Firsthand Opiates Abuse on Social Media: Monitoring Geospatial Patterns of Interest Through a Digital Cohort

Duilio Balsamo
University of Turin & ISI Foundation
Turin, Italy
duilio.balsamo@unito.it

Paolo Bajardi
ISI Foundation
Turin, Italy
paolo.bajardi@isi.it

André Panisson
ISI Foundation
Turin, Italy
andre.panisson@isi.it

ABSTRACT

In the last decade drug overdose deaths reached staggering proportions in the US. Besides the raw yearly deaths count that is worrisome *per se*, an alarming picture comes from the steep acceleration of such rate that increased by 21% from 2015 to 2016. While traditional public health surveillance suffers from its own biases and limitations, digital epidemiology offers a new lens to extract signals from Web and Social Media that might be complementary to official statistics. In this paper we present a computational approach to identify a digital cohort that might provide an updated and complementary view on the opioid crisis. We introduce an information retrieval algorithm suitable to identify relevant subspaces of discussion on social media, for mining data from users showing explicit interest in discussions about opioid consumption in Reddit. Moreover, despite the pseudonymous nature of the user base, almost 1.5 million users were geolocated at the US state level, resembling the census population distribution with a good agreement. A measure of prevalence of interest in opiate consumption has been estimated at the state level, producing a novel indicator with information that is not entirely encoded in the standard surveillance. Finally, we further provide a domain specific vocabulary containing informal lexicon and street nomenclature extracted by user-generated content that can be used by researchers and practitioners to implement novel digital public health surveillance methodologies for supporting policy makers in fighting the opioid epidemic.

CCS CONCEPTS

• **Information systems → Information retrieval**; • **Applied computing → Health informatics**; • **Human-centered computing → Social media**.

KEYWORDS

Online disease monitoring, Opioid epidemic, Reddit

ACM Reference Format:
Duilio Balsamo, Paolo Bajardi, and André Panisson. 2019. Firsthand Opiates Abuse on Social Media: Monitoring Geospatial Patterns of Interest Through a Digital Cohort. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308558.3313634

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313634

1 INTRODUCTION

The opioid crisis emerged in the US from the interplay of several determinants and evolved over time through three major phases: from the abuse of drug prescriptions of late 80's [26], to heroin injecting users soaring in 2010 [31] and the very recent rise of synthetic opioids flooding the drug market [23]. The rate of overdose deaths involving opioids shows heterogeneous geographical patterns ranging from 4.9 per 100,000 inhabitants in Texas to 43.4 in West Virginia in 2016 [45]. Unfortunately, despite the effort of national agencies in monitoring the phenomenon, state-level estimates suffer biases due to unequal coverage of the surveillance system as well as intrinsic biases in counting overdose deaths that lead to low specificity of drugs involved [17, 25, 41]. Besides drug overdose maps, opioid prescription maps are provided on yearly basis by the Centers for Disease Control and Prevention (CDC), showing heterogeneous patterns also due to different prescribing policies implemented by individual states. In this perspective, the gold standard represented by official surveillance statistics has to be carefully considered in the light of the known biases of the reported numbers and estimates. A complementary way to investigate these phenomena is by field studies, through ethnographic approaches and structured interviews: sadly these methods require an exceptional effort in order to gather insights from firsthand users [28] or subjects under opioid substitution maintenance treatment [47], and are often limited to small sample sizes ranging from tens to hundreds of individuals. In this context, a digital epidemiology approach [42] aimed at integrating and complementing existing knowledge about the opioid crisis gathering information with a bottom-up unsolicited approach might be extremely valuable.

In recent years, Social Media changed the way drug users share information about opiates, give online warnings and avoid potentially toxic drug batches [21], pointing at "Reddit: the front page of Internet" [39] as a promising digital source of information. Reddit is a social content aggregation website, the 5th most popular website in the US (`Alexa.com`, Fall 2018), on which users can post, comment and vote content. It is structured in cross-referenced yet independent sub-communities (i.e. *subreddits*), centered around a variety of topics ranging from science to drugs and porn [30]. Reddit is constantly growing, with a total of almost 800 million comments in the year 2016 and almost a billion comments in the year 2017. Given the ease to register with a "throwaway" account, Reddit is often used to discuss topics otherwise considered socially unacceptable or unsuitable for the mainstream; users can actively engage with others, talking uninhibitedly about personal experiences [27], receiving back social support and even life-saving tips from sub-communities of peers [21]. However, navigating such massive platforms and finding areas of specific interest is usually cumbersome since topics

are self-organized bottom-up through users' interactions, leaving the users to find relevant topics by word of mouth or using a basic search feature.

The main contributions of this work are summarized as follows:

- Design a general purpose information retrieval algorithm able to identify regions of interest when conducting epidemiological surveillance and monitoring on social media
- Provide an open domain specific vocabulary related to opiates discussions
- Demonstrate how information disclosed by Reddit users can be used to estimate their geographical location
- Identify a novel digital cohort of individuals addressing in a pseudonymous fashion health related topics
- Provide prevalence maps of interest in opiates consumption

2 RELATED WORK

Digital epidemiology [42], also referred to as *digital disease detection* [4] and *infodemiology* [13], broadly includes the use of the Internet and digital technologies for collecting and processing health information, both in aggregate form for public health surveillance or from individuals for personal health monitoring. Data from a so called *digital cohort* might be collected with active participation of individuals, as in the case of participatory epidemiology [14], but might also be collected passively, e.g. from Social Media, as a byproduct of platforms designed for different purposes. Participatory systems have been implemented through the use of Web platforms [35] and signals collected from such systems have been shown to be useful for epidemic forecasting [49]. Data collected from the Web and Social Media have also shown to be useful for monitoring different infectious diseases [3, 5, 9, 22, 46].

Reddit has already proven to be suitable for a variety of research purposes, ranging from the study of user engagement and interactions between highly related communities [18, 48] to post-election political analyses [1]. Also, it has been useful to study the impact of linguistic differences in news titles [20] and to explore recent web-related issues such as hate speech [43] or cyberbullying [38] as well as health related issues like mental illness [8], also providing insights about the opioid epidemics [36]. It is worth to notice that since the platform is collectively generated by the users as in most Social Media, there is not a blueprint of Reddit to perimeter the area under study, and only few attempts tried to overcome this issue extracting meaningful maps or suitable embedding spaces [29, 34].

We approach the problem of selecting subreddits that are relevant for a specific topic as an information retrieval (IR) problem, where it is possible to retrieve topic-specific documents by expressing a limited set of known keywords. Language models [10, 37, 44] tackle this problem using a probabilistic approach with the idea that words that are relevant for a given topic would be more likely to appear in a relevant document. While language modeling is not a common approach for ranking documents collected from Social Media due to the inherent sparsity of the documents – e.g. for Twitter, more elaborated IR approaches are needed to resolve the sparsity of short texts for tweets [33] – subreddits, on the other hand, can be seen as very rich documents from which topic-specific word distributions can be built.

The set of keywords expressed by the user may not include some yet unknown terms that are relevant to the topic but very specific to the language models of the documents, pointing the necessity of query expansion techniques. Relevance feedback [40] and pseudo-relevance feedback [6] are common approaches for query expansion; many of these approaches use human judgment to manually select a subset of the top retrieved documents, and use them to expand the query. More recently, word embeddings [24] have been used to expand the query with terms that are semantically similar to the query terms. Techniques that are based on language modeling might incorporate term proximity information [12] to address the automatic query expansion problem, or use an approach based on Information Theory, exploiting the *Kullback-Leibler distance* for query reweighing [7] and for training local word embeddings [11].

3 DATASET AND DATA PREPARATION

A dataset [1] containing the list of all submissions and comments published in Reddit since 2007 is publicly available online [2] and is maintained monthly by adding recent entries. The dataset is not 100% complete and it contains some gaps [15], but these are very small for the years 2016 and 2017 (around 1% missing data per month). We use the union of all submissions and comments from years 2016 and 2017, for a total of 1,980,497,553 entries. Only subreddits with at least 100 entries were selected, resulting in a set of 1,973,863,886 entries with 74,810 distinct subreddits and 15,747,502 distinct users. The text of each entry is parsed and tagged using the spaCy NLP library [2] v1.9.0. For the part-of-speech tagging, a greedy averaged perceptron model was used [19]. Finally, lemmatization is applied to each POS tag; the English lemmatization data is taken from WordNet [32] and lookup tables are taken from Lexiconista [3]. After all terms are lemmatized, we select those that appear at least 100 times in the corpus, resulting in a vocabulary of 762,746 lemmas.

4 IDENTIFYING RELEVANT DOCUMENTS VIA DOCUMENT RANKING AND QUERY EXPANSION

As discussed in Section 2, the wealth of information contained in Reddit data is not readily available and has to be thoroughly mined. In this section we describe an iterative methodology of semi-automatic retrieval of documents in heterogeneous corpora, in which human intervention is as little as possible. It is worth to stress that the approach is general and fully unsupervised. However, we added a human-in-the-loop to include domain expert knowledge in the process and reach better results. On the other hand, a domain expert alone without the aid of the algorithmic pipeline for document ranking and query expansion would have been hopeless in navigating the Reddit world by hand. The steps are summarized in Algorithm 1. We start with a small set of keywords Q provided by the user, or *query* in the following. At each iteration, we select documents which are both relevant to the query and informative, and enrich the *query* terms set until we arrive to a stable list of documents and query terms. While this methodology works well on Reddit in the domain of topics related to the opioid epidemics

[1] https://files.pushshift.io/reddit/
[2] https://spacy.io/
[3] http://www.lexiconista.com/datasets/lemmatization/

(see Section 4.1), it is also sufficiently general to be used for other information retrieval tasks and might be valuable for different epidemiological research questions.

Algorithm 1: IR steps for document ranking

Input: Corpus C, query Q
Parameters: n, m, α
Initialize the vocabulary $\mathbf{V}$;
foreach *word* $w \in \boldsymbol{V}$ **do**
  calculate $p_C(w)$;
  foreach *document* $d \in C$ **do**
    calculate $p_d(w)$;
$Q_{new} \leftarrow Q$;
$K_{new} \leftarrow \emptyset$;
repeat
  $Q \leftarrow Q_{new}$;
  $K \leftarrow K_{new}$;
  $R_d \leftarrow$ Rank documents using score$(d \mid Q, C)$ (Eq. 2);
  $K_{new} \leftarrow$ top n documents in R_d;
  $R_w \leftarrow$ Rank terms using score$(w \mid K_{new})$ (Eq. 3);
  $Q_{candidate} \leftarrow$ top m terms from R_w;
  $Q_{new} \leftarrow$ manual selection of terms in $Q \cup Q_{candidate}$;
until $Q = Q_{new}$ *and* $K = K_{new}$;
Output: R_d, R_w

First, we create a general vocabulary $\mathbf{V}$ by collecting terms from the entire corpus in a bag-of-words fashion. We compute the probability of occurrence of a term w in the entire corpus C as the ratio $p_C(w) = f_C(w)/\sum_w f_C(w)$ between its raw count $f_C(w)$ in the corpus and the total number of words in the corpus. Let us also define the regularized marginal probability of occurrence of term w in a document d as

$$p_d(w) = \frac{f_d(w)}{\sum_w f_d(w) + \alpha} + p_C(w). \tag{1}$$

where $f_d(w)$ is the count of w in d. In case of corpora with very heterogeneous document sizes, the regularization term α is added to control "the size" of the language model of the documents (small documents will result in small marginal probabilities). In our experiments, we use $\alpha = 10^4$, so documents with total number of words lower than 10^4 have "flattened" probabilities in their language model. Adding $p_C(w)$ to the marginal probability of w reduces the impact of words that are rare or not present in a document, and only words that are more likely to appear in the document will impact the document ranking. The use of these two regularization terms (α and $p_C(w)$) are effective in low-count scenarios and have a small impact for words with high probability in the document or in case of documents with large language model.

With the intuition that a document will result to be relevant in the context of the query if it contains query terms more likely to appear in the document than in the general corpus, we evaluate

$$\text{score}(d \mid Q, C) = KLD_Q(p_d, p_C) = \sum_{w \in Q} p_d(w) log \frac{p_d(w)}{p_C(w)} \tag{2}$$

which is the total contribution of the query terms in the *Kullback-Leibler divergence* between the document and the whole corpus C. We consider the top n documents ranked by relevance as measured by Eq. (2) as the *set of relevant documents* K. Once n is chosen (and thus K is obtained), in order to enrich the query terms Q we assign a score to each term based on the logarithm of the likelihood ratio

$$\text{score}(w \mid K) = \sum_{k \in K} log \frac{p_k(w) + p_C(w)}{p_C(w)}. \tag{3}$$

Whenever the the maximum likelihood estimate $p_k(w)$ of a term in a relevant document is higher than its estimate in the general context $p_C(w)$ – i.e. the term is more informative in the document than in the general context – the contribution to the score will be high and positive. Conversely, whenever a term is less likely to appear in the document than in general context, its contribution to the score will be smaller, highlighting that the term is common or not relevant to the context. Considering the top m terms ranked by Eq. (3), a subset of previously unknown relevant terms is added to Q with the supervision of an expert. With the newly enriched set of query terms, the entire pipeline can be iteratively evaluated until convergence, reached if no new documents are added to the top n documents and no new relevant terms among the top m can be added to Q. These steps are summarized in Algorithm 1, and the document ranking resulting from the last iteration of the algorithm is used to select the most relevant documents.

4.1 Opioid related subreddits

Our assumption is that authors who post content in a subreddit related to a particular topic are *interested* in that topic. Therefore we consider all authors participating in threads on subreddits related to opioid consumption as those *interested* in the topic, and we discover such subreddits by applying the algorithm described in Section 4.1. After data preparation steps described in Section 3, starting from a list of opiates related keywords q = `[fentanyl, oxycodone, suboxone, morphine]` we extract the top $n = 10$ subreddits: `[suboxone, fentanyl, Opiatewithdrawal, TarkovTrading, heroinaddiction, ChronicPainPlayhouse, OpiatesRecovery, opiates, Methadone, PoppyTea]`. We then proceed with the iterative procedure of query enrichment and document ranking, considering as relevant terms only opioid drug names, i.e. chemical names (e.g. *oxycodone*), brand names (e.g. *Percocet*) and street slang (e.g. *percs*), disregarding drugs that might be abused together with opiates (like benzodiazepines) but are not in the opiates domain. The final set of *opioid related subreddits* used in this paper is K = `[fentanyl, suboxone, Opiatewithdrawal, Methadone, opiates, pillhead, lean, ChronicPainPlayhouse, heroinaddiction, OpiatesRecovery]`. All the 37,009 users who posted on such subreddits in 2016 or 2017 are then considered as interested in opiates. Figure 1 (top) shows the final ranking score for the subreddits, plotted against their size in terms of number of words.

4.2 Opioid specific vocabulary

When applied to Reddit, the query expansion approach is particularly useful in revealing how a topic is discussed on subreddits. Given the large user base of Reddit and the tendency of the users in employing slang and street names alongside proper drug names,

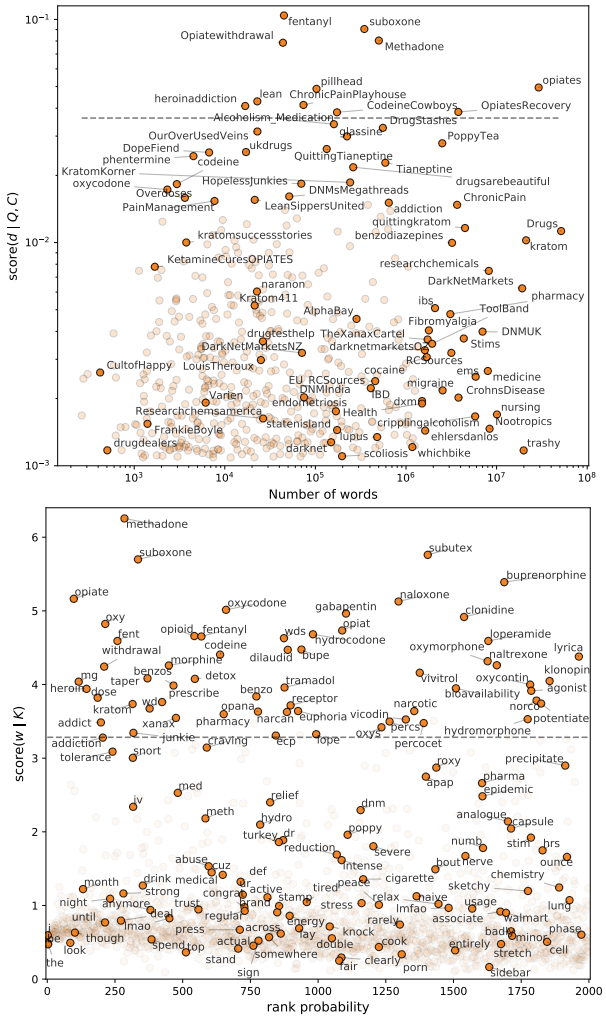

Figure 1: *Opiates subreddits* (top): individual subreddits are shown in a coordinate space of the number of words in the subreddit and the final ranking score. The subreddits above the dashed line are those selected as K opiates related subreddits. *Opiates vocabulary* (bottom): Top 2,000 terms sorted by rank probability of term in the set K. Terms above the dashed line were selected as query term candidates in the last step of the subreddit retrieval algorithm.

this method is very helpful in acknowledging alternative names of drugs, like *sub* for *suboxone* and *bth* for *black tar heroin*.

As a byproduct of the proposed methodology applied to the opiates domain, we extracted a topic-specific vocabulary by weighting each term of the vocabulary with Eq. 3. Very specific opioid-related terms (i.e. with high probability in opiates subreddits and low probability in the whole corpus) have large positive values of $score(w \mid K)$, as shown in Figure 1 (bottom), while stop-words and common terms have small score values. A total of 2,616 terms out of the original 762,746 (0.3%) have score higher than 1. The full list of vocabulary terms ranked by score is available for research purposes [4].

5 GEOLOCATING USERS ON REDDIT

Reddit does not provide any explicit information about users' location, therefore we apply three methodologies to assign a location to users. Finally, we merge the mappings in a single user-location matching.

1. Self reporting via regular expression: Reddit users often declare geographical information about themselves in submission or comment texts. We selected all texts containing the expression `'I live in'` (3,337,850 instances in 2016 and 2017) and extracted candidate expressions from the text that follows, to identify which ones represented US states and cities. We started with a set of US cities from the GeoNames database [5] with population higher than 20k, and selected only the candidate expressions that included both the city name and the state (e.g. 'Newark, California' or 'Newark, CA') to avoid confusion with cities with same name (e.g. 'Newark, New Jersey'). Once the US state for these expressions were assigned and removed from the candidate expressions, we proceeded to all US cities with population higher than 200k, selecting expressions with the name of the city and their variants (e.g. 'New York', 'Big Apple'). After assigning the corresponding US state for these expressions and removing them from the candidates, we proceeded to select the expressions with a state name on it (e.g. 'Alabama', 'California'). Among the initial set of candidate expressions, 886,919 (27%) had a state associated to it. By removing inconsistent self reporting (13,374 users who reported more than one US state) we geolocated 378,898 distinct users.

2. Self reporting via *user flairs*: In Reddit, *user flairs* are attributes (usually selected by the users) that are attached to their submissions or comments in a specific subreddit. In some subreddits flairs might be limited to a set of geographical locations (countries, states, cities and city neighborhoods), meaning that users should identify themselves with one of these locations. A user selecting a location flair is therefore considered equivalent to a user self reporting its location. We mapped the users participating in subreddits with location flairs referring to US states to their *flaired* positions. Using this approach, we mapped 206,125 users to the 51 US states (including District of Columbia) by selecting the most common among the position flairs expressed by a user.

3. Posting on subreddits specific to locations: Reddit includes subreddits discussing topics specific to geographical locations (e.g. `r/Alabama` or `r/Seattle`). The subreddit `r/LocationReddits` keeps a curated list of these local subreddits. We collected from the page corresponding to North America[6] the mappings of 2,844 subreddits to 51 US states. By assuming that a user who posts comments in one of these subreddits is likely to live that location, we estimated the position of 1,198,096 authors.

After retrieving US states positions using the three methods above, we found that about 12% of mapped users expressed multiple locations. In order to uniquely map authors and states, for location flairs and LocationReddits sources we assigned each author a unique

[4] https://github.com/ISIFoundation/WWW19OpiatesAbuseSocialMedia
[5] http://www.geonames.org/
[6] https://www.reddit.com/r/LocationReddits/wiki/faq/northamerica

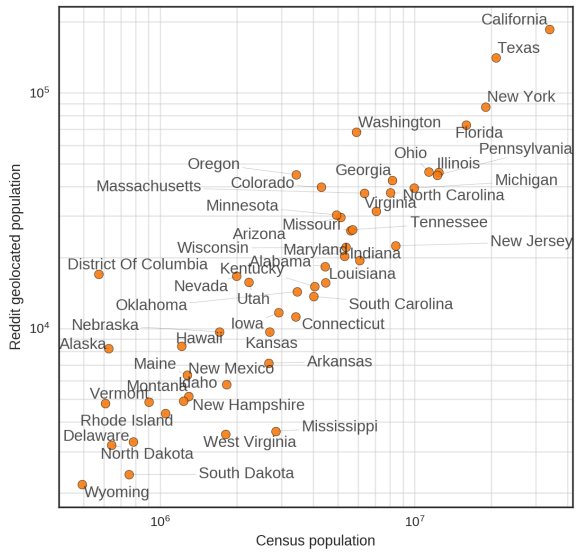

Figure 2: *Reddit geolocated population*: scatter plot of the number of geolocated Reddit users and census population.

location, it being the most frequent among the ones expressed by the author. We discarded authors whose most frequent location was not unique. This resulted in 194,008 authors retrieved via location flairs (5.9% loss) and 1,077,516 via LocationReddits (1.4% loss).

We evaluated Pearson's r correlation between the log of the 2000 US Census population and the log of the population assigned to the same US states using the three methodologies. Results are in good agreement for all sources, with r= 0.85, 0.91 and 0.86 for respectively User flairs, Regular expression and LocationReddits, and all p-values below 1e-12.

Finally, we merged the information from all three sources in a unique location for each author. We considered the regular expression technique to be the most reliable due to its unambiguous self reporting nature, resulting in the highest correlation with census data. We proceeded in the merging process by first assigning the authors their regular expression location, if present. If missing, we assigned them their position from the joint information of location flairs and LocationReddits by summing the occurrences of locations expressed in the two sources and verifying the uniqueness of the most frequent location. Although some approaches have been proposed to geolocate users using language models [16], we rely on a conservative approach with the aim of reducing misclassification, considering only explicit geographical information directly provided by the authors. The full set of users geolocated using the above methodology consists of 1,408,388 users, with state representativeness in the order of 5.5 Reddit users per thousand U.S. residents (median value among all U.S. states). Although we acknowledge a potential bias due to heterogeneities in Reddit population coverage and users demographics, the number of Reddit users has good linear correlation of r = 0.89 and p-value below 1e-12 with census population (Figure 2).

Region	Division	Opiates authors	Reddit authors	Interest prevalence
Northeast	Middle Atlantic	1,186	154,418	768.05
Northeast	New England	455	69,132	658.16
Midwest	East North Central	1,082	172,902	625.79
Midwest	West North Central	424	92,931	456.25
South	East South Central	457	63,269	722.31
South	South Atlantic	1,656	242,470	682.97
South	West South Central	1,079	177,856	606.67
West	Mountain	793	119,742	662.26
West	Pacific	1,894	315,668	599.10

Table 1: *Interest prevalence by US regional division*: number of opiates authors, total number of authors, and interest prevalence per 100,000 individuals measured on Reddit.

6 PREVALENCE OF INTEREST IN OPIATES

Conversations in opiates related subreddits branch off in many topics, mostly regarding opiates usage, dosages, interactions with other substances, safe practices and withdrawal, usually with a personal perspective. Users share their health and addiction status and provide support among each other. In general, they share a common and firsthand interest in opiates experiences. Thus, the number of authors participating in the conversations in opiates related subreddits (as identified in Section 4) is not to be considered as a crude number of opiates users and addicts but it is rather to be considered as a proxy of users personally interested in opiates in the broadest sense.

Using the geographical information of Reddit population estimated in Section 5 and having identified the opiates authors in Section 4.1, we were able to evaluate the *opiates interest prevalence* at the US state level as the fraction of geolocated users engaged in opiates subreddits and the total estimated population.

We mapped to US states the 9,026 geolocated users who posted in opioid related subreddits, equivalent to 24% of the opiates authors, for a mean interest prevalence of 636 per 100,000 Reddit users (CDC data for 2016 reported an age-adjusted rate of overdose deaths of 19.8 per 100,000 , and opioid prescription rate of 66.5 prescriptions dispensed per 100 persons). The areas of greater interest according to our estimates is the South Region (Table 1), with high prevalence for Mississippi, Arkansas, Louisiana and the highest measured value of 1,180 interested users per 100,000 population in West Virginia (Figure 3, green map). Middle Atlantic and New England states like Pennsylvania, New Jersey and Rhode Island are also largely involved, showing high interest rates ranging between 850 to 900 individuals per 100,000. In line with official statistics about drugs overdose deaths, West North Central states are those with the lowest interest rate measured on Reddit, ranging from 341 per 100,000 in Nebraska to 510 per 100,000 in Minnesota.

We confronted the estimated interest prevalence with official statistics from the Centers for Disease Control and Prevention [7] grasping different angles of the opioid crisis. In particular we focused

[7] https://www.cdc.gov/drugoverdose/data/index.html

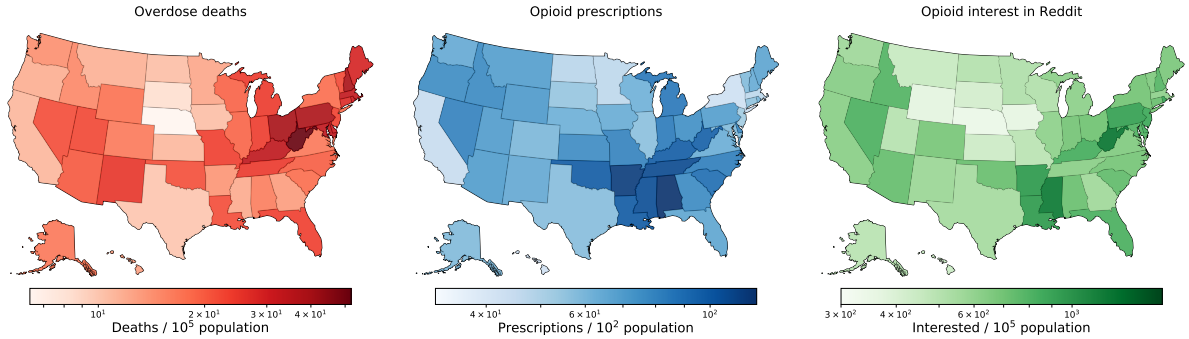

Figure 3: *US states distribution maps:* choropleth maps representing the overdose deaths rate for 2016 (red), the opioid prescription rate for 2016 (blue), the opioid interest rate in Reddit for 2016 and 2017 (green).

on opioid drug overdose deaths rates and retail opioid prescribing rates, both regarding 2016 (the most recent available dataset at the time of writing), shown in Figure 3 in red and blue respectively. These two phenomena seem fairly uncorrelated, with a Pearson's correlation of 0.068 (p-value of 0.637). It is worth to stress that those "gold-standard" data are the only ones provided by CDC that allows for comparisons between different states: although the counting of drugs overdose deaths includes every drug and is not broken down by drug type, state-level estimates of opiate-related overdose deaths are affected by heterogeneities in the surveillance system. On the other hand, official statistics about prescribing rates that includes both appropriate prescriptions and drug abuse are affected by different prescription policies in place at different states.

The interest prevalence shows fairly high positive linear correlations with CDC rates, respectively r = 0.45 (p-value = 8.4e-04) with the opioid overdose deaths rate and r = 0.506 (p-value = 1.6e-04) with the retail opioid prescribing rate. These correlations suggest that the signal of interest in opiates measured on Reddit partially explains the observed phenomena around opioid epidemics measured by the standard surveillance system. Moreover, we trained a linear regression model to fit the estimated prevalence using drug overdose death rates and prescriptions rates as features and predicted new values of interest prevalence, resulting in an higher correlation of r = 0.655 (p-value = 1.7e-07) with the estimated interest prevalence. This result confirms that we are sensing a broad signal, tied to drug overdose deaths and opioid prescriptions but probably accounting for more complex aspects of the phenomenon.

Leveraging the geolocated cohort, we also evaluated the temporal variation of interest prevalence between 2016 and 2017 broken down by state. According to Figure 4, the interest prevalence decreased only in 8 states while in general we observe that in areas with a good coverage of opiates-related users (namely, California, Texas, New York, Florida), the interest prevalence increased by 10% to 20%. It is worth to stress that, at the time of writing, no official data about 2017 drug overdose deaths and associated trends are available, highlighting the tremendous potential of a digital epidemiology approach to gather timely insights about hard-to-reach information of health-related topics at the population level.

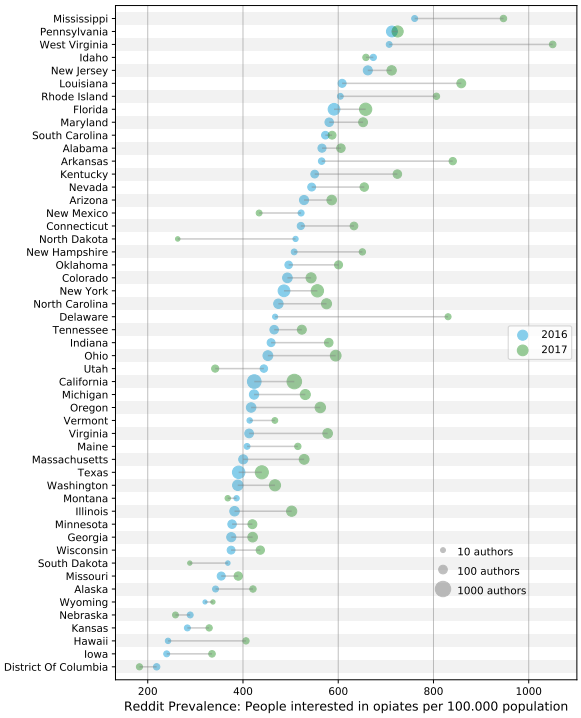

Figure 4: *Opioids interest prevalence:* number of Reddit authors per 100,000 Reddit population. Prevalence values for years 2016 and 2017 are reported on the x-axis. The size of the bubbles is proportional to the number of Reddit authors.

7 CONCLUSION

This study provides an analysis of Reddit content related to personal opiates abuse in the period of 2016 and 2017. Starting from almost 2 billion posts over 74k distinct subreddits, we applied a general information retrieval algorithm to identify specific subreddits of interest thus selecting 37,009 users that show an explicit interest in the topic. 1.5 million pseudonymous Reddit users were geolocated at US state level by looking at the content they generated in the platform. The number of mapped users for each state are in good agreement with census data, with some differences in terms of coverage. Such cohort might represent a biased yet valuable digital

observatory on several social, political and health-related topics. The prevalence of opiates interest extracted with the presented approach shows a complementary perspective to official surveillance, and its geographical heterogeneity partially encodes signals from opioid prescribing rates and drug overdose deaths.

ACKNOWLEDGMENTS

DB acknowledges support from the Lagrange Project and CRT Foundation (http://isi.it/en/lagrange-project/project).

REFERENCES

[1] Michael Barthel. 2016. How the 2016 presidential campaign is being discussed on Reddit. Retrieved November 5, 2018 from http://www.pewresearch.org/fact-tank/2016/05/26/how-the-2016-presidential-campaign-is-being-discussed

[2] Jason Baumgartner. 2015. I have every publicly available Reddit comment for research. 1.7 billion comments @ 250 GB compressed. Any interest in this? – r/datasets. Retrieved November 5, 2018 from https://redd.it/3bxlg7

[3] John S Brownstein, Shuyu Chu, Achla Marathe, Madhav V Marathe, Andre T Nguyen, Daniela Paolotti, Nicola Perra, Daniela Perrotta, Mauricio Santillana, Samarth Swarup, Michele Tizzoni, Alessandro Vespignani, Anil Kumar S Vullikanti, Mandy L Wilson, and Qian Zhang. 2017. Combining participatory influenza surveillance with modeling and forecasting: three alternative approaches. *JMIR public health and surveillance* 3, 4 (2017).

[4] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. 2009. Digital disease detection - harnessing the Web for public health surveillance. *New England Journal of Medicine* 360, 21 (2009), 2153–2157.

[5] John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. 2009. Influenza A (H1N1) virus, 2009 –online monitoring. *New England Journal of Medicine* 360, 21 (2009), 2156–2156.

[6] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. 1995. Automatic query expansion using SMART: TREC 3. *Proceedings of The third Text REtrieval Conference (TREC-3)* (1995), 69–69.

[7] Claudio Carpineto, Renato De Mori, Giovanni Romano, and Brigitte Bigi. 2001. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems (TOIS)* 19, 1 (2001), 1–27.

[8] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity. In *Proceedings of the Eighth Intl Conf. on Weblogs and Social Media, ICWSM 2014*. The AAAI Press.

[9] Rumi Chunara, Jason R Andrews, and John S Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene* 86, 1 (2012), 39–45.

[10] W Bruce Croft and John Lafferty. 2013. *Language modeling for information retrieval*. Vol. 13. Springer Science & Business Media.

[11] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query expansion with locally-trained word embeddings. *arXiv preprint arXiv:1605.07891* (2016).

[12] Liana Ermakova, Josiane Mothe, and Elena Nikitina. 2016. Proximity relevance model for query expansion. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. ACM, 1054–1059.

[13] Gunther Eysenbach. 2009. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of medical Internet research* 11, 1 (2009).

[14] Clark C Freifeld, Rumi Chunara, Sumiko R Mekaru, Emily H Chan, Taha Kass-Hout, Anahi Ayala Iacucci, and John S Brownstein. 2010. Participatory epidemiology: use of mobile phones for community-based health reporting. *PLoS medicine* 7, 12 (2010), e1000376.

[15] Devin Gaffney and J Nathan Matias. 2018. Caveat Emptor, Computational Social Science: Large-Scale Missing Data in a Widely-Published Reddit Corpus. *arXiv preprint arXiv:1803.05046* (2018).

[16] Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research* 49 (2014), 451–500.

[17] H. Hedegaard, M. Warner, L. Paulozzi, and R. Johnson. 2014. Issues to Consider When Analyzing ICD-10 Coded Data on Drug Poisoning (Overdose) Deaths. National Center for Health Statistics and the National Center for Injury Prevention and Control.

[18] Jack Hessel, Chenhao Tan, and Lillian Lee. 2016. Science, AskScience, and BadScience: On the Coexistence of Highly Related Communities.. In *Proc. of the 10th Intl AAAI Conf. on Web and Social Media, ICWSM 2016*. The AAAI Press, 171–180.

[19] Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1373–1378.

[20] Benjamin D Horne and Sibel Adali. 2017. The impact of crowds on news engagement: A Reddit case study. (2017). arXiv:1703.10570v2

[21] Fernando Alfonso III. 2017. How a Reddit forum has become a lifeline to opioid addicts in the US. https://www.theguardian.com/society/2017/jul/19/opioid-addiction-reddit-fentanyl-appalachia.

[22] Taha A Kass-Hout and Hend Alhinnawi. 2013. Social media in public health. *Br Med Bull* 108, 1 (2013), 5–24.

[23] Andrew Kolodny, David T Courtwright, Catherine S Hwang, Peter Kreiner, John L Eadie, Thomas W Clark, and G Caleb Alexander. 2015. The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction. *Annual review of public health* 36 (2015), 559–574.

[24] Saar Kuzi, Anna Shtok, and Oren Kurland. 2016. Query expansion using word embeddings. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. ACM, 1929–1932.

[25] Michael G Landen, Stuart Castle, Kurt B Nolte, Mary Gonzales, Luis G Escobedo, Barbara F Chatterjee, Karen Johnson, and C Mack Sewell. 2003. Methodological issues in the surveillance of poisoning, illicit drug overdose, and heroin overdose deaths in New Mexico. *American journal of epidemiology* 157, 3 (2003), 273–278.

[26] Pamela TM Leung, Erin M Macdonald, Matthew B Stanbrook, Irfan A Dhalla, and David N Juurlink. 2017. A 1980 letter on the risk of opioid addiction. *New England Journal of Medicine* 376, 22 (2017), 2194–2195.

[27] Lydia Manikonda, Ghazaleh Beigi, Huan Liu, and Subbarao Kambhampati. 2018. Twitter for Sparking a Movement, Reddit for Sharing the Moment:# metoo through the Lens of Social Media. *arXiv preprint arXiv:1803.08022* (2018).

[28] Sarah G Mars, Philippe Bourgois, George Karandinos, Fernando Montero, and Daniel Ciccarone. 2014. "Every 'never' I ever said came true": Transitions from opioid pills to heroin injecting. *Int 'l Journal of Drug Policy* 25, 2 (2014), 257–266.

[29] Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 27–31.

[30] Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2018. The anatomy of Reddit: An overview of academic research. *arXiv preprint arXiv:1810.10881* (2018).

[31] Thomas Michel and Joseph Loscalzo. 2015. Shifting patterns of prescription opioid and heroin abuse in the United States. *Mars* 372 (2015), 241–8.

[32] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[33] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Searching microblogs: coping with sparsity and document quality. In *Proc. of the 20th ACM int'l conf. on Information and knowledge management*. ACM, 183–188.

[34] Randal S Olson and Zachary P Neal. 2015. Navigating the massive world of reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science* 1 (2015), e4.

[35] Daniela Paolotti, Annasara Carnahan, Vittoria Colizza, Ken Eames, John C. Edmunds, Gabriel Gomes, Carl E Koppeschaar, Moa Rehn, Ronald Smallenburg, Clément Turbelin, Sander P. van Noort, and Alessandro Vespignani. 2014. Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. *Clinical Microbiology and Infection* 20, 1 (2014), 17–21.

[36] Albert Park and Mike Conway. 2017. Towards Tracking Opium Related Discussions in Social Media. *Online journal of public health informatics* 9, 1 (2017).

[37] Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 275–281.

[38] Tazeek Bin Abdur Rakib and Lay-Ki Soon. 2018. Using the Reddit Corpus for Cyberbully Detection. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 180–189.

[39] Reddit. 2018. Reddit: The front page of the internet. https://www.reddit.com/.

[40] Joseph John Rocchio. 1971. Relevance feedback in information retrieval. In *The Smart retrieval system: experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall, 313–323.

[41] Christopher J Ruhm. 2017. Geographic variation in opioid and heroin involved drug poisoning mortality rates. *American journal of preventive medicine* 53, 6 (2017), 745–753.

[42] Marcel Salathé, Linus Bengtsson, Todd J. Bodnar, Devon D. Brewer, John S. Brownstein, Caroline Buckee, Ellsworth M. Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L. Mabry, and Alessandro Vespignani. 2012. Digital Epidemiology. *PLOS Computational Biology* 8, 7 (07 2012), e1002616.

[43] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159* (2017).

[44] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press.

[45] Puja Seth, Lawrence Scholl, Rose A Rudd, and Sarah Bacon. 2018. Overdose deaths involving opioids, cocaine, and psychostimulants -United States, 2015–2016. *American Journal of Transplantation* 18, 6 (2018), 1556–1568.

[46] Megha Sharma, Kapil Yadav, Nitika Yadav, and Keith C Ferdinand. 2017. Zika virus pandemic - analysis of Facebook as a social media health information platform. *American journal of infection control* 45, 3 (2017), 301–302.

[47] Luis Sordo, Gregorio Barrio, Maria J Bravo, B Iciar Indave, Louisa Degenhardt, Lucas Wiessing, Marica Ferri, and Roberto Pastor-Barriuso. 2017. Mortality

risk during and after opioid substitution treatment: systematic review and meta-analysis of cohort studies. *bmj* 357 (2017), j1550.

[48] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1056–1066.

[49] Qian Zhang, Nicola Perra, Daniela Perrotta, Michele Tizzoni, Daniela Paolotti, and Alessandro Vespignani. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. 311–319.