the other hand, has a better fit with only a slightly positive slope, which might indicate a relatively static popularity. Cocaine presents a more interesting case, however, in that it seems to have an overall positive trend but poor fit, which is likely the result of the large burst in usage around day 38. After reviewing the tweets during this time period (June 6-9), we discovered this peak was the result of a large spike in retweets of the form: "RT @<user>: ZAYN MALIK NOW DOING COCAINE..." where Zayn Malik is a member of the popular band One Direction. To account for this outlier, we removed these retweets, denoting this new data as *Cocaine*† (Figure 3) and refit, yielding both a higher slope and significantly higher R^2, indicating cocaine seems to becoming increasingly popular (at least on social media).

Table 1: Drug Popularity Trends

Drug	Slope	R^2
Cocaine	0.349	0.019
Cocaine†	0.539	0.295
Heroin	−0.118	0.017
Meth	−0.215	0.105
Yaba	0.024	0.195

5.2 Trends in Location

Every tweet contains a metadata field that can optionally include its geographic location in the form of global positioning system (GPS) coordinates. While only about 2% of all tweets contain populated versions of this coordinate field, we can now ask interesting questions like how these tweets are distributed geographically and how that distribution compares to existing data on geographic drug usage. To this end, we extracted geolocated tweets within the United States, divided them into sets for four popular drugs (cocaine, heroin, meth, and oxy), and reverse-geocoded them to identify the US state from which each tweet was posted. Table 2 lists the number of geocoded tweets used for each drug.

We could not simply color each US state by its number of tweets containing each drug because states with larger populations would naturally produce more tweets. To account for this population density effect, we normalized each state's tweet count by the expected number of tweets for that state. To calculate this expectation, we counted the number of tweets found in each US state in the 1% sample stream from April to July and determined the per-state probability distribution $P_s(x) = \frac{N_s}{N}$ where $s \in \{$ set of US states $\}$, N_s is the number of tweets in state x, and N is the total number of tweets. We then take the product of a state's likelihood $P_s(x)$ and the total number of tweets as the expected number of tweets in that state.

Figure 4 illustrates the per-state distributions with a heat map for each drug (blue corresponds to no tweets about the drug in the given state, and magenta corresponds to maximum number of tweets found in a given state).

Table 2: Geocoded Tweet Counts

Drug	Geocoded Tweet Count
Cocaine	281
Heroin	1,490
Meth	3,462
Oxy	223

From these figures, we get an immediate sense that the

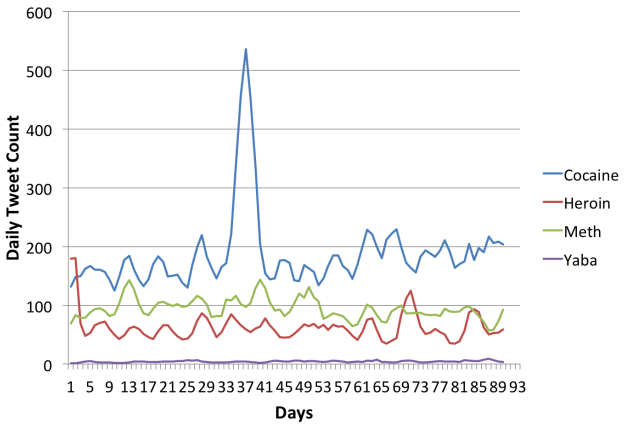

Figure 2: Smoothed Daily Tweet Frequencies

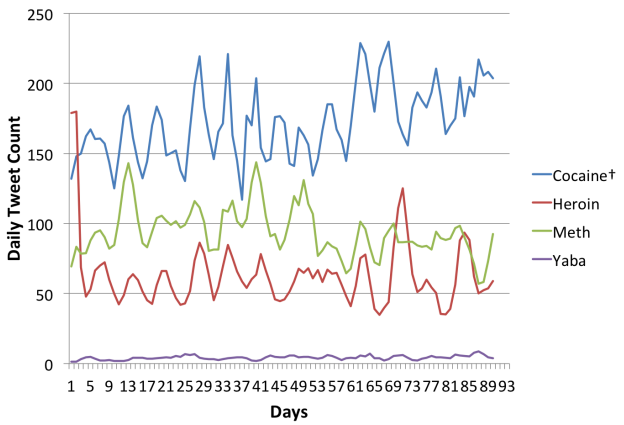

Figure 3: Frequencies w/o Zayn Malik Retweets

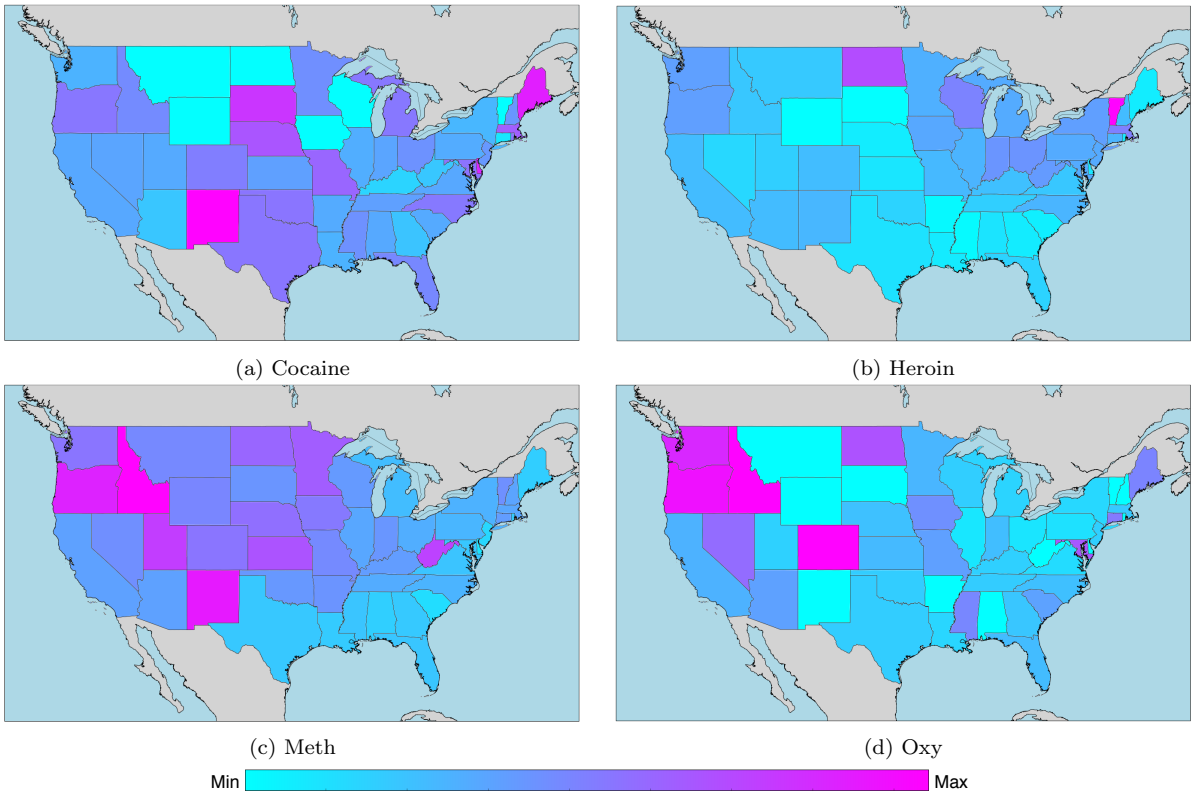

Figure 4: Geolocated Drug Mentions

popularity of different drugs vary widely based on geographic location. For instance, cocaine and meth seems relatively popular throughout the country while heroin's popularity seems more concentrated in the upper-midwest and northeastern states. Likewise, oxy seems much more concentrated in the pacific northwestern states like Idaho, Oregon, and Washington.

When compared against existing geographic drug abuse statistics, some results depicted in these figures are consistent. For instance, results from the 2012-2013 National Survey on Drug Use and Health show New Mexico and Massachusetts as having some of the highest percentages of cocaine use among people over the age of 12 for cocaine (Figure 4a) [1]. Similarly, according to Withdrawal.net, Seattle, Washington leads the United States in searches related to oxycodone [2].

We see a surprising result for heroin (Figure 4b) with particularly high references in North Dakota and Vermont. According to the Associated Press, however, North Dakota is in fact experiencing a rise in heroin usage [11]. Other reports also indicate Vermont has recently seen a large increase in heroin use, with a recent article from Politico Magazine calling the state "America's Heroin Capital" [17] and the New York Times coverage of Peter Shumlin, governor of Vermont, highlighting the state's drug issues [16].

6. DISCUSSION AND CONCLUSIONS

We have presented results from our initial work on creating a social media tool for detecting emerging trends in illicit drug use. Using terminology provided by subject matter experts, we built an architecture to collect posts from Twitter, filter them, and then analyze the temporal and geographic trends for specific drugs.

There are many challenges ahead in this space. As mentioned above, disambiguation is one of the largest. Some terms, like "weed" are mostly used to describe drug use while others, like "skittles", a slang term for ecstasy, or "horse", slang for heroin, are overwhelmingly used in non-drug contexts. It may be the case that even aggressive filtering will be unable to reliably pull drug-related tweets from the pool of data collected for such terms. One challenge we need to address soon will be understanding which terms are likely to be useful, which are not, and where to draw the line. How to determine that boundary and what kinds of tools to use in cases where there is useful data are all important next steps.

We believe geospatial information will be especially important for our work, but only about 2% of tweets actually come with location information. Fortunately, there are other techniques for inferring an approximate location for tweets (e.g. using friends' locations as in [4]). We will examine how this and other location-based analysis can be incorporated into our process.

Furthermore, it is generally difficult to acquire recent and timely statistics on drug abuse across a wide geographic region. This challenge both motivates and hinders our work: it motivates us in that a social media-based tool could provide more rapid situational awareness about new drug trends and impedes us in our ability to evaluate results against independent ground truth. For instance, most of the data available comes from surveys of at least one year old or older (for example, the latest nationwide release from the United States'

Substance Abuse and Mental Health Services Administration[5] is 2013).

Our project is focused on drug use, but we believe the architectural and analytic strategies we are developing will be useful well beyond this domain. Other public health applications and even entirely different domains that want to follow trends for defined topics are likely to benefit from using Twitter data in a similar way.

7. REFERENCES

[1] 2012-2013 national survey on drug use and health: National maps of prevalence estimates, by state. Technical report, National Survey on Drug Use and Health, 2014.

[2] Illicit searches: Visualizing drug-related web searches across the us. *Withdrawal.net*, 2014.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[4] R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401, Oct 2014.

[5] Michael D Conover, Clayton Davis, Emilio Ferrara, Karissa McKelvey, Filippo Menczer, and Alessandro Flammini. The geospatial characteristics of a social movement communication network. *PloS one*, 8(3):e55957, 2013.

[6] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013.

[7] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[8] David Hauger and Markus Schedl. Exploring geospatial music listening patterns in microblog data. In *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, pages 133–146. Springer, 2014.

[9] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE, 2010.

[10] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10:90–97, 2010.

[11] James MacPherson. Heroin use, sale on rise in north dakota. *Associated Press*, April 2014.

[12] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[13] Brett Meyer, Kevin Bryan, Yamara Santos, and Beomjin Kim. Twitterreporter: Breaking news detection and visualization through the geo-tagged twitter network. In *CATA*, pages 84–89, 2011.

[14] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, 2013.

[15] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[16] Katharine Seelye. In annual speech, vermont governor shifts focus to drug abuse. *New York Times*, January 2014.

[17] Gina Tron. How did idyllic vermont become america's heroin capital? *Politico Magazine*, February 2014.

[18] AR Winstock. The global drug survey 2014. findings. *Global Drug Survey*, 2014.

[19] ED Wish, EE Artigiani, and AS Billing. Community drug early warning system: The cdews pilot project. *Office of National Drug Control Policy. Executive Office of the President Washington, DC*, 2013.

[5] `http://www.samhsa.gov`