Correlating Twitter Language with Community-Level Health Outcomes

Arno Schneuwly
EPFL
arno.schneuwly@epfl.ch

Ralf Grubenmann
SpinningBytes
rg@spinningbytes.com

Séverine Rion Logean
Swiss Re
severine_rion@swissre.com

Mark Cieliebak
ZHAW
ciel@zhaw.ch

Martin Jaggi
EPFL
martin.jaggi@epfl.ch

Abstract

We study how language on social media is linked to diseases such as atherosclerotic heart disease (AHD), diabetes and various types of cancer. Our proposed model leverages state-of-the-art sentence embeddings, followed by a regression model and clustering, without the need of additional labelled data. It allows to predict community-level medical outcomes from language, and thereby potentially translate these to the individual level. The method is applicable to a wide range of target variables and allows us to discover known and potentially novel correlations of medical outcomes with life-style aspects and other socioeconomic risk factors.

1 Introduction

Surveys and empirical studies have long been a cornerstone of psychological, sociological and medical research, but each of these traditional methods pose challenges for researchers. They are time-consuming, costly, may introduce a bias or suffer from bad experiment design.

With the advent of big data and the increasing popularity of the internet and social media, larger amounts of data are now available to researchers than ever before. This offers strong promise new avenues of research using analytic procedures, obtaining a more fine-grained and at the same time broader picture of communities and populations as a whole (Salathé, 2018). Such methods allow for faster and more automated investigation of demographic variables. It has been shown that Twitter data can predict atherosclerotic heart-disease risk at the community level more accurately than traditional demographic data (Eichstaedt et al., 2015). The same method has also been used to capture and accurately predict patterns of excessive alcohol consumption (Curtis et al., 2018).

In this study, we utilize Twitter data to predict various health target variables (AHD, diabetes, various types of cancers) to see how well language patterns on social media reflect the geographic variations of those targets. Furthermore, we propose a new method to study social media content by characterizing disease-related correlations of language, by leveraging available demographic and disease information on the community level. In contrast to (Eichstaedt et al., 2015), our method is not relying on word-based topic models, but instead leverages modern state-of-the-art text representation methods, in particular sentence embeddings, which have been in increasing use in the Natural Language Processing, Information Retrieval and Text Analytics fields in the past years. We demonstrate that our approach helps capturing the semantic meaning of tweets as opposed to features merely based on word frequencies, which come with robustness problems (Brown and Coyne, 2018; Schwartz et al., 2018). We examine the effectiveness of sentence embeddings in modeling language correlates of the medical target variables (disease outcome).

Section 2 gives a generalized description of our method. We apply the previously described method to the tweets and health data in Section 3 The system's performance is evaluated in Section 4 followed by the discussion in Section 5. Our code is available on github.com/epfml/correlating-tweets.

2 Method

We are given a large quantity of text (sentences or tweets) in the form of social media messages by individuals. Each individual—and therefore each sentence—is assigned to a predefined category, for example a geographic region or a population subset. We assume the number of sentences to be sig-

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 71–78
Florence, Italy, August 2, 2019.

nificantly larger than the number of communities. Furthermore, we assume that the target variable of interest, for example disease mortality or prevalence rate, is available for each community (but not for each individual). Our system consists of two subsystems:

1. *(Prediction)* The predictive subsystem makes predictions of target variables (e.g. AHD mortality rate) based on aggregated language features. The resulting linear predictions are applicable on the community level (e.g. counties) or on the individual level, and are trained using k-fold cross-validated Ridge regression.

2. *(Interpretability)* The averaged regression weights from the prediction system allow for interpretation of the system: We use a fixed clustering (which was obtained from all sentences without any target information), and then rank each topic cluster with respect to a prediction weight vector from point 1). The top and bottom ranked topic clusters for each target variable give insights into known and potentially novel correlations of topics with the target medical outcome.

In summary, the community association is used as a proxy or weak labelling to correlate individual language with community-level target variables. The following subsections give a more detailed description of the two subsystems.

2.1 System Description

Let $\mathcal{S}$ be the set of sentences (e.g. tweets), with their total number denoted as $|\mathcal{S}| = S$. Each sentence is associated to exactly one of the A communities $\mathcal{A} = \{a_1, \ldots, a_A\}$ (e.g. geographic regions). The function $\delta : \mathcal{S} \to \mathcal{A}$ defines this mapping. Let $\mathbf{y} \in \mathbb{R}^A$ be the target vector for an arbitrary target variable, so that each community a_j has a corresponding target value $y_{a_j} \in \mathbb{R}$.

Preprocessing and Embeddings. The complete linguistic preprocessing pipeline of a sentence is incorporated by the function $\rho(s_i)$, $\forall\ i \in \{1, \ldots, S\}$, which represents an arbitrary sentence s_i as a sequence of tokens. Each sentence s_i then is represented by a D-dimensional embedding vector providing a numerical representation of the semantics for the given short text:

$$\mathbf{x}_i = \textit{Sent2Vec}(\rho(s_i)) \in \mathbb{R}^D. \tag{1}$$

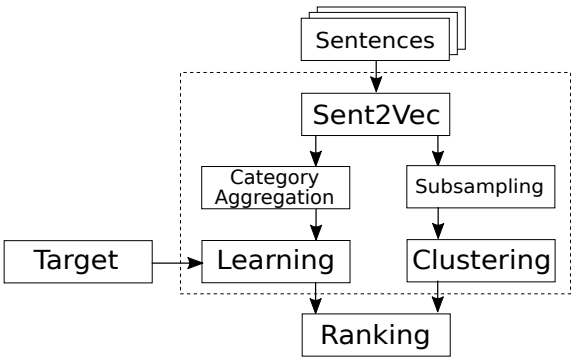


Figure 1: System Description.

While our method is generic for any text representation method, here Sent2Vec (Pagliardini et al., 2018) was chosen for its computational efficiency and scalability to large datasets.

2.2 Feature Aggregation

We use averaging of the sentence embedding vectors over each community to obtain the language features for each community. Formally, the complete feature matrix of all sentences is denoted as $\mathbf{X} \in \mathbb{R}^{S \times D}$. For our approach, the sentence embedding features are averaged over each community a_j. Formally, an individual feature $\overline{x}_{a_j,d}$ of the averaged embedding $\overline{\mathbf{x}}_{a_j} \in \mathbb{R}^{1 \times D}$ for a given community a_j is defined as

$$\overline{x}_{a_j,d} = \frac{1}{N_{a_j}} \sum_{x_i : s_i \in \mathcal{S}\ \wedge\ \delta(s_i) = a_j} x_{i,d}, \tag{2}$$

where $N_{a_j} = |\{s_i : s_i \in S \wedge \delta(s_i) = a_j\}|$ is the number of sentences belonging to community a_j. Consequently, the aggregated community-level embedding matrix is given by

$$\overline{\mathbf{X}} = \begin{bmatrix} \overline{\mathbf{x}}_{a_1}^\top \\ \vdots \\ \overline{\mathbf{x}}_{a_A}^\top \end{bmatrix} \in \mathbb{R}^{A \times D}. \tag{3}$$

2.3 Train-Test Split

Leveraging the targets available for each community, our regression method is applied to the aggregated features $\overline{\mathbf{X}}$ and the target $\mathbf{y}$. We employ K-fold cross-validation: the previously defined set $\mathcal{A}$ is split into K as equally sized pairwise disjoint subsets $\mathcal{A}_k$ as possible such that: $\mathcal{A} = \bigcup_{k=1}^{K} \mathcal{A}_k$, $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset\ \forall i, j \in 1, \ldots, K, i \neq j$ and $|\mathcal{A}_1| \approx \cdots \approx |\mathcal{A}_K|$. The training set for a fold k is $\text{TR}_k = \left(\bigcup_{i=1}^{K} \mathcal{A}_i\right) \setminus \mathcal{A}_k$ with the corresponding test set $\text{TE}_k = \mathcal{A}_k$, where $N_k^\theta = |\text{TR}_k|$ and $N_k^\Lambda =$

$|\text{TE}_k|$. The operators $\theta_k : \{1, \ldots, N_k^\theta\} \rightarrow \text{TR}_k$ and $\Lambda_k : \{1, \ldots, N_k^\Lambda\} \rightarrow \text{TE}_k$ uniquely map the indexes to the corresponding communities a_j for the k^{th} train-test split. For each split k the train and test embedding matrices respectively are defined as

$$\overline{\mathbf{X}}_{\theta_k} = \left[\overline{\mathbf{x}}_{\theta_k(1)}, \ldots, \overline{\mathbf{x}}_{\theta_k(N_k^\theta)}\right]^\top, \quad (4)$$

$$\overline{\mathbf{X}}_{\Lambda_k} = \left[\overline{\mathbf{x}}_{\Lambda_k(1)}, \ldots, \overline{\mathbf{x}}_{\Lambda_k(N_k^\Lambda)}\right]^\top. \quad (5)$$

Accordingly, we define the target vectors

$$\mathbf{y}_{\theta_k} = \left[y_{\theta_k(1)}, \ldots, y_{\theta_k(N_k^\theta)}\right]^\top, \quad (6)$$

$$\mathbf{y}_{\Lambda_k} = \left[y_{\Lambda_k(1)}, \ldots, y_{\Lambda_k(N_k^\Lambda)}\right]^\top. \quad (7)$$

2.4 Ridge Regression

For each train-test split k we perform linear regression from the community-level textual features $\overline{\mathbf{X}}_{\theta_k}$ to the health target variable $\mathbf{y}_{\theta_k}$. We employ Ridge regression (Hoerl and Kennard, 1970). In our context, the Ridge regression is defined as the following optimization problem:

$$\min_{\boldsymbol{\omega}_k \in \mathbb{R}^D} \frac{1}{2A} \sum_{i=1}^{N_k^\theta} \left[y_{\theta_k(i)} - \overline{\mathbf{x}}_{\theta_k}^\top \boldsymbol{\omega}_k\right]^2 + \lambda \|\boldsymbol{\omega}_k\|_2^2, \quad (8)$$

where the optimal solution is

$$\boldsymbol{\omega}_k^\star = \left(\overline{\boldsymbol{X}}_{\theta_k}^\top \overline{\mathbf{X}}_{\theta_k} + 2N_k^\theta \lambda \mathbf{I}\right)^{-1} \overline{\mathbf{X}}_{\theta_k}^\top \quad \in \mathbb{R}^D. \quad (9)$$

Within each each fold we tune the regularization parameter λ.

2.5 Prediction Subsystem

Let $\overline{\boldsymbol{y}}_{\Lambda_k} = \overline{\mathbf{X}}_{\Lambda_k} \boldsymbol{\omega}_k^\star = [\overline{y}_{\Lambda_k(1)}, \ldots, \overline{y}_{\Lambda_k(N_k^\Lambda)}]^\top$ be the predicted values for the test set of the split k. The concatenated prediction vector for all splits is

$$\overline{\mathbf{y}}_\Lambda = \begin{bmatrix} \overline{\mathbf{y}}_{\Lambda_1}^\top \\ \vdots \\ \overline{\mathbf{y}}_{\Lambda_K}^\top \end{bmatrix} \in \mathbb{R}^A \quad (10)$$

Accordingly, we define the concatenated true target vector as

$$\mathbf{y}_\Lambda = \begin{bmatrix} \mathbf{y}_{\Lambda_1}^\top \\ \vdots \\ \mathbf{y}_{\Lambda_K}^\top \end{bmatrix} \in \mathbb{R}^A, \quad (11)$$

i.e., the set of individual scalars is identical to the entries in the original target vector $\boldsymbol{y}$. The predictive performance of the system can be assessed through the following metrics:

- Pearson Correlation Coefficient
- Mean Average Error of prediction (MAE)
- Classification Accuracy for Quantile Prediction

The first two metrics are evaluated with the vectors $\overline{\mathbf{y}}_\Lambda$ and $\mathbf{y}_\Lambda$ from all folds. In the quantile-based assessment we independently bin the true values $\mathbf{y}_\Lambda$ and the predicted values $\overline{\mathbf{y}}_\Lambda$ into C different quantiles. Each individual true and predicted value is assigned to a quantile $c_j \in \{c_1, \ldots, c_C\}$. These assignments can be used to visually compare results on a heat-map or as regular evaluation scores in terms of accuracy.

2.5.1 Ridge-Weight Aggregation

For the final prediction model, the regression weights $\boldsymbol{\omega}_k^\star$ from Ridge regression are averaged over the K folds, i.e. $\overline{\boldsymbol{\omega}} = \frac{1}{K} \sum_{k=1}^K \boldsymbol{\omega}_k^\star$.

For every sentence embedding $\mathbf{x}_q$, the prediction is computed as $\overline{y}_q = \mathbf{x}_q^\top \overline{\boldsymbol{\omega}} \in \mathbb{R}$.

2.6 Interpretation Subsystem: Cluster Ranking

We employ predefined textual topic clusters—which are independent of any target values—in order to enable interpretation of the textual correlates. Each cluster is a collection of sentences and should, intuitively, be interpretable as a topic, e.g. separate topics about indoor and outdoor activities as shown in Fig. 4. For each cluster m a ranking score can be computed with respect to a linear prediction model $\overline{\boldsymbol{\omega}}$ such as defined above. Let $\mathcal{Q}_m = \{q : \zeta(q) = m \ \wedge \ q \ \in \mathcal{Q}\}$ be the set of sentences assigned to cluster m. The score ι_m for the cluster m is the average of all predictions $\overline{y}_q = \mathbf{x}_q^\top \overline{\boldsymbol{\omega}}$ within the cluster m:

$$\iota_m = \frac{1}{|\mathcal{Q}_m|} \sum_{\overline{y}_q :\ q \in \mathcal{Q}_m} \overline{y}_q \quad (12)$$

By ordering the scores ι_m of all clusters, we obtain the final ranking sequence of all clusters, with respect to the target-specific model $\overline{\boldsymbol{\omega}}$.

Clustering Preprocessing. For obtaining the fixed clustering, as $\mathbf{X}$ is a very large matrix, clustering might require subsampling to reduce computational complexity. Hence, Q out of the S embeddings in $\mathcal{S}$ are randomly subsampled into the set $\mathcal{Q}$. The mapping $\Phi(Q) = [\phi(1), \ldots, \phi(Q)]^\top$

is a uniformly random selection of row indexes in $\mathbf{X}$ out of $\binom{N}{Q}$. We define the subsampled data matrix as $\mathbf{X}_Q = \left[\mathbf{x}_{\phi(1)}, \ldots, \mathbf{x}_{\phi(Q)}\right]^\top \in \mathbb{R}^{Q \times D}$.

The subset $\mathbf{X}_Q$ is clustered with the Yinyang K-Means algorithm (Ding et al., 2015). We use M centroids and the cosine similarity as a distance function. The cluster assignment vector $\mathbf{M} \in [1, \ldots, M]^\top$ assigns one cluster for each embedding in $\mathbf{X}_Q$. Accordingly, the operator $\zeta : \{1, \ldots, Q\} \rightarrow \{1, \ldots, M\}$ indicates the assigned cluster m for a given sentence s in $\mathcal{Q}$ (see cluster ranking above). The cluster centers are defined in $\mathbf{M}_Q \in \mathbb{R}^{Q \times D}$.

3 Data sources

We apply the method described in Section 2 to the following setting: The pool of sentences $\mathcal{S}$ consists of geotagged Tweets. The assigned locations are in the United States. The geotags are categorized into US-counties which represent the set of communities $\mathcal{A}$. The target variables $\mathbf{y}$ are health-related variables, for example normalized mortality or prevalence rates. We focus on cancer and AHD mortality as well as on diabetes prevalence. Hence, the quantile-based predictions give a categorization of the Ridge regression predictions on a US-county level. The ranked topics assess what language might relate to higher or lower rates of the corresponding disease. Table 1 provides an overview of the size of the data sources, the year the data was collected in and the mean μ and standard deviation σ of the target variables. Not all counties are covered in the publicly available datasets, usually being limited to more populous counties. The collected Tweets are from 2014 and 2015. The target variables are the union-averaged values from 2014 and 2015: if the target variable is available for both years the two values are averaged. Conversely, if a county data point is only available for one, but not both years, we use this standalone value.

3.1 Datorium Tweets

Tweets are short messages of no more than 140 characters[1] published by users of the Twitter platform. They reflect discussions, thoughts and activities of its users. We use a dataset of approximately 144 million tweets collected from first of June 2014 to first of June 2015 (Datorium, 2017).

[1]Twitter increased the limit to 280 characters in 2017, which doesn't affect our data.

Name	**# tweets**	**Year**	
Datorium	147M	14/15	
Name	**# counties**	**Year**	$\boldsymbol{\mu}$, $\boldsymbol{\sigma}$
AHD	803	14/15	43.0, 16.1
Diabetes	3129	13	9.7, 2.2
Breast	487	13/14	12.4, 2.8
Colon	490	13/14	12.1, 3.0
Liver	293	13/14	7.5, 2.4
Lung	1612	13/14	52.4, 16.2
Melanoma	162	13/14	3.8, 1.2
Prostate	351	13/14	8.5, 2.0
Stomach	136	13/14	3.6, 0.9

Table 1: Overview of data sources.

Each tweet was geotagged by the submitting user with exact GPS coordinates and all tweets are from within the US, allowing accurate county-level mapping of individual tweets.

3.2 AHD & Cancer Mortality

Our source of the statistical county-level target variables is the CDC WONDER[2] database (CDC, 2018) for AHD and cancer. Values are given as deaths per capita (100'000).

3.3 Diabetes Prevalence

We use county-wise age-adjusted diabetes prevalence data from the year 2013 (CDC, 2016), provided as percent of the population afflicted with type II diabetes. The data is available for almost all the 3144 US counties, making it a valuable target to use.

4 Results

The results of our method for the various target variables are listed in Table 2 along with the performance of the baseline model outlined in Section 4.1. We provide the Pearson correlation ($\boldsymbol{\rho}$) and the mean absolute error (MAE) of our system along with the baseline model's Pearson correlation.

4.1 LDA Baseline Model

We reimplemented the approach proposed by Eichstaedt et al. (2015) as a baseline for comparison, and were able to reproduce their findings about AHD with recent data: similar results were

[2]US Centers for Disease Control and Prevention - Wide-ranging Online Data for Epidemiologic Research.

found with the Datorium Twitter dataset (Datorium, 2017) and CDC AHD data from 2014 and 2015. Their approach averages topics generated with Latent Dirichlet Allocation (LDA) of tweets per county as features for Ridge regression. We do not use any hand-curated emotion-specific dictionaries, as these did not impact performance in our experiments. We used the predefined *Facebook* LDA coefficients of Eichstaedt et al. (2015), updated them with the word frequencies of our collected Twitter data (Datorium, 2017). Our results are computed with a 10-fold cross-validation and without any feature selection.

Type	ρ	ρ *LDA*	*MAE*
AHD	**0.46**	0.31	13.4
Diabetes	**0.73**	0.72	1.1
Breast	**0.44**	0.42	1.80
Colon	**0.55**	0.51	1.87
Liver	0.29	**0.40**	1.59
Lung	**0.68**	0.63	8.44
Melanoma	**0.72**	0.61	0.68
Prostate	**0.39**	0.38	1.34
Stomach	0.44	**0.51**	0.72

Table 2: Results of predictions on different health targets. ρ: our system (Section 2.5), ρ LDA: topic model baseline (Eichstaedt et al. (2015), Section 4.1), MAE: mean absolute error of our system (Section 2.5).

4.2 Detailed Results

In this section we discuss a selection of our results in detail, with additional information available in Appendix A.1.

Diabetes has a strong demographic bias, with a higher prevalence in the south-east of the US, the so called *diabetes belt*. Compared to the national average, the african-american population in the diabetes belt has a higher risk of diabetes by a factor of more than 2 (Barker et al., 2011) and the south-east of the US has a large african-american population. Therefore, linguistic features (Green, 2002) common in african-american are a strong predictor of diabetes rates. The model learns these linguistic features, as seen in Figure 3, and its predictions closely match the actual geographic distribution, as seen in Figure 2. A moderate alcohol consumption is linked to a low risk of type II diabetes compared to no or excessive consumption (Koppes et al., 2005). The strongest negatively correlated word clouds in Figure 3 support this finding.

The most positively related word clouds for melanoma in Figure 4 are related to outdoor activities (Elwood et al., 1985). Conversely, the strongest negatively correlated word clouds suggest indoor activity related language.

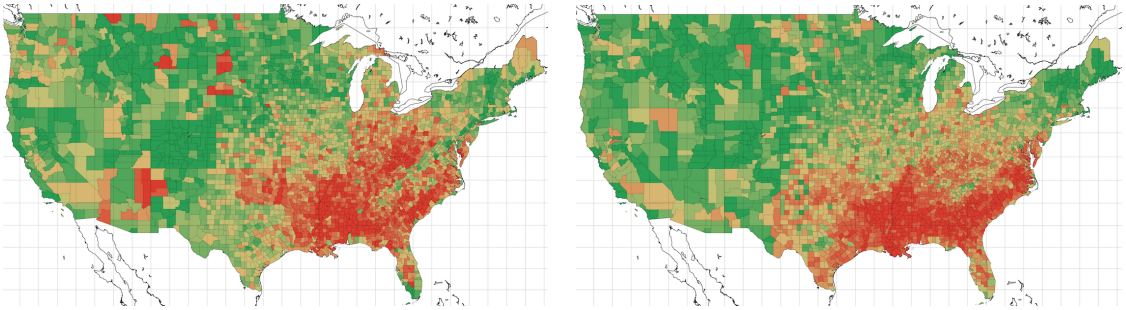
5 Discussion

In this paper, we introduced a novel approach for language-based predictions and correlation of community-level health variables. For various health-related demographic variables, our approach outperforms in most cases (Table 2) similar models based on traditional demographic data by using only geolocated tweets. Our approach provides a method for discovering novel correlations between open-vocabulary topics and health variables, allowing researchers to discover yet unknown contributing factors based on large collections of data with minimal effort.

Our findings, when applying our method to AHD risk, diabetes prevalence and the risk of various types of cancers, using geolocated tweets from the US only, show that a large variety of health-related variables can be predicted with surprisingly high precision based solely on social media data. Furthermore, we show that our model identifies known and novel risk or protective factors in the form of topics. Both aspects are of interest to researchers and policy makers. Our model proved to be robust for the majority of targets it was applied to.

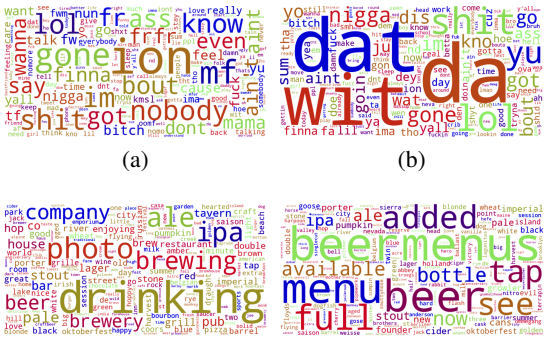
For AHD risk, we show that our approach significantly outperforms previous models based on topic models such as LDA or traditional statistical models (Eichstaedt et al., 2015), achieving a ρ-value of 0.46, an increase of 0.09 over previous approaches. For diabetes prevalence our model correctly predicts its geographic distribution by identifying linguistic features common in high-prevalence areas among other features, with a ρ-value of 0.73. For melanoma risk, it finds a high-correlation with the popularity of outdoor activities, corresponding to exposure to sunlight being one of the main risk factors in skin cancer, with an overall ρ-value of 0.72.

One of the main limitations of our approach is the need for a large collection of sentences for each community as well as a large number of communities with target variables, leading to potentially unreliable results when this is not the case, such as for social media posts by individuals



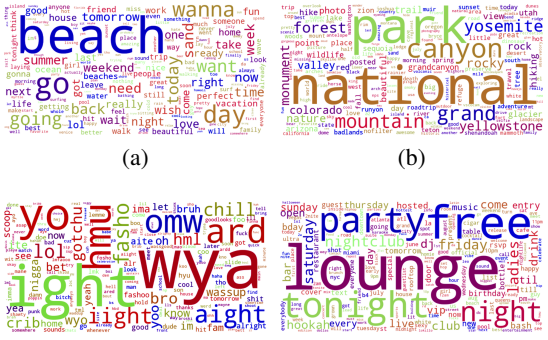
(a) (b)

Figure 2: Quantiles of the prevalence of **diabetes**. (a) Target values (b) Predicted values from tweets



(a) (b) (c) (d)

Figure 3: Word clouds of topics correlating with **diabetes**: (a) (b) strongest positive correlation (c) (d) strongest negative correlation among $M = 2000$ clusters.



(a) (b) (c) (d)

Figure 4: Word clouds of topics correlating with **melanoma**: (a) (b) strongest positive correlation (c) (d) strongest negative correlation among $M = 2000$ clusters.

or when modeling target values which are only available in e.g. few counties. Further research is needed to ascertain whether significant results can also be achieved in such scenarios, and if robustness of our approach is improved compared to bag-of-words-based baselines (Eichstaedt et al., 2015; Brown and Coyne, 2018; Schwartz et al., 2018). Furthermore, all mentioned approaches rely on *correlation*, and thus do not provide a way to determine any *causation*, or ruling out of potential underlying factors not captured by the model. Even though using social media data introduces a non-negligible bias towards users of social media, our approach was able to predict target variables tied to very different age-groups, which is encouraging and supports the robustness of our approach.

Our method captures language features on a community scale. This raises the question of how these findings can be translated to the individual person. Theoretically, a community-based model as described above could be used to rank social media posts or messages of an individual user, with respect to specific health risks. However, as we currently do not have ground truth values on the individual level, and since user's social media history has very high variance, this is left for future investigation.

Future research should also address the applicability of our model to textual data other than Twitter and potentially from non-social media sources, to communities that are not geography based, to the time evolution of topics and health/lifestyle statistics, as well as to targets that are not health related. The general methodology offers promise for new avenues for data-driven discovery in fields such as medicine, sociology and psychology.

Acknowledgements. We would like to thank Ahmed Kulovic and Maxime Delisle for valuable input and discussions.

References

Lawrence E. Barker, Karen A. Kirtland, Edward W. Gregg, Linda S. Geiss, and Theodore J. Thompson. 2011. Geographic distribution of diagnosed diabetes in the us: a diabetes belt. *American journal of preventive medicine*, 40(4):434–439.

Nicholas JL. Brown and James C. Coyne. 2018. Does Twitter language reliably predict heart disease? a commentary on eichstaedt et al.(2015a). *PeerJ*, 6:e5656.

CDC. 2016. County data. *National Center for Chronic Disease Prevention and Health Promotion, Division of Diabetes Translation*.

CDC. 2018. CDC WONDER. *WONDER – Wide-ranging Online Data for Epidemiologic Research*.

Brenda Curtis, Salvatore Giorgi, Anneke EK. Buffone, Lyle H. Ungar, Robert D. Ashford, Jessie Hemmons, Dan Summers, Casey Hamilton, and H. Andrew Schwartz. 2018. Can Twitter be used to predict county excessive alcohol consumption rates? *PloS one*, 13(4):e0194290.

Datorium. Geotagged Twitter posts from the united states: A tweet collection to investigate representativeness [online]. 2017.

Yufei Ding, Yue Zhao, Xipeng Shen, Madanlal Musuvathi, and Todd Mytkowicz. 2015. Yinyang K-means: A drop-in replacement of the classic K-means with consistent speedup. In *ICML'15 - Proceedings of the 32nd International Conference on International Conference on Machine Learning*.

Johannes C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, et al. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169.

J. Mark Elwood, Richard P. Gallagher, GB. Hill, and JCG. Pearson. 1985. Cutaneous melanoma in relation to intermittent and constant sun exposurethe western canada melanoma study. *International journal of cancer*, 35(4):427–433.

Lisa J. Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Lando LJ. Koppes, Jacqueline M. Dekker, Henk FJ. Hendriks, Lex M. Bouter, and Robert J. Heine. 2005. Moderate alcohol consumption lowers the risk of type 2 diabetes: a meta-analysis of prospective observational studies. *Diabetes care*, 28(3):719–725.

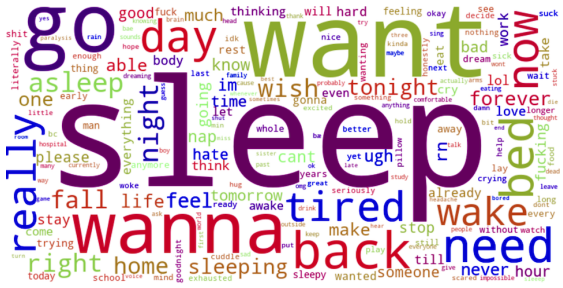
Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Marcel Salathé. 2018. Digital epidemiology: what is it, and where is it going? *Life sciences, society and policy*, 14(1):1.

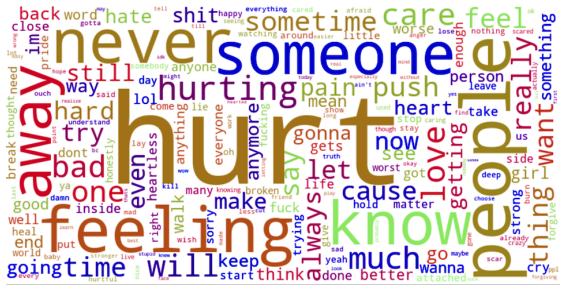
H. Andrew Schwartz, Salvatore Giorgi, Margaret L. Kern, Gregory Park, Maarten Sap, Darwin R. Labarthe, Emily E. Larson, Martin Seligman, Lyle H. Ungar, et al. 2018. More evidence that Twitter language predicts heart disease: a response and replication.

A Appendices

A.1 Additional Figures

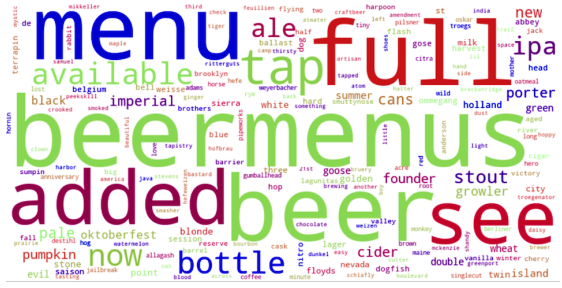


(a)



(b)

(c)



(d)

Figure 5: Word clouds of topics correlating with **colorectal cancer**: (a) (b)strongest positively correlated topics (c) (d) strongest negatively correlated topics among $M = 2000$ clusters.

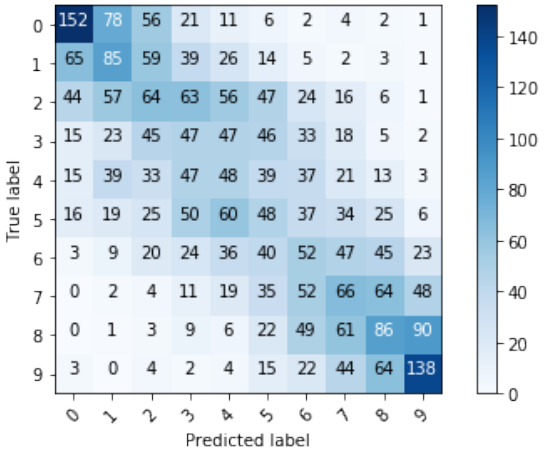


Figure 6: Confusion matrix for decile-based prediction of **diabetes** prevalence.

A.2 Implementation Details

Tweets were collected according to the provided datorium IDs using the Tweepy[3] library. The tweets were then imported into Google BigQuery[4] and processed using Apache Beam[5]. The sentence embeddings were computed using the official Sent2Vec source code and the provided 700-dimensional pre-trained model for tweets (using bigrams)[6]. Clustering was performed by libKMCUDA[7]. Scikit-learn[8] was used for 10-fold cross validation, Ridge regression, calculating the correlation and hyperparameter search.

[3] https://www.tweepy.org/
[4] https://cloud.google.com/bigquery/
[5] https://beam.apache.org/
[6] https://github.com/epfml/sent2vec
[7] https://github.com/src-d/kmcuda
[8] https://scikit-learn.org/stable/