

ツイート内の発言による夜間活動休止時間の推定に関する研究

児嶋 拓己[†] 吉田 稔[‡] 松本 和幸[‡] 北 研二[‡]

徳島大学大学院先端技術科学教育部[†]

徳島大学大学院社会産業理工学研究部[‡]

1. はじめに

近年の情報処理技術やネットワーク技術の発達に伴いスマートフォンやタブレット端末などが普及してきている。また、これらの普及に伴い LINE や Twitter, Facebook などといった SNS も発達している。中でも Twitter はユーザが現在行っていることや、それに対して感じたことを投稿する傾向が強くリアルタイム性に優れている、かつアクティブアカウントが世界では 3 億 3300 万人、国内では 4500 万人超存在している。またユーザ同士のコミュニケーションも容易に行うことができる。

そんな中、インターネットや SNS の普及、日本人の睡眠教育の意識の低さなどが原因となっている「日本人の睡眠時間の減少」が問題として挙げられている[1]。そこで、Twitter を利用し睡眠時間を推定する手法を提案することによって、Twitter ユーザの生活習慣・睡眠不足の改善を行うことができると思われる。Twitter の投稿から正確な睡眠時間を取得するのが困難なため、就寝時と起床時のツイートの投稿時間の差を夜間活動休止時間とし、睡眠時間と近似していると考えこの時間を推定する手法を提案する。

2. 関連研究

田中ら[2]は、マイクロブログの投稿内容だけでなく各単位時間の投稿数に基づきユーザをクラスタリングし、投稿内容、生活習慣と投稿時間帯から職業属性を推定する手法を提案している。同じ職業でもライフスタイルによっていくつかのグループに分類されることが想定されているため、「同じ職業でも多様なライフスタイルが存在する問題」への対応も検討されている。職業推定の実験の結果から、特徴的なライフスタイルのユーザに対しては、推定の精度の向上

が見られている。しかし、ライフスタイルに顕著な特徴を持たない職業(「会社員、公務員、自営業」や「パート、アルバイト、無職」など)については、特徴語やライフスタイルなどでは適切に職業推定することが難しいという結果も出ている。

本研究において、投稿内容と投稿間隔を用いて生活習慣である睡眠時間(本研究では夜間活動休止時間)を推定するところで相違点があると考ええる。

3. 提案手法

投稿やライフスタイルの性質は職業によって大きく異なることが考えられるため、本研究では、対象を学生に絞り学生の投稿内容からの夜間活動休止時間の推定を目指す。

まず、TwitterAPI を利用して起床を表す単語(具体的には「おはよ」と「おきた」等)を含むツイートを投稿しているユーザを取得する。この時、プロフィール文に「学生」と含むユーザを選択する。

次に、選択したユーザからツイートを収集し、その中から投稿内容、投稿時間を参考に人手によって就寝前のツイートを取得する。収集したツイートデータをトレーニングデータとテストデータに分け、それぞれ形態素解析を行い、「名詞」「動詞」「名詞と動詞」のみを抽出したデータをそれぞれ用意する。このデータを利用して、Doc2Vec[3]と SVM[4]を用いた下記の 2 つの分類方法を用いて分類を行う。

(手法 1)Doc2Vec を用いた文書類似度による分類

(手法 2)SVM による 2 値分類

手法 1 では、2 値分類(夜間活動休止時間が 8 時間以上か未満)と 6 種類の分類(夜間活動休止時間が 2, 4, 6, 8, 10, 12 で分割)の 2 つの分類を行う。手法 2 では、2 値分類を行う。

Doc2Vec では学習済みモデル[5]を用いてトレーニングデータとテストデータの類似度を求める。テストデータそれぞれに対して最も類似しているトレーニングデータとのラベルを比較し

A Study on the Estimation of Nighttime Pause Time by Remarks in Tweets

[†]Takumi Kojima, Faculty of Engineering Tokushima University

[‡]Minoru Yoshida, Kazuyuki Matsumoto, Kenji Kita, Graduate School of Technology, Industrial and Social Sciences, Tokushima University

その分類精度により評価を行う。SVM ではトレーニングデータを用いて 2 値分類の分類器を作成し、テストデータが正解ラベル通り分類できているかの精度により評価を行う。

4. 実験データ

本研究で利用する実験データについて述べる。まず、トレーニングデータには 50 人のユーザ、テストデータには 10 人のユーザを選択し、それぞれ一人あたり 3000 ツイートを取得する。その中から、就寝前のツイートを抽出する。

次に Doc2Vec, SVM に用いたツイートデータ数を示す。

	トレーニングデータ	テストデータ
Doc2Vec	1817 ツイート	484 ツイート
SVM	1699 ツイート	464 ツイート

5. 実験結果

実験結果を示す。

		名詞	動詞	名詞と動詞
Doc2Vec	2 値	55.0%	50.3%	54.4%
	6 種類	31.5%	26.0%	32.5%
SVM		62.3%	61.7%	63.3%

表に示したように、SVM の分類精度が約 62%、Doc2Vec においては 2 値分類では約 50%、6 種類の分類では約 30%という全体的に低い結果となった。

6. 考察

結果に示したように Doc2Vec においての 2 値分類では精度が全体を通して約 50%、6 種類の分類では全体を通して約 30%、比較対象とした SVM が約 62%と最も精度が良いという結果になった。

このような結果になった理由としては、就寝前のツイートの取得方法に問題があったと思われる。自動で「おはよ」や「おきた」を含む起床時のツイートの 1 つ前のツイート就寝時のツイートとして取得するとそのツイートをする前に何かツイートをしているとそのツイートが就寝時のツイートとして取得してしまう問題点があったため、人手で就寝時のツイートを抽出したが、そのツイートをしたからといって今から寝る、または今起きたといったものの確証がないといった問題点が挙げられる。

7. おわりに

本研究では、就寝前のツイートの内容から睡眠時間に近似すると考えられる「夜間活動休止時間」を推定する手法を提案した。実験の結果としては、Doc2Vec における 2 値分類では約 50%、6 種類の分類では約 30%、SVM を用いた 2 値分類では約 62%という全体的に低い結果となった。

問題点としては、就寝時のツイートが本当に就寝時のツイートであるという確証がない、という問題点が挙げられた。

今後の課題としては、正確な夜間活動休止時間を取得する必要がある。そのため、本研究で利用したユーザよりもよりツイート頻度が多いユーザを選択しツイートを取得する、ツイートを取得するユーザに対してアンケートを取り正確な夜間活動休止時間を取得する必要があると思われる。

参考文献

- [1]睡眠 TOPICS,
<https://www.nishikawasangyo.co.jp/company/laboratory/topics/01/>
- [2]田中成典, 中村健二, 加藤諒, 寺口敏生: マイクログログの投稿時間に着目したユーザの職業推定に関する研究, 情報処理学会論文誌, データベース, Vol. 6, No. 5, pp71-84, Dec. 2013
- [3]鈴木千尋, 佐藤直行: Doc2Vec を用いた国語記述答案の自動評価, 情報処理学会第 81 回大会, pp135-136, 2019
- [4]Classias,
<http://www.chokkan.org/software/classias/index.html>. ja
- [5]日本語 WIKIPEDIA で学習した DOC2VEC モデル,
https://yag-ays.github.io/project/pretrained_doc2vec_wikipedia/