SNSの投稿内容に基づく生活習慣病発症リスクの分析

田渕 尚道　松本 和幸　吉田 稔　西村 良太　北 研二
徳島大学大学院創成科学研究科
c612135001@tokushima-u.ac.jp, {matumoto, mino, kita, nishimura}@is.tokushima-u.ac.jp

概要

国内における死亡原因は，がん・循環器疾患・糖尿病などで全体の約 6 割を占めている.これらの病気はいずれも食事・運動・睡眠などの生活習慣と深い関りがあり，生活習慣病と呼ばれている．生活習慣病には自覚症状がほとんどないため，病気の悪化に気づきにくいという特徴がある．本研究では，より簡易的に生活習慣病を予防するシステムの構築を目指し，ソーシャル・ネットワーキング・サービス(SNS)を用いて，ユーザの投稿内容から食事・運動・体調・精神状態に関するツイートの分析を行った．調査の結果，本研究の成果がユーザの投稿内容から生活習慣病発症の危険因子を検出するシステムの構築につながることが分かった．

1 はじめに

生活習慣病は国内における，死亡者数の約 6 割を占めている[1]．生活習慣病の予防は私たちの健康を守るために非常に重要であるが，生活習慣病の早期発見は難しい．生活習慣病にはほとんど自覚症状がなく，気づいた時には病気が悪化しているケースが少なくない．本研究では，定期的に生活習慣を見直すきっかけを与えるシステムが必要であると考え，多くの人が自身の日常生活について投稿しているSNS に注目する．本論文では，年々ユーザが増加傾向にある Twitter[2]を対象とし，ユーザの投稿内容から食事，運動，体調，精神状態に関するツイートの分析を行い，ユーザの特徴，ユーザ間の共通点の分析手法について述べる．これらの分析結果から，ユーザの投稿内容から生活習慣病発症の危険因子を検出するシステムの構築について考察する．

2 関連研究

高橋ら[3]は，ユーザの生活パターンと介入が受け入れられる可能性を考慮して環境のモデルを推定し，後ろ向き帰納法アルゴリズムで目標から逆算して最適な方策を得る方法を構築している．彼らの研究では，ユーザの生活パターンへの介入をユーザ自身が記録したカレンダーアプリの生活記録データを分析することで行う.そのため,介入を受けるためには，ユーザが毎日の生活を記録する必要がある．また，彼らのシステムは，ある程度生活習慣の改善が必要であることを自覚しているユーザには有効であるが，ユーザへの負担が大きいため,ユーザ数は限られる．本研究では，生活習慣を改善する必要性に気づいていないユーザを対象として生活習慣改善のきっかけを与えるシステムを目指している点で異なる．

Gael Pérez Rodríguezt ら[4]は，Twitter の糖尿病コミュニティの患者や近親者の健康状態を理解するために，糖尿病患者の投稿内容に焦点を当てた分析を行っている．Twitter の投稿内容からユーザの健康状態の分析を行う点では，本研究は彼らの研究と類似している．彼らは糖尿病に罹っているユーザのみを分析対象としているが，本研究は分析対象とするユーザを限定しない点で異なる．

3 提案手法

提案手法は，まず，半年間で健康状態（体調・精神状態）と生活習慣（食事・運動）のそれぞれ 1 項目以上を含むツイートを定期的に行っているユーザを選定し，半年分のツイートを収集する．これらの情報の収集には Twitter API [5]を使用する．

収集したツイートに対して，自動および手動でラベル付けを行う．自動ラベル付けではラベル別にキーワードを設定し，キーワードが含まれるツイートに対してラベルを自動付与する．手動ラベル付けでは，自動ラベル付けでラベルが付与されなかったツイートを対象に，各ユーザからランダムでツイートを抽出し，手動ラベル付け用のデータを作成し，手作業によるラベル付けを行う．モデル作成のためにラベル付けしたデータのテキストに対して，BERT (Bidirectional Encoder Representations from Transfomers)[6]を用いてツイート単位でテキストをベクトル化した．これを用いてマルチラベル分類モデルの作成と評価を行った．最後に，作成したモデルを用いて，収集した全データにラベルを付与し，データ分析を行う．

3.1 Twitter API

Twitter API を利用することで，公式サイトを介さずに，ツイートやタイムラインなどのデータを取得することができる．キーワードや期間を指定してツイートを収集することが可能であり，分析に必要なツイートを効率的に収集することができるため，本研究では Twitter API を用いてデータ収集を行う．

3.2 ラベル付きデータ作成

ラベルの種類はその他(0)，運動(1)，食事(2)，体調(3)，精神状態(4)の 5 種類とした．自動ラベル付けでは，表 1 に示すように各ラベルのキーワードを設定する．キーワードが含まれたツイートには各キーワードに対応するラベルを付与した．その他(0)と精神状態(4)のツイートはキーワードの設定が困難であるため，自動ラベル付けは行わない．

手作業によるラベル付けでは自動ラベル付けでラベルが付与されなかったツイートを対象に，各ユーザからランダムで一定数のツイートを抽出し，得られたデータを手作業ラベル付け用のデータとする．ラベル付与の偏りを抑えるために，同じデータに対して 4 人の作業者が付与を行い，最終的なラベルは多数決で決定する．多数決で決定できない場合は，複数のクラスに属するツイートとして扱う．

表 1　ラベルの種類とキーワード

ラベル	キーワード
その他(0)	
運動(1)	トレーニング・ジョギング・ウォーキング・ランニング・筋トレ・エクササイズ
食事(2)	食事・朝食・昼食・夜食・ランチ・ディナー・ご飯
体調(3)	体調・健康状態
精神状態(4)	

3.3 テキストのベクトル化

東北大学乾研究室が公開している BERT の事前学習済みのモデル[7]を用いて，ツイートの分散表現ベクトルを取得する．このモデルは日本語の Wikipedia 記事を対象に事前学習したものであり，得られる分散表現ベクトルの次元数は 768 次元である．

3.4 モデル作成と評価

モデルの作成は BERT モデルをサポートしている Simple Transformers[8]を用いた．Simple Transformers は Transformer モデルが，シンプルな実装で使えるライブラリである．今回のモデルの構造は BERT の出力次元がラベルの種類数である 5 になるように最終層だけ全結合層に取り換えたものである．また，モデルの評価には 5 分割交差検証法を用いる．

3.5 データ分析

データ分析では，まずデータ分析の対象ユーザを抽出した．ラベル(0)のツイートは特徴を把握することが難しいため分析対象外とする．データ分析の対象ユーザを 1 日当たり 10 件以上のツイート，且つ，その 10 件以上のツイート内に複数のラベルがあるユーザとした．これは 1 日当たりのラベル出現割合のパターンでユーザを比較するためである．

分析対象のユーザを抽出後，ユーザごとに 1 日当たりのラベルの出現割合を出し，K-Means 法[9]により出現割合のパターンで適当なクラス数にユーザを分類し，各クラス内の単語の出現傾向を調べる．分類したクラス内の単語傾向を調べる手法はトピックモデルを用いる．トピックモデルの学習は LDA(Latent Dirichlet Allocation)[10]により行う．

4 結果と考察

4.1 データ収集

Twitter API を用いて 3 節で述べた条件に当てはまるユーザの選定を行う．まず，食事・運動のキーワードを設定し，食事・運動関連のツイートをしているユーザを特定し，過去に体調・精神状態に関するツイートを投稿したユーザの特定を行う．表 2 が収集時のキーワードの例である．最後に，特定したユーザの中で bot の可能性があるユーザ，鍵付きアカウントのユーザ，半年以内に作られたユーザを除外することで，収集対象のユーザを決定する．ユーザを選定した後，各ユーザの半年間分のツイートを収集したところ，ユーザ数は 5297 名となり，収集できたツイート数は約 276 万件となった．

表 2　ツイート収集時のキーワード

項目	キーワード
食事・運動	食べ過ぎ・食う・食べる・食べた・めし・飯・ランチ・ディナー・運動・トレーニング・エクササイズ・ジョギング・ランニング・ウォーキング・走る・走った・走って
体調･精神状態	疲労・疲れ・倦怠・寝れない・体調・具合・調子・健康状態

4.2 ラベル付きデータ

ラベル付きデータの作成は，自動ラベル付けと手動ラベル付けで行った．自動ラベル付けでは運動，食事，体調の 3 つの項目に対してキーワードを設定し，キーワードを含むツイートには自動でラベル付けを行った．手動ラベル付けでは自動ラベル付けでラベルが付与されなかったツイートを対象に，各ユーザから 2 件のツイートを抽出し，得られた 10,594 件のツイートを手作業ラベル付け用のデータとした．ラベル付与の偏りを抑えるために，同じデータに対して 4 人でラベル付けを行い，最終的なラベルは多数決で決めた．ラベル割合が 1:1 に分かれる場合は重複して学習させ，1:1:1:1 に分かれる場合は除外した．ラベル付けの結果は表 3 のようになった．

表 3　ラベル付け結果

ラベル	自動	手動
その他(0)	-	15,616
運動(1)	38,297	388
食事(2)	36,255	780
体調(3)	8,936	770
精神状態(4)	-	1,596

4.3 モデル作成と評価

モデルに学習させるラベル付きデータは自動ラベル付けデータを各項目ランダムに 700 件抽出したデータと手動ラベルデータで構成されている(表 4).

モデルの作成では Simple Transformers を用いてマルチクラス分類モデルの作成をした．モデルの評価は Python の機械学習用ライブラリである scikit-learn [11]を用いた交差検証法で行った．今回は 5 分割の交差検証を行った．また，学習データのクラスに偏りがあるため，学習の際にアンダーサンプリング，オーバーサンプリングの処理を加えた．アンダーサンプリング処理を加えたモデルの分類精度は 77%，オーバーサンプリング処理を加えたモデルは分類精度が 87%となった．今回，分類精度が高かったオーバーサンプリング処理を加えたモデルを分析に用いることにした．

表 4　学習データの内訳

ラベル	件数
その他(0)	15,616
運動(1)	1,088
食事(2)	1,480
体調(3)	1,470
精神状態(4)	1,596

4.4 分析結果

データ分析の対象ユーザを 1 日当たり 10 件以上のツイート，かつ，その 10 件以上のツイート内に(0)ラベルを除く複数のラベルがあるユーザとする．その結果，対象ユーザは 64 名となった．各ユーザに対して，1 日当たりの各ラベルの出現割合を求める．各ラベルの出現割合は以下の式(1)で算出した．

$$R_i = 100 \times \frac{L_i}{L_s} \quad (1)$$

R_i :ラベルiの出現割合
L_i :ラベルiの出現回数
L_s : 全ラベルの出現回数の合計

ユーザのクラス分類では K-Means 法を用いて，ラベルの出現割合のパターンでユーザを 5 つのクラスに分類した．クラス名はクラス内での出現割合の高いラベルから参照し，A:[運動・食事]，B:[食事・精神状態]，C:[食事]，D:[体調]，E:[精神状態]とした．

各クラスに対し，トピック数を 30 とし，名詞，形容詞のみを対象にトピックモデル(LDA)と pyLDAvis[12]を用いてトピック内の単語傾向やトピック間の共通点を調査した. pyLDAvis の散布図において，円の中心に配置されている数字は各トピックの番号を表し，円の大きさは，分析対象としている単語が各トピックにどの程度含まれているかを表す．また，円の中心間の距離はトピック間の距離を表している．棒グラフは各トピック内の出現頻度の高い上位 30 語と各単語の出現数を示す. 図 1 はクラス B の「眠い」の出現分布を表している．図 2 は図 1 のトピック 26 内の上位単語を表している. 棒グラフの赤色部はトピック 26 内での各単語の単語数を示し，青色部は全体の単語数を示している．また，各クラス内の上位頻出単語 5 語を表 5 にまとめた．

図 1　pyLDAvis によるトピックモデルの可視化

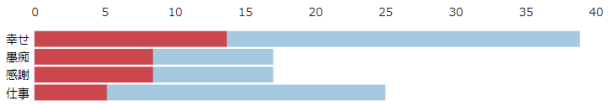

図 2　pyLDAvis による単語の出現頻度の可視化

表 5　各クラス内の上位頻出単語

クラス名	上位単語
A(運動・食事)	ダイエット・お腹・運動・効果・トレーニング
B(食事・精神状態)	幸せ・美味しい・ビーマルワン・運動・健康
C(食事)	ダイエット・運動・多い・カロリー・糖質制限
D(体調)	下剤・体重・痛い・悪い・笑顔
E(精神状態)	嬉しい・仕事・幸せ・最高・楽しい

4.5 考察

pyLDAvis によるトピックモデルの可視化結果に基づき，クラスごとに考察を行う．クラス A では，「ダイエット」が全トピック，「糖質制限」が 16 トピック，「良い」が 26 トピックに分布していた．ダイエット目的でのトレーニングや食事制限関連の単語が多く，健康意識が高いユーザが多いと考えられる．クラス B では，図 1，図 2 から分かるように「眠い」が 14 トピック，その他には「怖い」が 9 トピック，「愚痴」が 10 トピックに分布していた．クラス B は A，C，E と比べ，ネガティブな単語の分布が広いことから生活習慣に何らかの問題を抱えているユーザが多いと考えられる．クラス C では上位頻出単語以外に「アップ」が 12 トピック，「良い」が 26 トピック，「楽しい」が 18 トピックに分布していた．ポジティブな単語の分布が多く，食事を中心とした健康意識の高いユーザが多いと考えられる．

クラス D では「下剤」「悪い」が 29 トピック，「痛い」が 28 トピックに分布していた．その他にも「無理」「腹痛」などのネガティブな単語が広く分布し，他のクラスと比較し，顕著にネガティブな単語が広く分布していた．さらに「お菓子」「アイス」「ラーメン」といった食生活に悪影響を及ぼす可能性のある単語も確認できた．これらを考慮すると，クラス D 内のユーザは食生活の乱れから体調不良を起こしている可能性があり，生活習慣病のリスクが最も高いユーザが多く含まれると考えられる．クラス E では，「嬉しい」が 23 トピック，「最高」が 20 トピックに分布し，ポジティブな単語が広く分布している一方で，「嫌」が 15 トピック，「怖い」が 10 トピック，その他にも「メンタル」「我慢」「必死」といったネガティブな単語も広く分布していた．また，他のクラスには無かった「仕事」が 17 トピックに分布している点も特徴的であった．以上を考慮すると，仕事などのストレスによって今後生活習慣病を発症する可能性のあるユーザが多いと考えられる．これらの考察から 5 つのクラスを生活習慣病発症リスクの観点から比較すると、[A・C]<[E]<[B]<[D]の順で生活習慣病発症リスクは高くなるのではないかと考えられる．

5 おわりに

本研究では，Twitter を対象として，ユーザの投稿内容から食事，運動，体調，精神状態に関するツイートの分析を行い，ユーザの特徴やユーザ間の共通点の調査を行った．

その結果，ユーザをトピックモデルによって適当なクラスに分類することで，クラスごとのユーザの特徴を把握することが可能であることが分かった．また，生活習慣病発症のリスクという観点から結果を考察することによって，各クラスの生活習慣病リスクの比較を行うことが可能であった．この結果は生活習慣病発症の危険因子を検出するシステムの構築へつながると考えられる．

今後の課題として，ツイート収集時やラベル付与時のキーワードの見直しと分類モデルの精度向上を行う．また，体調の変化を表す語とネガティブやポジティブワードの共起頻度や生活習慣病を発症するリスク表現をあらかじめ決定しておき，そのような表現がユーザごとにどのように表れているのかについての調査を行う．

謝辞

本研究は公益財団法人 JKA 令和 3 年度複数年研究補助事業により実施されました. 深く謝意を表します.

参考文献

[1] 生活習慣病を知ろう!(引用日:202 年 12 月 24 日) https://www.smartlife.mhlw.go.jp/event/disease/

[2] 【2021 年 12 月版】人気ソーシャルメディアのユーザ数まとめ (引用日:2021 年 12 月 24 日)https://www.comnico.jp/we-love-social/sns-users

[3] 高橋公海，辛島匡宏，倉島健，戸田浩之: 強化学習を用いた健康行動促進における介入戦略の学習，第 12 回データ工学と情報マネジメントに関するフォーラム，2020 年 3 月 2 日，I2-3，day1 p59

[4] Mining the sociome for Health Informatics: Analysis of therapeutic lifestyle adherence of diabetic patients in Twitter, Gael Pérez Rodríguez, Martín Pérez Pérez, Florentino Fdez-Riverola, Anália Maria Garcia Lourenço, Future Generation Computer Systems, Volume110, September 2020, Pages214-232

[5] Twitter API https://developer.twitter.com/en/docs/twitter-api

[6] BERT: Pre-training of Deep Bidirectional Transformers for Language, Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp.4171-4186, 2019.

[7] BERT 日本語学習済みモデル 東北大学 乾・鈴木研究室 https://github.com/cl-tohoku/bert-japanese

[8] Simple Transformers: https://simpletransformers.ai/

[9] sklearn.cluster.KMeans: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[10] models.ldamodel-Latent Dirichlet Allocation-gensim: https://radimrehurek.com/gensim/models/ldamodel.html

[11] scikit-learn: https://scikit-learn.org/stable/

[12] pyLDAvis: https://github.com/bmabey/pyLDAvis