Vol.2017-DBS-165 No.27
Vol.2017-IFAT-128 No.27
2017/9/20

ソーシャルメディアを用いた 依存症者の発言分類とその空間分析

村山 太一[1]　若宮 翔子[1]　荒牧 英治[1]

概要：これまで，都道府県より小さな単位での疫学調査は困難であった．このため，ニコチン依存症など環境の影響を受ける疾患の場合，治療において参考となる統計指標が存在しないことが多かった．本研究では，ニコチン，アルコール，ギャンブルという 3 つの依存症を題材に，ソーシャルメディアから依存症者の所在地の詳細な空間的情報を抽出し，これを可視化することを目指す．まず，ソーシャルメディア上の依存症者と非依存症者の発言を分類するために，クラウドソーシングから得られた擬似ツイートデータを学習データとし，Recurrent Convolutional Neural Network を用いて分類器を構築する．次に，構築した分類器を用いて，位置情報付きツイート 12,237 件を依存症者と非依存症者の発言に分類し，地図上にマッピングしてそれぞれの地理的分布を可視化する．ニコチン依存症を対象にした実験では，喫煙者の中でもニコチン依存症者の発言は駅を中心とした繁華街に集まる傾向があるなど，直感的に妥当な結果が得られた．今後，この可視化の妥当性の詳細な検討や活用を行う予定である．

キーワード：疫学調査，依存症，ソーシャルメディア，クラウドソーシング，Twitter，自然言語処理

Addiction-related Tweet Classification and Spatial Analysis

Taichi Murayama[1]　Shoko Wakamiya[1]　Eiji Aramaki[1]

Abstract: The epidemiological survey is conducted for each prefecture and much more detailed surveys are not done. Addiction in the epidemiology largely depends on environment. It is important for addiction treatment to figure out the relationship between addicts and behavioral environment. The purpose of this research is to visualize and analyze addicts' behavior and environment in detail. Firstly, we construct a classifier based on Recurrent Convolutional Neural Network from quasi-tweet data obtained through crowdsourcing, to distinguish addicts' tweets from non-addicts' tweets in social media. Next, we visualize a geographic distribution of the classified tweets with location information on a map. The geographic distribution of the tweets by nicotine addicts suggests a pattern that nicotine addicts tend to gather in busy streets centered on stations than non-nicotine addicts. In future work, we plan to analyze individual tweets on a long term basis for figuring out characteristic patterns of addicts' behavior.

Keywords: Epidemiological study, Addiction, Social media, Crowdsourcing, Twitter, Natural Language Processing

1. はじめに

依存症（一般には，**中毒**と呼ばれることもあるが，本稿では**依存症**と呼ぶ）は，物質的に豊かである先進国に共通した大きな社会問題の一つである．例えば，アルコール依存症者は日本で約 58 万人 [1]，ニコチン依存症者は喫煙者の 68.6 %の約 1,487 万人 [2] にのぼると考えられており，多くの日本人が何かしらの物質の依存症になっているといえる．このような依存症に対する調査は集団を対象とする疫学の手法を取られており，都道府県単位で依存症者の実態を調査するのが一般的である．しかし，依存症は行動や環境といった要素との関連性が強いことから，一般的な疫学調査以上に，より詳細な空間分析を行い，依存症者の行

[1] 奈良先端科学技術大学院大学

Vol.2017-DBS-165 No.27
Vol.2017-IFAT-128 No.27
2017/9/20

動と環境との関係を解き明かすことが重要であると考えられる．

今回，従来の統計手法では取得できなかったより詳細な依存症者の情報取得のために，ソーシャルメディアデータを活用する．一般的にソーシャルメディアデータの利用は，人々の発言や振る舞いに関するデータを大規模に取得できることや，リアルタイムな位置情報を取得できることが特徴として挙げられ，これらの利点に着目した研究が多い．本研究では，依存症者の発言の位置情報の可視化を行い，依存症者の行動範囲などの新たな知見の取得を目標とする．

まず，クラウドソーシングで依存症者のテキストデータを取得し，そのデータを元に，依存症者者の発言を分類する分類器の作成を行う．次に，分類器を用いてソーシャルメディアのユーザによる発言を分類し，依存症者の空間的分析を行う．図 1 で本研究の全体図を表す．

本研究の新規性は以下の 2 点である．

(1) クラウドソーシングによる疑似ツイートを用いて，依存症者の発言に対する分類器を構築した点
(2) これまでよく知られていなかった依存症者の地理的分布をソーシャルメディアデータを利用して明らかにした点

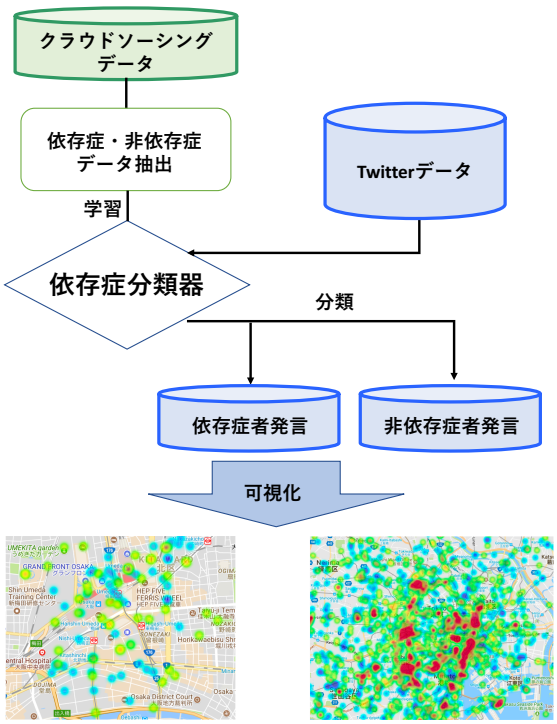

図 1 全体図

2. 関連研究

2.1 様々な依存症

依存症にはニコチン，アルコール，ギャンブルといった様々な種類がある．これまで，依存症者ではなく単なる喫煙者や飲酒者の動向の把握や禁煙啓蒙活動といった目的で，ソーシャルメディアを用いた研究が行われてきた．

ニコチンとソーシャルメディアを扱った研究としては，タバコ関連の発言をジャンル，テーマや感情といった観点で機械学習を用いて分類し，タバコ関連の用語がどのような人々によってどのように用いられるかを分析することで，公衆衛生に関する運動への利用可能性について考察したもの [3] や，ラテン系アメリカ人を対象にソーシャルメディアを用いた禁煙キャンペーンに関する考察を行ったもの [4] などが挙げられる．他にもソーシャルメディアとタバコに関連した多くの研究があり，詳しくはシステマティックレビュー [5] にまとめられている．しかし，喫煙者に焦点を当てた研究は多くあるものの，依存症者に焦点を当てたものは，依存症者と非依存症者の判別が難しいため多くない．

アルコールに関しては，アルコール摂取者の行動の傾向をソーシャルメディアデータから読み取るものなどがある．問題となる飲酒行為のツイートの時間的分析を行ったもの [6] や，インスタグラムの写真から飲み過ぎの判定を行うもの [7] などの研究が挙げられる．しかし，喫煙者に関する研究と同様の理由から，ソーシャルメディアを用いたアルコール依存症に焦点を当てた研究は見られない．

ギャンブルに関しては，ギャンブルが盛んなアメリカではソーシャルメディアを通じた研究が多く行われている．その中でも，ネットギャンブルが最近盛んであることから，ソーシャルメディアの利用とネットギャンブルの関係を分析した研究 [8] などが見られる．

2.2 依存症と疫学

医療分野における依存症研究は，個人の治療に焦点を当てたものが多い．しかし，依存症の治療には，他の依存症者との接触など環境が重要な要因 [9] とされており，対象を集団とみなす大きな視点も重要である．

このように患者を集団で扱う学問としては，**疫学**と呼ばれる分野がある．疫学は，集団を対象に疾病の発生原因や予防などを研究する学問であり，主に感染症が扱われるが，交通事故などの人災，地震や火災といった天災，病気や依存症など，人間の健康を損ねる原因となる対象を幅広く扱っている．疫学において，依存症が扱われることはまだ少ないが，アルコール依存症やニコチン依存症は，行動や環境といった要素との関連性が強いことから，本研究では，疫学と同様に地域ごとの特性を調査する．ただし，一般的な疫学調査で行われる都道府県ごとの調査よりも，より詳細な位置情報の分析を試みる．

3. 手法：依存症／非依存症の分類器構築

本研究は，依存症者の行動分析をソーシャルメディアデータを利用して行うことを目的としている．そのために依存症者かどうかを判別する必要がある．本章では，ソーシャルメディアにおけるユーザの発言を依存症者によるものと非依存症者によるものに分類するための分類器の構築について述べる．

まず，分類器作成用の学習データとして，クラウドソーシングを用いて収集したテキストデータを紹介する (3.1 節)．次に，実装した分類モデルを述べ (3.2 節)，最後に分類器の精度 (3.3 節) とその考察 (3.4 節) を述べる．

3.1 利用データ

依存症者と非依存症者の区別を行いながら，大量のテキストデータを収集するために，クラウドソーシングを利用した．今回は，代表的な 3 つの依存症であるニコチン依存症，アルコール依存症，およびギャンブル依存症に関するテキストデータの収集を行った．調査対象となる依存症に関する行為，つまりニコチン依存症ならば喫煙，アルコール依存症ならば飲酒，ギャンブル依存症ならばギャンブルを日常的に行っている者を調査対象とした．年齢や性別，在住地などの基本情報，依存症かどうかの判定設問，テキストデータを得るための質問をそれぞれ設定した．依存症かどうかの判定に使用する指標を表 1 に示す．

表 1 依存症判定の基準

対象	判定指標	基準点
ニコチン依存症	TDS [10]	5 点以上
アルコール依存症	AUDIT [11]	20 点以上
ギャンブル依存症	SOGS [12]	5 点以上

ニコチン依存症用のテキストデータを収集するために，以下のような質問を設定した．

(1) もし，あなたがツイッターでタバコを楽しく吸っている状況を報告するとすれば，どのように書きますか?
(2) もし，あなたがツイッターでタバコを吸いたくて吸えない状況を報告するとすれば，どのように書きますか?
(3) もし，あなたがツイッターでタバコが大好きだということを主張するなら，どのように書きますか?
(4) もし，あなたがツイッターでタバコの良い点を語るならば，どのように主張しますか?

他の依存症に対しては，上記の下線部を変更した設問を用いる．ここで得られたテキストデータをソーシャルメディア上の個人の擬似的な発言とみなし，学習データを構築する．

Yahoo!クラウドソーシング[*1]で上記の設問を用いて，各依存症につき 1000 人分のテキストデータの取得を行った．なお，チェック設問により，不適切なユーザのフィルタリングを行なった．また，「ツイートしない」や「呟かない」といった発言はノイズとして目視にて削除を行った．その結果，ニコチンのクラウドソーシングでは 2,401 発言，アルコールのクラウドソーシングでは 2,683 発言，ギャンブルのクラウドソーシングでは 1,949 発言をそれぞれ取得した．表 2 に取得した擬似ツイートの統計量を示す．

表 2 発言取得数

	ニコチン	アルコール	ギャンブル
取得発言数	2,401	2,683	1,949
（うち，依存症者発言数）	(1,808)	(243)	(921)

3.2 実装モデル

クラウドソーシングで収集したテキストデータを学習データとして，ソーシャルデータ上の発言から依存症者による発言か非依存症者の発言かを判別する分類モデルを作成する．データの前処理としてワードエンベディングには単語をベクトル化する Word2Vec を用いた．実装モデルには，Recurrent Convolutional Neural Network (RCNN)[13] を隠れ層を 100 次元として利用した．RCNN は，文章中の各単語が位置する場所によってバイアスがかかる Recurrent Neural Network を改良したモデルであり，より文脈を取得しやすいという利点がある．実装は，scikit-learn[*2] と keras[*3] のライブラリを利用して行った．

3.3 結果

実装モデルの精度を確認するために，依存症ごとにテキストデータを 5 分割し交差検証を行った．結果，アルコール依存症分類器の Accuracy は 85.8%，ニコチン依存症分類器の Accuracy は 63.6%，ギャンブル依存症分類器の Accuracy は 53.3%であった（表 3）．

本研究における分類器作成の目的は，ソーシャルメディアデータから依存症者の発言を抽出することであるため，適合率が重要である．そこで，適合率に着目すると，アルコール依存症に関しては 24.3%，ニコチン依存症に関しては 77.6%，ギャンブル依存症に関しては 43.0%であった．この結果から，アルコール依存症とギャンブル依存症では，依存症者の発言の識別は困難であるが，ニコチン依存症に対しては実用可能であるといえる．

[*1] https://crowdsourcing.yahoo.co.jp/
[*2] http://scikit-learn.org/stable/
[*3] https://keras.io

Vol.2017-DBS-165 No.27
Vol.2017-IFAT-128 No.27
2017/9/20

表 3 分類精度

対象	Accuacy	適合率	再現率	F1-score
アルコール依存症	85.8%	24.3%	15.5%	0.189
ニコチン依存症	63.6%	77.6%	75.1%	0.763
ギャンブル依存症	53.3%	43.0%	49.4%	0.460

3.4 考察

結果に示した通り，アルコール依存症とギャンブル依存症の分類器は低い Accuracy であった．

アルコール依存症に関しては，他の依存症と比較し，飲酒する人の絶対数が多いため，クラウドソーシングでは依存症者の十分なサンプル数を取得できなかったことが原因であると考えられる．アルコールおよびニコチン依存症は物質的依存症であり，その物質の摂取を行わないとイライラなどの離脱症状を引き起こすためテキストデータにもこれを反映した発言 (1)(2) が多い．

一方で，ギャンブル依存症は物質的依存症と比較し離脱症状が少ないため，テキストデータにも離脱症状を表現するような発言 (3) が少ない．そのため，依存症者と非依存症者の違いが出にくいことが原因だと考えられる．

アルコールとギャンブルに対し，ニコチン依存症に関しては，クラウドソーシングで十分な依存症者のサンプル数を取得でき，離脱症状も明確である．このことから，ニコチン依存症に関しては一定の精度を持つ分類器の作成を行うことができたと考えられる．

それぞれの依存症者の発言例を以下に示す．

(1) 飲みたくてたまらない、酒を盗んでしまうかも...
(2) タバコ吸えないイライラ
(3) あ～パチンコ行きたいけどお金がない～ やりたいことができずイライラする

4. 結果：依存症者の発言分布可視化

前章で作成したニコチン依存症の分類器を用いて，代表的なソーシャルメディアである Twitter 上での位置情報付き発言を，依存症者と非依存症者の発言に分類する．その結果を用いて，地理的分布の可視化を行う．

4.1 利用データ

ニコチン依存症の分類器にかけるテストデータとして，2011 年 7 月～2012 年 7 月に Twitter 上に投稿された位置情報付きツイート 24,817,903 件のうち，タバコに関するキーワード（「タバコ」「たばこ」「煙草」「喫煙」など）が含まれ，チェックインサービス Foursquare[*4] などを用いた位置情報のみの発言を除いた 12,237 ツイートを用いる．

分類器による依存症者の発見が目的であることから，より適合率が高くなるように分類の閾値を 0.95 に設定し，データを分類した．この結果，5,246 ツイートをニコチン依存症者の発言，6,991 ツイートを非依存症者の発言として分類した．

4.2 位置情報に基づいた可視化とその考察

ニコチン依存症者の発言分布を可視化するために Google Map API[*5] を用いて，ヒートマップを描画した．ニコチン依存症者による発言を赤～緑，ニコチンに関連した発言のうち非依存症者による発言を濃青～淡青で表現する．

図 2 は東京都全体におけるニコチン関連の発言分布を可視化した結果，図 3 は大阪府全体における発言分布を可視化した結果である．

駅を中心とした繁華街をより詳細に示した図 2(a) と図 3(a) を見ると，非依存症者，依存症者両者とも繁華街の駅を中心とした場所での発言が多く見られる．一方で，依存症者は非依存症者以上に繁華街の駅の中心部分に集まりやすいように見られる．これは，駅構内や駅周辺の広場に喫煙所が多く見られることから，依存症者が外出時に駅周辺の喫煙所に集まりやすいことを示唆している．

図 2(b) と図 3(b) のような繁華街ではない地域においては，依存症者はまばらに散らばっている．これに対し，非依存症者は特に線路沿いに多く見られるといった特徴が見られる．

5. 考察: 可視化の応用について

これまで，都道府県や市区町村といった大きな単位についての指標（タバコ消費量や禁煙外来受診者数など）は存在したが，それよりも小さい単位での統計についてはよく知られていなかった．このような背景の中，本研究は，これまでよく知られていなかったニコチン依存症者の詳細な位置を得ることができた．この情報の応用を本章で行う．

素朴には，依存症の治療において，対象物や同じ依存症者との距離を取ることが重要であるため，この地点を避けることが効果的である．そこで，ニコチン依存症者の密集地帯を避けるために可視化結果をそのまま用いることが考えられる．そこで，繁華街同士の比較を行う．具体的には，新宿に住むにしても，西側と東側のどちらがよいか，または，新宿と品川がどちらがよいか，といった情報を取得する．

そこで，市区町村まで大きい単位ではなく，数キロメートル程度の単位で可視化情報を数値化することが有効である．様々な数値化の方法が考えられるが，本研究では，依存症者と非依存症者の遭遇する割合を指数化して**分煙指標**として定義し，これを用いて地域間の比較を行う．

分煙指標の定義

依存症の治療において，依存症者との距離が重要である．

[*4] https://ja.foursquare.com/

[*5] https://developers.google.com/maps/

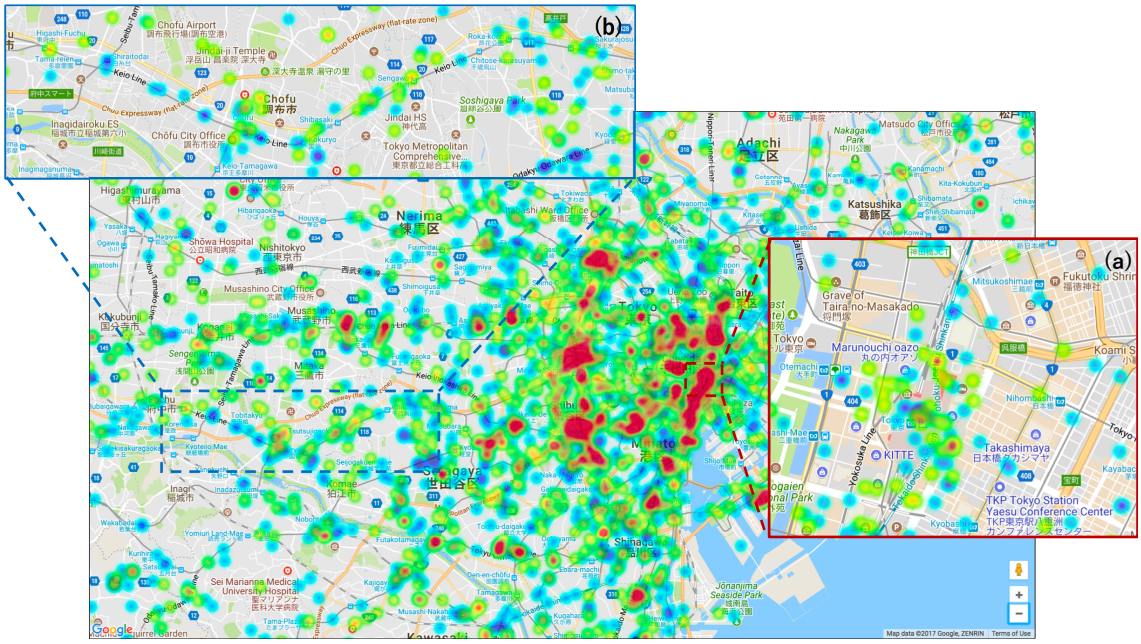

図 2　東京都の可視化．(a) 市街地，(b) 非市街地．

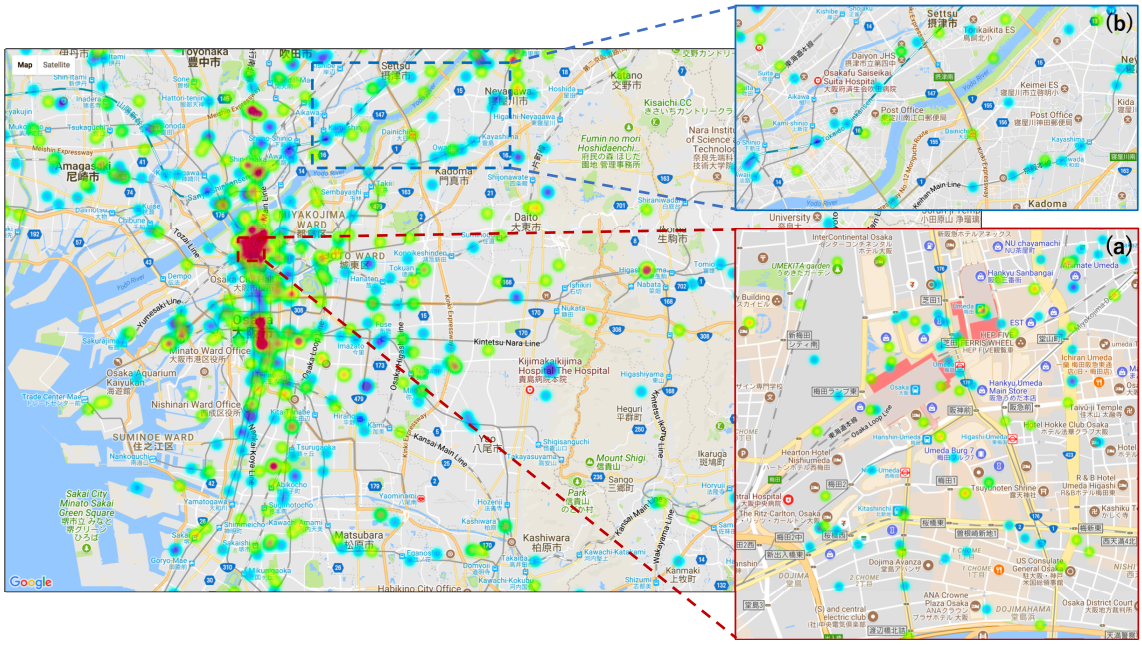

図 3　大阪府の可視化．(a) 市街地，(b) 非市街地．
繁華街では，依存症者の発言は非依存症者の発言分布と比較し，駅の中心部分に集まって見られる．繁華街ではない地域では，依存症者の発言はまばらに散らばっている一方で，非依存症者の発言は特に線路沿いに多く見られる．

そこで，ある地域において，最も近い依存症者への距離の平均 ($MinDist$) を以下の式により求める．

$$MinDist = \frac{\sum_{n \in N} \min_{p \in P} DIST(n,p)}{|N|}$$

ここで，N は非依存症者の集合，P は依存症者の集合，$DIST(n,p)$ は n と p の距離を求める関数である．

ただし，そもそも過疎な地域ではこの値が必然的に大きくなる．そこで，以下のように，依存症者，非依存症者に関わらず，最も近い人間への平均距離で正規化する．

$$NormalizedMinDist = \sum_{n \in N} \frac{\min_{n \in N} DIST(p,n)}{\min_{a \in A} DIST(n,a)}$$

ここで，A は発言者の集合（依存症者と非依存症者の和集合 $N \cap P$），$DIST(n,a)$ は n と a の距離を求める関数（ただし，$n = a$ （距離 0）の場合は除く）．

結果

この分煙指標の数値の大きさは，非依存症者がその地域

を出歩く際に依存症者と出会う確率の低さを指す．東京や関西の駅を中心とした繁華街の分煙指標を示した結果を表 4 に示す．この結果から，渋谷駅は分煙指標が高く，出歩く際に依存症者と出会う確率が低い地域であるといえる．

この指標のように，発言位置をある程度の単位（市区町村よりは小さいが，詳細な発言位置よりも大きな単位）で集計することにより，環境と依存症者の関係をより密に解き明かすことも可能となるだろう．

しかし，図 4 に示した新潟県のような過疎地は，データ数が少なく，可視化によって読み取れる情報が少なくなるといった問題が存在するため，いかにデータ数を増加させるかの検討が必要となる．

同時に，この分煙指標の他にも，依存症者の分布を混合分布とみなし，依存症者の集中しやすい地点を求めるなど，様々な指標化が考えられ，目的やデータ規模に沿った指標のデザインが今後の課題となる．

表 4 駅を中心とした地域の分煙指標

各駅を中心とした各辺 1km 正方形の地域	分煙指標
渋谷駅	5.847
梅田駅	2.791
池袋駅	2.626
新宿駅	2.094
京都駅	2.073
六本木駅	1.921
品川駅	1.798
秋葉原駅	1.707
東京駅	1.675

6. おわりに

本研究は，ソーシャルメディア上の依存症者と非依存症者の発言を分類し，可視化した．まず，クラウドソーシングから得られた擬似ツイートデータを学習データとし，Recurrent Convolutional Neural Network を用いて分類器を構築した．次に，これを用いて，位置情報付きツイート 12,237 件を依存症者と非依存症者の発言に分類し，地図上に可視化した．ニコチン依存症を対象にした実験では，喫煙者の中でもニコチン依存症者の発言は駅を中心とした繁華街に集まる傾向など，直感的にも妥当な結果が得られた．今後は，学習データの改善による分類器の精度の向上と，その指標化が課題である．

謝辞

本研究の一部は，日本医療研究開発機構研究費 新興・再興感染症に対する革新的医薬品等開発推進研究事業（課題番号：16768699），戦略的情報通信研究開発推進事業（SCOPE）（課題番号：17934316），JSPS 科研費 JP16K16057 および JST ACT-I の支援を受けたものです．

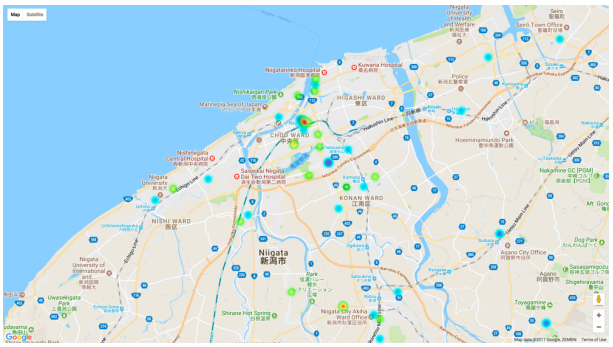

図 4 新潟県の可視化マップ

参考文献

[1] 樋口進：ＷＨＯ世界戦略を踏まえたアルコールの有害使用対策に関する総合的研究*6.
[2] ファイザー株式会社：日本全国の“ニコチン依存度チェック”2014*7.
[3] Mysln M, Zhu SH, Chapman W, Conway M:*Using twitter to examine smoking behavior and perceptions of emerging tobacco products*, J Med Internet Res, 2013.
[4] Anguiano B, Brown-Johnson C, Rosas LG, Pechmann C, Prochaska JJ*Latino Adults' Perspectives on Treating Tobacco Use Via Social Media*, J Med Internet Res, 2017.
[5] Lienemann BA, Unger JB, Cruz TB, Chu KH:*Methods for Coding Tobacco-Related Twitter Data: A Systematic Review*, J Med Internet Res, 2017.
[6] West, J. , Hall, P. , Hanson, C. , Prier, K. , Giraud-Carrier, C. , Neeley, E. and Barnes, M.:*Temporal variability of problem drinking on Twitter*, Open Journal of Preventive Medicine, 2012.
[7] Venkata Rama Kiran Garimella , Abdulrahman Alfayad , Ingmar Weber:*Social Media Image Analysis for Public Health*, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016.
[8] Daniel L. King, Paul H. Delfabbro, Dean Kaptsis, Tara Zwaans:*Adolescent simulated gambling via digital and social media: An emerging problem*, Computers in Human Behavior, 2014.
[9] Okoli, Chizimuzo T.C. et al:*econdhand Tobacco Smoke Exposure and Susceptibility to Smoking, Perceived Addiction, and Psychobehavioral Symptoms among College Students.*,Journal of American college health: J of ACH 64.2, 2016.
[10] Kawakami N, Takatsuka N, Inaba S, Shimizu H:*Development of a screening questionnaire for tobacco/nicotine dependence according to Icd-10, Dsm-III-r, And Dsm-IV*, Addictive Behaviors, 1999.
[11] Babor, Thomas F., et al.:*AUDIT: The alcohol use disorders identification test: Guidelines for use in primary health care.*, WHO, 1992.
[12] 斎藤　学:強迫的 *(*病的*)* 賭博とその治療–病的賭博スクリーニング・テスト *(*修正 *SOGS)* の紹介をかねて., 家族機能研究所研究紀要 1: 10-56., 1997.
[13] Lai, Siwei, et al.:*Recurrent Convolutional Neural Networks for Text Classification*, AAAI(Vol. 333, pp. 2267-2273), 2015.

*6 https://mhlw-grants.niph.go.jp/niph/search/NIDD00.do?resrchNum=201315050A (accessed:2017/08/08)
*7 http://www.pfizer.co.jp/pfizer/company/press/2014/2014_10_31.html (accessed:2017/08/08)