International Journal of Advanced Intelligence
Volume 12, Number 1, pp.1-9, March, 2021.

AIA International Advanced Information Institute

Toward Analyzing Relations between Sleeping Time and Social Networking Service Texts: Prediction of the Tweet Time Span Using the Last Tweet of the Day

Minoru Yoshida, Takumi Kojima, Kazuyuki Matsumoto, and Kenji Kita

Tokushima University, Minamijosanjima-cho 2-1
Tokushima, 770-8506, Japan
mino@is.tokushima-u.ac.jp
angel_busters@icloud.com
matumoto@is.tokushima-u.ac.jp
kita@is.tokushima-u.ac.jp

Received (25 Oct. 2020)
Revised (25 Dec. 2020)

Sleeping habits are one of the major issues in today's healthcare. In this paper, we consider the problem of analyzing sleeping habits of people using social networking service (SNS) texts. As the first step toward predicting user's sleeping time using SNS texts, we assume that the time span between the user's last post in one day and the first post the next day can be used as a pseudo-indicator for the user's sleeping time if the user posts the text sufficiently frequently. We call such tweet time spans "pseudo-sleeping time" if the first tweet of the next day include "Good morning" or similar words. We try to predict such pseudo-sleeping time using the text (tweet) of the preceding tweet (i.e., the last tweet of the day). Preliminary experiments show that the tweet text contains some useful information to predict the user's pseudo-sleeping time.

Keywords: sleeping time, SNS, text mining

1. Introduction

In this paper, we discuss the problem of predicting the sleeping time of a person given the social networking service (SNS) texts (e.g., tweets) of that person. Today, sleeping habits are one of the main issues in healthcare. It will contribute greatly to our quality of life if we can predict or obtain some insights about our sleeping habits from the SNS texts.

However, it is costly to collect accurate sleeping time data, resulting in very small size of obtained data to be useful for meaningful analysis or applicable to machine-learning algorithms.

To solve these issues, we instead propose a concept of *pseudo-sleeping time*, which can be easily collected using time stamps of SNS posts. Here, pseudo-sleeping time is defined as tweet time spans between the last tweet of one day and the first tweet of the next day. (See Figure 1.) Note that here "day" is not the span from 0:00 to 23:59 hours, but the span of awake hours of a person. In other words, we regard the tweet

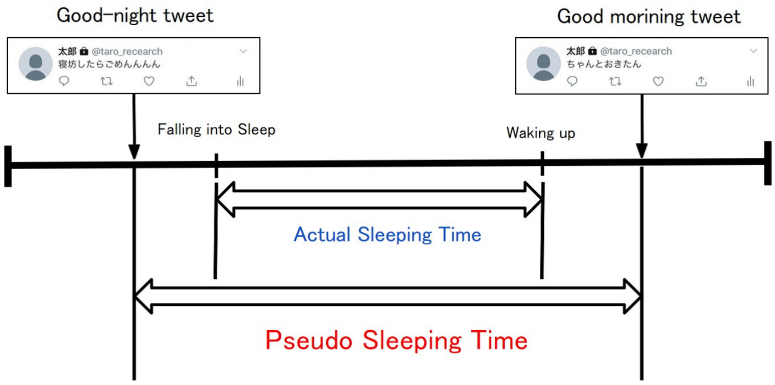

Fig. 1. Definition of Pseudo Sleeping Time

saying "good morning" or similar things as the first tweet of a day (which we call *good-morning tweets*) and the tweet before the first morning tweet as the last tweet of the previous day (which we call *good-night tweets*.) That is, the pseudo-sleeping time is the time span between a good-night tweet and the next good-morning tweet.

In this paper, we try to predict the pseudo-sleeping time using the contents of the good-night tweets using a standard support vector machine (SVM) classifier. We compare several settings using different datasets and different features to be used in the classifier.

The remainder of the paper is organized as follows. In Section 2, we discuss the related work. Section 3 explains our data and Section 4 describes our methods. We show our experimental results using two types of data in Section 5. Section 6 concludes the paper and lists the future work.

2. Related Work

Many researches for using Twitter for disease surveillance have been done so far[1]. For example, visualizing influenza epidemic by extracting tweets related to influenza is a major research topic in this area[6,7], while systems for mining more general public-health-related issues were also available[5]. Surveying public health using Twitter is also a major research topic. Sadilek and Kautz[3] proposed an automatic way to infer the health of people using geo-tagged tweets. They used features such as the word "sick" appearing in the tweets, the number of times the user visits gyms, etc.

Analyzing people with sleep troubles using Twitter was proposed by some researchers. Jamison-Powell et al.[2] used Twitter to find the people with their sleep disorder by Twitter content analysis. (For example, tweets containing the word "insomnia" strongly suggest that the user has sleep-related troubles.) McIver et al.[4] also proposed to use Twitter to investigate sleep issues by finding users who exhibited keywords like "can't sleep" and analyze their tweets. These researches focused

on people who have relatively serious troubles with sleep. In contrast, our research focuses on more general sleeping-time issues including simple insufficient sleeping time, which is not a severe disease currently but may cause other related diseases. Our contribution especially is to propose a method for analyzing the tweets that do not contain sleep-related words explicitly (like "can't sleep") for sleeping time surveillance.

3. Data and Problem Definition

We collected tweets via Twitter REST API v1.1 and then searched for "good morning" or other similar phrases. If the tweet m that contained such phrases (which we call *good-morning tweets*) were found, we obtained the previous tweet n (which we call *good-night tweets*), and calculated the time span t between m and n.

In this paper, we define our prediction problem as the binary classification problem. We give label $l = +1$ to the tweet n if $t > 8$ (i.e., pseudo-sleeping time is over 8 hours) and label -1 otherwise. We call the tweets with label $l = +1$ *positive tweets* and the tweets with label $l = -1$ *negative tweets*. Therefore, resulting (n, l) pairs (i.e., pairs of the good-morning tweet and time-classification label) are used as our dataset and the task is to predict label l given n.

4. Method

Each tweet is converted to the word list using the morphological analyzer, MeCab[a]. MeCab separates a sentence into a list of words, each of which is tagged with Part-Of-Speech (POS) labels such as nouns, verbs, and adjectives.

We tested the following three types of POS filters:

- nouns only,
- verbs only, and
- nouns and verbs only.

After obtaining a list of words for each sentence, words other than the selected POS (e.g., adjectives or prepositions) are discarded. The resulting list is converted to a Bag of Words (BoW) representation, which is a list of (word, frequency) pairs. The resulting BoW for each tweet is converted to tf-idf vectors by replacing a frequency value for each word with the corresponding tf-idf score. The tf-idf scoring function is defined as follows:

$$tfidf(t, d) = tf(t, d) \cdot idf(t, d)$$

$tf(t, d)$ and $idf(t, d)$ are defined as follows:

$$tf(t, d) = \frac{n_{t,d}}{\sum_{s \in d} n_{s,d}}$$

[a] https://taku910.github.io/mecab/

where $n(t, f)$ is the frequency of word t in document d.

$$idf(f) = \log \frac{N}{df(t) + 1}$$

where N is the number of documents and $df(t)$ is the number of documents in which t appears.

The obtained tf-idf vectors were divided into training and test data and used for classification. SVM[8] was used as a standard classifier for binary classification problems. SVMs were trained on training data, and the accuracy of the trained classifier was evaluated on test data. We used the Classias[9] implementation for the SVM classification.

5. Experiments

We collected two types of data described as follows:

Data-1: We collected good-morning tweets randomly using the Twitter search API and obtained corresponding good-night tweets.

Data-2: We collected good-morning tweets of selected users, and obtained corresponding good-night tweets.

In the following subsections, we describe these datasets and reports the experimental results on each data set.

5.1. *Results on Data-1*

Data-1 is a collections of tweets randomly obtained by searching for good-morning tweets and corresponding good-night tweets. As a result, we obtained the following number of tweets for training and test data.

- Training data: 1,500 tweets
- Test data: 100 tweets

We applied SVM on this data. The results are shown in Table 1.

Table 1. Results on Data-1

nouns	verbs	nouns and verbs
47%	83%	51%

We observed that the setting using verb features showed extremely high accuracy. To find the reasons, we checked what kinds of strings are found in positive ($l = +1$) or negative ($l = -1$) tweets.

We found that the following phrases were frequently found in the negative tweets:

(1) "Fukugyou" (second job)
(2) Japanese quotation marks
(3) "Eigyou" ((shop is) open)
(4) "Nyusu" (news)
(5) "DM" (direct mail)

These were typical strings found in bot accounts. For example, "Eigyou" (open) was found in the advertisement tweets such as "XX shop is open from 10:00 to 22:00 !". We also found that Japanese quotation marks were typically found in the bot accounts that quoted famous phrases or jokes. As a result, phrases that suggest that the account was for the user's lifelogs could be used as features to distinguish such lifelog tweets from the "noisy" bot tweets. Especially, verbs like "Neru" (sleep) strongly suggest that the accounts were for lifelogs, resulting in high accuracy for verb features.

Considering these results, we concluded that the high accuracy obtained on the data 1 was highly due to these "noises" that did not reflect the actual sleeping time of the user.

Therefore, we decided to construct a new data set (the data-2 mentioned above), which is "cleaner" in the sense that it:

(1) Consisted of reliable users only.
(2) Manually checked whether the good-night tweets are actually the last tweet of the previous day.

To satisfy condition 1, we limited the collected accounts to be of the student users only because student users show relatively regular life patterns compared to adults who have much various statuses (e.g., office workers, housewives/husbands, self-employed workers, etc.), resulting in more meaningful analysis. We also focused on the users with the sufficient numbers of tweets to be used for good-morining and good-night tweets extraction. Condition 2 is to filter out the "noisy" tweets that is before the morining tweet but is not the good-night tweet (e.g., the pre-good-morining tweets that do not say "good-morining", such as "I will sleep a little longer...") Although currently we check the condition 2 manually, we think we can automate the process using some patterns or classifiers to filter out such noisy good-night tweets.

5.2. *Results on Data-2*

In data-2, we obtained the user profile texts of the tweets, and retained the tweets of the user whose profile contains the word "student." It is for focusing on only tweets by students for the purpose of regularizing the lifestyle habits of the users, because lifestyle habits are much different among people with various status/jobs, such as students and office workers.

As a result, we obtained the following number of users, as shown in Table 2.

Table 2. The Number of Users and Tweets

# of data \ # of data types	training	test
# of users	50	10
# of data per user	3,000	3,000
# of all tweets	150,000	30,000

We obtained good-morning and good-night tweets from these tweets, resulting in the following number of training and test data.

Table 3. Description of Data-2

training	test
1817	484

We applied SVMs to the obtained data in the same way as data-1. Table 4 shows the results.

Table 4. Results on Data-2

nouns	verbs	nouns and verbs
62.3%	61.7%	63.3%

We observed that we could predict the label l with over 60% accuracy. This result suggests that the last tweet of a day contained some useful information for predicting the user's sleeping time.

To analyze the results in further detail, we extracted phrases frequently found in positive/negative tweets in the same way as data-1.

"Baito" (part-time job) were very frequently found in positive tweets (19 times in positive tweets while 0 times in negative ones). In contrast, various phrases were found in negative tweets such as:

(1) "Jikan" (time)
(2) "Testo" (exam)
(3) "2ji" (2 o'clock)
(4) "tokei" (clock)
(5) "neochi" (falling asleep)
(6) "netai" (want to sleep)

Phrase 1, 3, and 4 suggest that the user suddenly realized that it was late at

night. In contrast, phrases 2 and 6 suggest that they had work to do and thus could not go to bed. Phrase 5 suggests that the user fell asleep and would go to bed to sleep for the remaining time. In general, these situations tend to cause a reduced sleeping time.

6. Conclusions and Future Work

In this paper, we discussed the problem of predicting the sleeping time of persons using Twitter texts posted by users. For collecting sufficient data at low costs, we proposed the concept of pseudo-sleeping time, which can be considered as a kind of upper-bound of the true sleeping time and used as an approximation to the sleeping time. Using this concept, we defined our task to classify the pseudo-sleeping time using the final tweet of the day.

We applied SVM classifiers on different datasets using different features, and found that we can predict the pseudo-sleeping time to some extent using appropriately collected datasets by avoiding "noises" (i.e., bot tweets) and constrained user types (i.e., students in our case).

The future work includes analysis of the relation between the pseudo-sleeping time and true sleeping time. We also plan to increase the data size by including more users. The use of classification methods other than SVMs will also be an important future work.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18K11549, JP20K12027.

References

1. Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, Courtney D Corley, Using social media for actionable disease surveillance and outbreak management: a systematic literature review, PloS one, 10(10), 2015.
2. Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, Shaun W. Lawson: "I can't get no sleep": discussing #insomnia on twitter. CHI 2012: pp. 1501-1510, 2012.
3. Adam Sadilek, Henry A. Kautz: Modeling the impact of lifestyle on health at scale. WSDM 2013: pp. 637-646, 2013.
4. David J. McIver, J. Hawkins, R. Chunara, Arnaub K Chatterjee, Aman Bhandari, Timothy P. Fitzgerald, S. Jain, J. Brownstein: Characterizing Sleep Issues Using Twitter, Journal of Medical Internet Research, Vol 17, No 6, 2015.
5. Michael J. Paul, Mark Dredze: You Are What You Tweet: Analyzing Twitter for Public Health. ICWSM 2011
6. Shin Kanouchi, Mamoru Komachi, Naoaki Oka-zaki, Eiji Aramaki, Hiroshi Ishikawa, Who caught a cold ? - Identifying the subject of a symptom, Proceedings of ACL (1) 2015, pp. 1660-1670, 2015.
7. Hayate Iso, Shoko Wakamiya, Eiji Aramaki, Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction, Proceedings of COLING 2016, pp. 76-86, 2016.
8. Corinna Cortes, Vladimir Vapnik, Support-vector networks, Machine Learning vol. 20, pp. 273?297. 1995.

9. Naoaki Okazaki, Classias: A collection of machine-learning algorithms for classification, http://www.chokkan.org/software/classias/

Minoru Yoshida

He is a lecturer at the Department of Information Science and Intelligent Systems, University of Tokushima. After receiving his BSc, MSc, and PhD degrees from the University of Tokyo in 1998, 2000, and 2003, respectively, he worked as an assistant professor at the Information Technology Center, University of Tokyo. His current research interests include web document analysis and text mining for the documents available on the WWW.

Takumi Kojima

He received his bachelor's and master's degrees from Tokushima University in in 2018 and 2020, respectively. His research interests include natural language processing and social network analysis. He currently works at DOCOMO CS Shikoku Inc.

Kazuyuki Matsumoto

He received his PhD degree in 2008 from Tokushima University. He is currently an associate professor of Tokushima University. His research interests include affective computing, emotion recognition, artificial intelligence and natural language processing. He is a member of IPSJ, ANLP, IEICE and IEEJ.

Kenji Kita

He received the B.S. degree in mathematics and PhD degree in electrical engineering, both from Waseda University, Tokyo, Japan, in 1981 and 1992, respectively. From 1983 to 1987, he worked for the Oki Electric Industry Co. Ltd., Tokyo, Japan. From 1987 to 1992, he was a researcher at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan. Since 1992, he has been with Tokushima University, Tokushima, Japan, where he currently is a Professor at Faculty of Engineering. His current research interests include multimedia information retrieval, natural language processing, and speech recognition.