Recognizing Depression from Twitter Activity

Sho Tsugawa
University of Tsukuba
Ibaraki 305-8573, Japan
s-tugawa@cs.tsukuba.ac.jp

Yusuke Kikuchi
Kwansei Gakuin University
Hyogo 669-1337, Japan
bfc72150@kwansei.ac.jp

Fumio Kishino
Kwansei Gakuin University
Hyogo 669-1337, Japan
kishino@kwansei.ac.jp

Kosuke Nakajima[*]
Osaka University
Osaka 565-0871, Japan
nakajima.kosuke@ist.osaka-u.ac.jp

Yuichi Itoh
Osaka University
Osaka 565-0871, Japan
itoh@ist.osaka-u.ac.jp

Hiroyuki Ohsaki
Kwansei Gakuin University
Hyogo 669-1337, Japan
ohsaki@kwansei.ac.jp

ABSTRACT

In this paper, we extensively evaluate the effectiveness of using a user's social media activities for estimating degree of depression. As ground truth data, we use the results of a web-based questionnaire for measuring degree of depression of Twitter users. We extract several features from the activity histories of Twitter users. By leveraging these features, we construct models for estimating the presence of active depression. Through experiments, we show that (1) features obtained from user activities can be used to predict depression of users with an accuracy of 69%, (2) topics of tweets estimated with a topic model are useful features, (3) approximately two months of observation data are necessary for recognizing depression, and longer observation periods do not contribute to improving the accuracy of estimation for current depression; sometimes, longer periods worsen the accuracy.

Author Keywords

Social media; Depression; Twitter; Machine learning

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Depression has become recognized as a major public health problem around the world [21]. A 2012 survey conducted by the World Health Organization (WHO) suggests that depression affects 350 million globally [21]. Among other associated problems, depression can lead to suicide [32], and so the high number of people with depression is considered to be a serious problem. Also in Japan, depression and suicide are recognized as serious problems. An OECD (Organisation for Economic Co-operation and Development) report shows that the standardized suicide rate per hundred thousand people in Japan is 20.9 in 2011, which is much higher than OECD average of 12.4 [24]. Suicide of young people is particularly serious since for 15–39 year old people in Japan, the top cause of death is suicide [15]. To cope with such a serious problem, the government of Japan is making strong efforts on the care of depression [15] since depression is considered to be a major cause of suicide [32].

Before depression can be effectively treated, it must be recognized in individuals [18, 34, 36]. The WHO reports that the majority of people who need treatment for depression do not receive it [21]. Even during visits with primary care physicians, depression is often unrecognized and undiagnosed; as a result, many people with depression do not receive adequate treatment [36]. Hence, for the effective treatment of depression in the population, it is important that people know how to recognize symptoms of depression in themselves and those around them.

Several measures for estimating signs of depression in individuals have been proposed in the form of questionnaires [2, 17, 30, 40]. Among such questionnaires, there are many scales for estimating severity of depression: the Center for Epidemiologic Studies Depression Scale (CES-D) [30], Beck's Depression Scale (BDI) [2], Zung's Self-rating Depression Scale (SDS) [40], and Hamilton's Depression Rating Scale (HDRS) [17] are the more popular. Results on these scales are obtained from answers to the questionnaires, either through self-reporting or third-party observation.

These subjective measures are widely used and are expected to achieve high estimation accuracy. Nevertheless, developing techniques for recognizing depression in individuals by using objective, rather than subjective, information is important. One reason for this is that questionnaires are often costly, and sometimes significantly so. In contrast, a method for recognizing depression in individuals by examining objective information, such as records of daily activities, could be applied for a very low cost in many cases. A technique to identify symptoms of depression in individuals from objec-

[*]Currently with Mitsubishi Electric Corporation

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.
Copyright © 2015 ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702280

tive information would hasten recognition of depression by individuals and thereby allow earlier access to effective treatment.

As objective information for identifying depression in individuals, we focus on large-scale records of users' activities in social media. In recent years, social media has become increasingly popular, and therefore both large-scale and fine-grained records of users' activities are available [26]. Many researchers are interested in analyzing such large-scale data, and prior analyses have used those for purposes such as estimating the future movement of stock market indices and predicting the results of elections [6, 23, 38, 39, 33]. Since social media are often used for expressing emotions, the records of users' activities are a promising source for information that could be used to recognize symptoms of depression in the users.

The use of data from social media to estimate severity of depression in social media users has been studied before. For example, Tsugawa *et al.* [37] found that frequency of word usage in messages (hereinafter, tweets) on the popular microblogging service Twitter could be used for recognizing depression among users. Pioneering work by De Choudhury *et al.* [14] describes how to construct support vector machine (SVM) classifiers that can be used to identify users with depression using features obtained from the records of individual users' activities on Twitter. That research shows the potential for recognizing symptoms of depression in a user by examining the user's social media activities.

The results of prior research suggest that the use of social media for recognizing depression is a promising approach. However, research into these techniques is just beginning, and so extensive evaluation is necessary. As an example, the sensitivity of results to the period over which a user's social media activities are analyzed is not yet known. Additionally, the generality of the results of [14] should be verified by analyzing other datasets. It should be important to share and accumulate the results from several datasets among researchers particularly in such an emerging research field.

In this paper, we extensively evaluate the effectiveness of using a user's social media activities for estimating degree of depression. As ground truth data, we use the results of a web-based questionnaire for measuring degree of depression; this questionnaire was completed by the Twitter users who agreed to participate. From these data and user activity histories, we investigate the relation between features of each user's activity and that user's score on the depression questionnaire. We construct several models to predict users with depression from the features of the user's activity. For this purpose, we use SVM and investigate the estimation accuracy of the models. Our main contributions are summarized as follows.

- We show that features obtained from user activities can be used to predict depression of users with an accuracy of 69%, which is similar to the rate in [14]. While De Choudhury *et al.* [14] use data about English-speaking users, we use data about Japanese-speaking users. Together, these results suggest that the use of data from social media is an effective approach for users with different backgrounds (here, English-speaking and Japanese-speaking users).
- We show the effectiveness of using topics, as identified in tweets by a topic model [5], as features for the SVM. Our results show that applying a simple bag-of-words model (i.e., word frequencies) to tweets results in overfitting, but using topics inferred from word frequencies prevents overfitting and improves estimation accuracy.
- We investigate the relation between the number of tweets used for estimation and the estimation accuracy. We suggest an appropriate amount of tweets for recognizing depression. Additionally, we find that about two months of observation data are necessary for recognizing depression, and longer observation periods do not contribute to improving the accuracy of estimation for current depression; sometimes, longer periods worsen the accuracy.

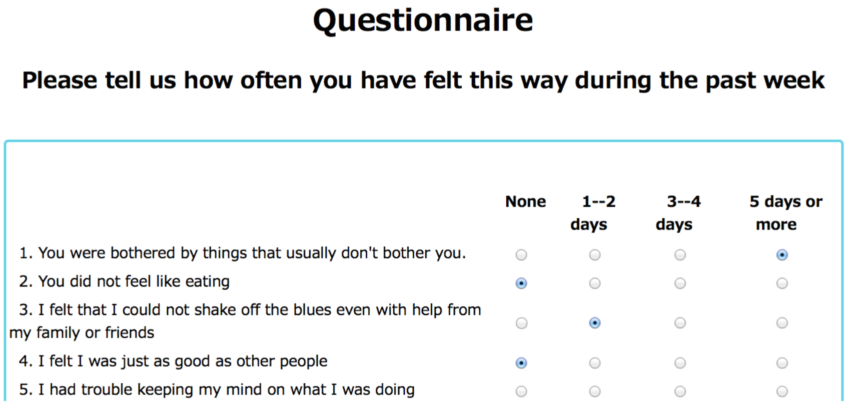
RELATED WORK

In the areas of medicine and psychology, several questionnaire-based measures for rating depression in individuals have been proposed [2, 17, 30, 40]. CES-D [30], BDI [2], and SDS [40] estimate the severity of depression in individuals from the self-reported answers to 20 questions. HDRS uses the answers given by a third party to 17 questions for the same purpose [17].

Approaches that use objective information, such as log data about an individual's activities to estimate depression severity have been studied recently. Resnik *et al.* [31] propose a method for identifying depression in individuals by analyzing textual data from essays written by the individuals. Resnik *et al.* obtain topics for the essays by applying a popular topic-extraction model, latent Dirichlet allocation (LDA) [5]. By using the discovered topics as features in machine learning, depression severity is estimated. That research is an important work that shows the potential for using texts written by an individual to estimate severity of depression in that person. However, the essays are typically difficult to obtain.

Due to the widespread adoption of social media and the availability of large-scale data from social media, approaches that use such data for depression screening are receiving increased attention from researchers. Moreno *et al.* [22] shows that college students display symptoms consistent with depression on Facebook, a popular social networking service. Park *et al.* [27] analyze differences between Twitter users with and without depression by analyzing their activities. In Park *et al.* [28], a similar analysis is performed by analyzing data from Facebook. Using multiple regression analysis, Tsugawa *et al.* [37] show that frequencies of word usage on Twitter are useful features for recognizing depression among users. The main objective of such research is to clarify which features that can obtained from user activity are useful for estimating the severity of depression.

De Choudhury *et al.* [14] are pioneering in demonstrating the estimation accuracy that could be achieved by using activities on Twitter to predict depression among users. In their study,



Questionnaire

Please tell us how often you have felt this way during the past week

	None	1--2 days	3--4 days	5 days or more
1. You were bothered by things that usually don't bother you.	○	○	○	●
2. You did not feel like eating	●	○	○	○
3. I felt that I could not shake off the blues even with help from my family or friends	○	●	○	○
4. I felt I was just as good as other people	●	○	○	○
5. I had trouble keeping my mind on what I was doing	○	○	○	○

Figure 1. A screenshot of our website. The questionnaires to determine degree of depression (CES-D questionnaires) were completed by participants through their web browsers. (Originally all messages shown in this figure were written in Japanese.)

De Choudhury *et al.* obtained training data for machine learning by appealing to large numbers of people for help (popularly known as crowdsourcing). Then, models that could be used to predict risk of depression were identified from several features obtained from the records of user activity on Twitter by using SVM. Experimental results show that depression can be recognized among users with an accuracy of approximately 70%. Such approaches are applied to prediction of postpartum depression from Facebook and Twitter data [11, 13]. De Choudhury *et al.* have also proposed a method for estimating the depressive tendencies of populations by a similar approach [12]. Twitter data are also suggested to be effective for estimating health-related statistics [10, 29].

Our study builds on the mentioned prior work and contributes to enhancing methods for predicting depression risk from objective information, particularly large-scale data from social media. Research into using social media data for recognizing depression in individuals is just beginning, and so whenever possible the effectiveness of such approaches should be validated for several datasets. In this study, we adopt an approach similar to that in De Choudhury *et al.* [14] and apply it to Japanese-speaking Twitter users. Note that this is not the first study on depression of non-English social media users since differences between Korean social media users with and without depression have been analyzed in [27, 28]. This study extends the prediction framework of [14] to Japanese-speaking users. We also examine the effectiveness of using topics of tweet messages as a feature for estimating depression risk. Additionally, we investigate the effects on accuracy of observation period and number of data items used for estimation; this has not yet been fully investigated.

METHODOLOGY

Data Gathering

In this study, we gathered information on depression levels of Twitter users and their activity histories. To do this, we published a website to administer a questionnaire and disseminated information about the website over Twitter[1]. In contrast to De Choudhury *et al.* [14], who collected data from English-speaking users through crowdsourcing, this study collected data from Japanese-speaking volunteers. This approach was used to investigate the extent to which depression risk can be estimated for a population different from the population considered by the prior research [14]. Figure 1 shows a screenshot of our website.

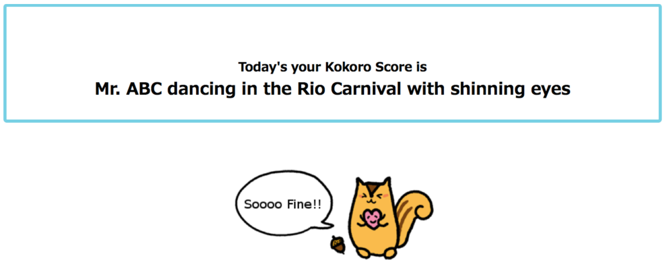
The website collected the responses to a questionnaire to evaluate the degree of depression of the Twitter users who participated (hereinafter, the participants) and to collect the histories of participants activities on Twitter. The activity histories of participants were collected through the Twitter application programming interface (API)[2], and the questionnaires to determine degree of depression were completed by participants through their web browsers.

Before data collection, visitors to the website were presented with a written explanation of the aims of the experiment, the information that would be collected, and how that information would be handled. Those who consented to become participants after receiving the explanation logged into their individual Twitter accounts through the OAuth authorization process. Next, participants were surveyed on gender, age, occupation, and history of depression, following which they answered a questionnaire designed to evaluate degree of depression. A message called the “kokoro score,” (“kokoro” is a Japanese word meaning “heart”) determined on the basis of answers to the questionnaire and information in the collected tweets, was displayed to participants after completion of the questionnaire (Fig. 2). Experiment participants were able to tweet the message displayed, which made it possible to promote the website over Twitter by word-of-mouth in a type of snowball sampling.

The CES-D questionnaire was used to evaluate the degree of depression [30]. In the CES-D test, participants answered 20 questions on a Likert-type 4-point scale. Each answer was assigned a score of 0-3 points, with the sum of the points from all answers used as the score to estimate likelihood of depression. Several standards exist by which to determine the appro-

[1] Administration of the website and processing of the obtained data were performed under the approval by the ethical committee in the organization to which authors belong.

[2] `https://dev.twitter.com/docs/api/1.1`



Today's your Kokoro Score is

Mr. ABC dancing in the Rio Carnival with shinning eyes

Figure 2. A screenshot of our website showing *kokoro score* to user Mr. ABC. Several candidate messages for display were prepared in advance, and a message was selected from among these and displayed to the participant according to the questionnaire answers and textual information contained in the tweets. The aim of this message was to provide participants an incentive to answer the questionnaire. (Originally all messages shown in this figure were written in Japanese.)

priate cutoff score for identifying depression. In this research, we regarded a score of 22 points or higher as indicating active depression and a score of 21 points or lower as indicating no active depression; these are the same values as used in [14] and give a cutoff score of 22. In addition, answers to BDI [2], a depression scale used with characteristics similar to CES-D, were collected to ensure the reliability of data. For each participant, scores were calculated on both scales, with poor correlation regarded as indicating unreliable answers. The time taken to answer the questionnaires was also recorded, and those completed in too brief a time were excluded. After each participant answered the questionnaire, the activity history of that participant on Twitter was collected from Twitter by using the API. At most 3,200 tweets were collected for each participant, and the number of users following the participant and being followed by the participant were recorded. Tweets published after the questionnaire was taken were discarded.

The website was opened to the public on 4 December 2013, at which time the authors publicized it on their Twitter accounts. Between 4 December 2013 and 8 February 2014, 219 people participated in the experiment. After eliminating participants who did not tweet and participants who answered the questionnaire in fewer than 30 seconds (as previously mentioned, to ensure the reliability of the questionnaire answers), 214 sets of answers remained. Only the first set of answers was used for participants who completed the questionnaire more than once. As a result, data about 209 experiment participants (male: 121; female: 88) aged 16 to 55 (mean: 28.8 years; standard deviation: 8.2 years) were analyzed. The correlations between CES-D score and BDI score for these participants were high, 0.87, and there were no participants with uncorrelated scores, so the data for all 209 participants were used; excluded datasets are not discussed any further. Figure 3 shows the histogram of CES-D scores of 209 participants. Among the participants, 81 (resp. 128) were estimated to have (resp. not have) active depression, for an incidence of approximately 39%. This incidence is similar to that found by De Choudhury *et al.* [14], who identified depression in approximately 36% of participants. Table 1 gives statistics on the activity histories of participants.

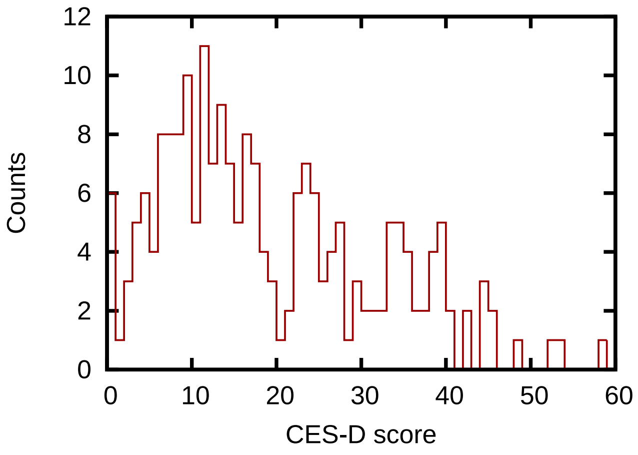


Figure 3. Histogram of CES-D scores of 209 participants

Table 1. Basic statistics about activities of participants

	Mean	Median	Std. dev.
Num. of tweets	2749.1	3191	850.2
Num. of tweets per day	18.0	7.8	27.5
Days spanned by tweets	569.2	348.2	502.0

Features for Recognizing Depression

It is possible to extract various features from the activity histories of Twitter users. This section explains what kinds of features are used to estimate degree of depression and the way in which these quantities are extracted. Table 2 shows the features used in this study. A detailed explanation of each feature follows.

The frequencies of words in a tweet (i.e., its bag of words) are used as a basic feature relating to the content of the tweet. Tsugawa *et al.* showed that the word frequencies are useful for identifying depression [37]. MeCab [20] was used to for morphological stemming and categorization of the Japanese tweet text to obtain accurate word frequencies. Particles, auxiliary verbs, adnominal adjectives, and visual symbols were excluded for extracting content words. Words used by only

Table 2. Features used for predicting depression

Name	Description
bag of words	Frequencies of words used in the tweet
topic	Ratio of tweet topics found by LDA [5]
positive	Ratio of positive-affect words contained in the tweet
negative	Ratio of negative-affect words contained in the tweet
hour	Hourly posting frequency
tweet frequency	Tweets per day
num. of words	Average number of words per tweet
RT	Overall retweet rate
mention	Overall mention rate
URL	Ratio of tweets containing a URL
followee	Number of users following
follower	Number of users followed

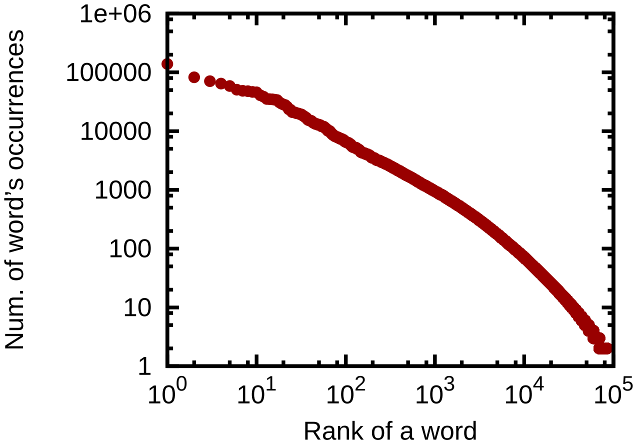


Figure 4. Distribution of word frequencies: Rank of a word based on its frequency vs. the total number of the word's occurrences in all tweets

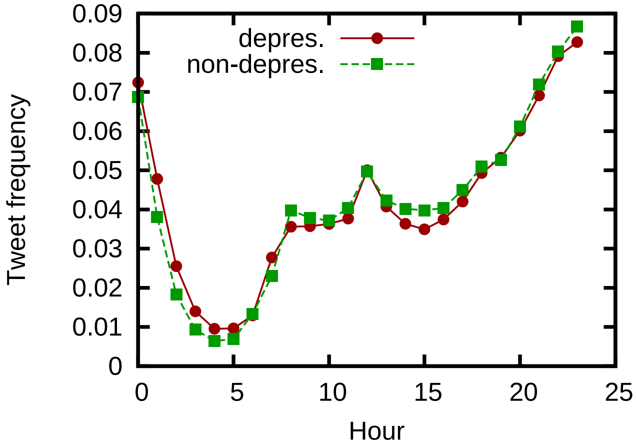


Figure 5. Hourly posting frequency: Comparison between participants with and without depression

one participant were also excluded, resulting in a total of 84,255 distinct words. However, most of these words were rarely used, and the distribution of word frequencies is extremely biased (see Fig. 4). Because words with a low rate of use were regarded as unlikely to be associated with depression for most users, the frequencies of only the 20,000 words with the highest rate of use (corresponding to 25 or more uses across all participants) were used as a feature in this study. Furthermore, because the number and length of tweets differed by participant, the word frequencies were normalized by the total number of words in the tweets.

The topics of the tweets of each user, as estimated by using a representative topic model LDA [5], were used as a second feature relating to the content of the tweets. With LDA, the distribution of topics in each document is estimated from the word frequencies in each text through unsupervised learning on the assumption that the text and the words in it are generated according to a particular topic [5]. In LDA, the number of topics to identify and a set of documents (as bags of words) are used as input, and a topic distribution is output for each document. As mentioned in Related Work Section, the topics of essays written by university students were estimated by using LDA and found to be useful in evaluating degree of depression [31]. From that study, topics are expected to be a useful feature. A set of all tweets of each user was used as the user document for input in LDA, and the 20,000 words selected as described above were used as the words. We used LDA with collapsed Gibbs sampling [16]. As the parameters of LDA, we used $\alpha = 50/K$ and $\beta = 0.1$, where K is the number of topics [16]. All extracted topics were used as the features.

The ratio of positive words and the ratio of negative words used in the tweet text are used as the final features relating to tweet content. Users with depression are intuitively expected to use negative words more frequently than users without depression do. To categorize words, a dictionary of affective words [19], which is compiled by manual evaluation of a dictionary of positive and negative words extracted according to a technique proposed in the literature [35], is used. The dictionary contains 760 positive words and 862 negative words.

The user's timing of tweets, frequency of tweets, average number of words, retweet rate (rate of republishing other users' tweets), mention rate (rate of directly referencing at least one other user), ratio of tweets containing a uniform resource locator (URL), number of users being followed, and number of users following are used as features independent of the content of the tweet. The relative ratios of tweets posted during each hour of the day were used to characterize the timing of tweets; the number of posts per day was used as the posting frequency; and the ratio of qualifying tweets to all tweets were used for the retweet ratio, mention ratio, and ratio of tweets containing a URL. These features are used in prior research [14].

RESULTS

Comparison of Features between Users with and without Depression

Differences in the features described in the previous section according to the presence or absence of depression were investigated. A complete discussion of individual-level differences in word frequencies and topics according to the presence of depression would be lengthy and could obfuscate the underlying results. Accordingly, in this section we focus on features other than the bags of words and topics. In the next section, we use the bags of words and topics to predict depression among participants and check those results.

First, a comparison was made of the posting frequency by time for all participants (Fig. 5). Figure 5 shows that the posting frequency was the highest at 11 p.m. and the lowest at 4 a.m. through 5 a.m. (all times are Japan Standard Time) among all participants, independent of the presence of depression. Almost no difference was found in posting frequency of the two groups. This differs from the results of [14], in which

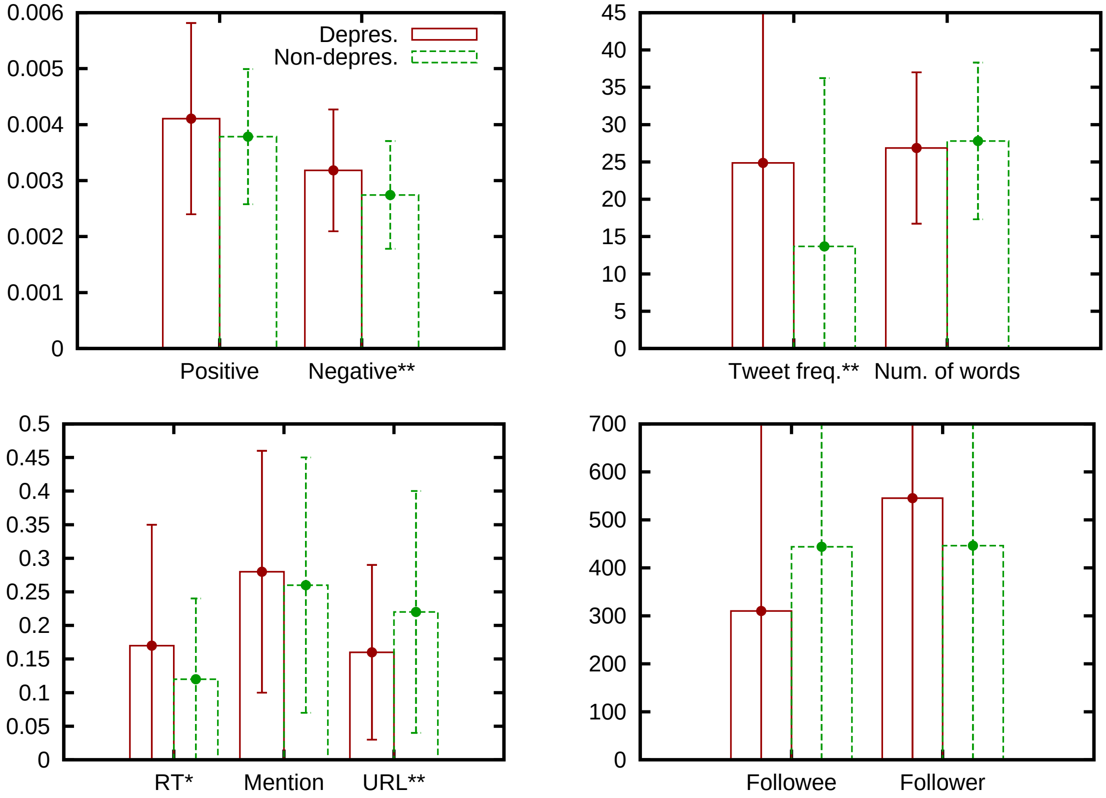


Figure 6. Feature means: Comparison between participants with and without depression (: $p < 0.01$, * : $p < 0.05$)**

a large difference between participants with and without depression was observed in relative frequencies by time. The difference in observed trends suggests that the feature relating to posting time is not robust.

Next, the averages of each feature, partitioned by depression status, are examined (see Fig. 6). A significant difference between the two groups was found for the relative ratio of negative words, posting frequency, retweet rate, and ratio of tweets containing a URL. Individual t-tests confirmed significant differences in these features according to the presence of depression, with a significance level of 5%. Furthermore, although not statistically significant in our study, a numerical difference can be seen between groups in the number of followees and followers. These features were found to be significant in [14], and these are expected to be useful features in the general case. Similarly, although a significant difference was observed in mention frequency in [14], no such difference was observed in this study. This suggests that mention frequency is not a robust feature, in the same way as posting time. The differences between this study and [14] might be caused by the cultural differences between Japanese-speaking and English-speaking users. However, further investigations are necessary to clarify the reason of the differences; i.e., whether these are really due to the cultural differences, or just due to the differences of the participants.

The above analysis identifies several features that may be helpful in recognizing depression in Twitter users. In the next section, we investigate the accuracy of estimating the presence of active depression from these features.

Prediction Accuracy

In this section, we investigate the degree of accuracy with which the presence of active depression can be ascertained from the features extracted from a user's activity history. Classifiers were constructed by machine learning with SVM for estimating the presence of active depression, and their classification accuracy was evaluated by 10-fold cross-validation. Several classifiers were constructed from the extracted features, and the classification accuracy of each was investigated. The precision, recall, F-measure, and accuracy of the estimations (where the CES-D scores are taken as ground truth with a cutoff score of 22) are used as indices to evaluate classification accuracy. The precision is the proportion of participants with depression among those classified as having depression; the recall is the proportion of participants with depression who were classified as having depression; the F-measure is the harmonic mean of the precision and the recall; and the accuracy is the rate of correct classification across all participants. We note here that classifying all participants as not having depression yields precision, recall, F-measure and accuracy of 0, 0, 0, and 0.61, respectively. In the complementary case, in which all participants are classified as having depression, the corresponding values are 0.39, 1, 0.56, and 0.39, in the same order.

Table 3 shows the classification accuracy of the constructed model. The classification accuracies are the average values given by 10-fold cross-validation carried out 100 times. There are a large number of combinations of all features, so only the results of the most accurate models and those models useful for discussion are shown. A radial basis function kernel was used for the SVM kernel.

Table 3 shows that the presence of active depression can be estimated by the most accurate model with 0.61 precision, 0.37 recall, 0.46 F-measure and 66% accuracy. Because the research of De Choudhury *et al.* and that of this study use different data, a simple comparison cannot be carried out, but these accuracies can be considered to be comparable level to those reported by De Choudhury *et al.* [14]. Since similar results were obtained from different datasets, estimating the presence of depression by examining users' activity histories on Twitter can be considered to be a useful technique.

Looking at the features used, it is found that the estimation accuracy of models using the bag of words as a feature is low. The cause of this may be that this approach results in overfitting. When using the bag of words feature, the feature dimensions are too high, which is known to degrade performance. In Tab. 3, limiting the number of words to 2,000 (rather than 20,000) improves accuracy with the bag of words feature, which confirms that overfitting is occurring. However, even after this adjustment, accuracy with that feature is low in comparison to the accuracy of models without it. In contrast, models using topics as a feature achieved a high accuracy. This means it may be possible to improve estimation accuracy by reducing the dimensions to the broad feature of topics, rather than using the narrow feature of word frequencies. Not only topics but also other features (positive, negative, tweet frequency, RT, URL, followee, and follower from Tab. 2) were found to result in more accurate models than using the bag of words, and these additional features may be useful when predicting the presence of depression. Together with topics, models using the number of followees and the posting frequency achieved the highest accuracy of the models investigated here.

Looking at the number of topics in the topic model, it was found that accuracy was the highest in this experiment when the number of topics was set at 10. This is probably because using a very small number of topics lowered the expressiveness of the model using a large number of topics made it difficult to capture the predominant topics.

Effects of the Amount of Data on Prediction Accuracy

Next, we varied the quantity of tweets used for learning and estimation, and investigated the resulting estimation accuracy in order to find the quantity of data needed to make accurate predictions. Features were extracted using the N most-recent tweets (prior to study participation) and applied these features to model learning and prediction. For users with fewer than N tweets, all available tweets were used. Here, we discuss the results for the following models: the 10-topic model (Model 1); a model using the features positive, negative, tweet frequency, RT, URL, followee, and follower (Model 2); and a 10-topic model including the features positive, negative, tweet frequency, RT, URL, followee, and follower with the highest estimation accuracy from the previous section (Model 3). Figure 7 shows the relation between N (the number of tweets used) and the F-measure (used to representing estimation accuracy).

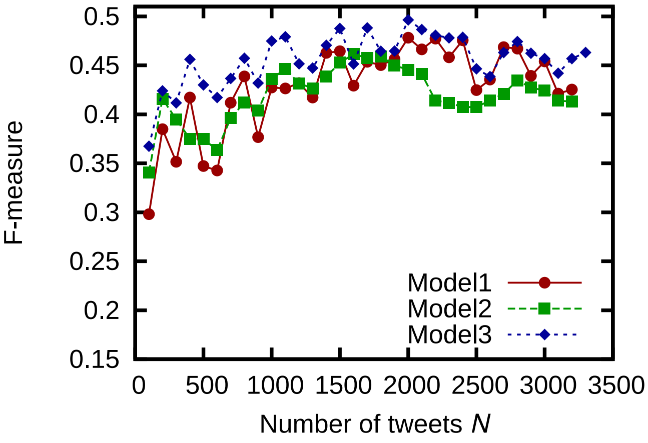


Figure 7. Number of tweets used for training and prediction vs. F-measure

From these results, it is found that the F-measure reaches approximately 0.45 when the number of tweets used is approximately 500–1000, and it remains almost unchanged even when the number of tweets used is further increased. This implies that observing around 500–1000 tweets is sufficient for recognizing depression, and observation of a higher number of tweets contributes negligibly to estimation accuracy.

The time span over which tweets were observed was also controlled, and the F-measure was calculated to assess the resulting estimation accuracy. The tweets used in learning and prediction were limited to those made in the W weeks prior to the day on which the participant answered the questionnaire, and the previously mentioned models were used for prediction. Figure 8 shows the relation between the time span W and the F-measure.

From this result, it is found that the F-measure reaches 0.4–0.5 when the tweets of the prior 6–16 weeks (1.5–4 months) are used, and lengthening the time span does not improve this; in some cases, longer time spans result in worse results. This indicates that using tweets from the most recent 6–16 weeks is sufficient for recognizing depression, and observations spanning a longer period may be less accurate. This could be because out-of-date information may not reflect the current state of depression; in such cases, the additional data do not contain information useful for prediction. The highest accuracy is achieved when using tweets from the recent 8 weeks, and the presence of active depression can be estimated by Model 3 with 0.64 precision, 0.43 recall, 0.52 F-measure and 69% accuracy.

Table 3. Classification accuracy of constructed models

Features	Precision	Recall	F-measure	Accuracy
bag of words (20,000 words)	0.04	0.0005	0.024	61%
bag of words (2,000 words)	0.58	0.13	0.21	62%
5 topics	0.50	0.13	0.21	61%
10 topics	0.55	0.35	0.43	64%
20 topics	0.54	0.33	0.41	63%
positive + negative + tweet frequency + RT + URL + followee + follower	0.57	0.33	0.41	64%
10 topics + positive + negative + tweet frequency + RT + URL + followee + follower	0.61	0.37	0.46	66%

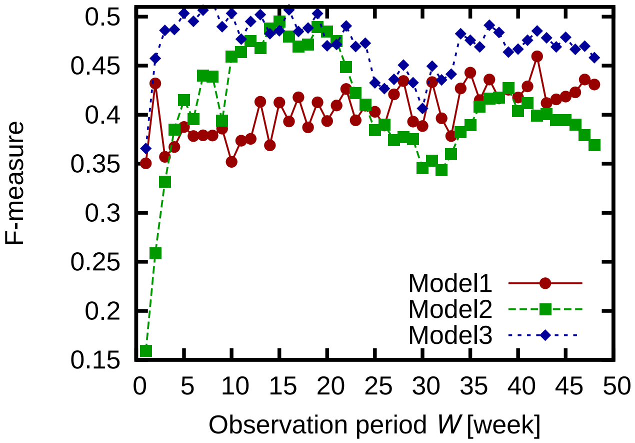


Figure 8. Observation period W [week] vs. F-measure

DISCUSSION

This study investigated the degree to which the activity history of Twitter users is useful in recognizing depression. The results showed that the activity history of users can be used to recognize the presence of depression with an accuracy of approximately 69%. The fact that similar results were obtained between user groups with different languages and backgrounds, as here and in [14], strongly suggests that using social media activity is a valid approach to recognizing depression. Sociological studies show that the rate of self-disclosure of Japanese is lower than that of American (e.g., [1]); i.e., Japanese people talk less about themselves than American people. In contrast, Japanese people post tweets very frequently. The ratio of Japanese tweets around the world is 14% in 2011[3] although the population of Japanese is only 130 million. It should be noted that using social media data is an effective approach for people with such different cultural backgrounds. Note that we do not claim that our models achieved higher performance than those in [14], but just claim that our models achieved comparable performance with [14]. While it is difficult to directly compare the results in [14] and ours since the data are different from each other, the values of precision and recall reported in [14] are higher than ours.

There were several challenges to apply the framework of [14] to Japanese-speaking twitter users, and these challenges might affect the performance of our models. Our models didn't include several features used in [14]. Choudhury *et al.* [14] show that usage of 1st person pronoun and usage of 3rd person pronoun are useful features. However, as linguistic researches discuss (e.g., [25]), Japanese personal pronouns are significantly different at least from those in the Western languages. Japanese have a large number of forms for each person pronoun. For instance, in [25], 21 types of Japanese 1st person pronoun are listed, and we can observe more variety of forms of pronoun in social media texts. Moreover, the subject words are often missing in Japanese texts, and using pronouns for subject words is sometimes unnatural particularly in twitter-like short and casual texts. We therefore decided not to include such features in our models. Moreover, to the best of our knowledge, there are no available standard Japanese dictionaries like ANEW lexicon [7], and we therefore used a simple list of positive and negative words. These limitations might affect our model performance.

While linguistic-based features and some dictionary-based features were not used in our study, we adopted a topic modeling approach and demonstrated the usefulness of it. Our experimental results showed that the model including topics as the features of SVM achieves 2% higher accuracy and 0.05 point higher F-measure than the model without topics (see Tab. 3). Resnik *et al.* [31] show the usefulness of a topic model when recognizing depression by analyzing essays written by university students. This study extends that result to Twitter. A single tweet is short, but because a large number of tweets (around 1,000 here) can be used, it is possible to extract topics successfully, which should make Twitter data useful for recognizing depression. Using topic model is expected to be a promising approach also for recognizing depression of non-Japanese twitter users since Resnik *et al.* [31] have already shown the usefulness of a topic model to English documents, and topic modeling approaches are successful in other tasks such as predicting stock market prices [33] and public health statistics [29] from social media text data. We therefore expect that the accuracy of the model found in [14] could be increased by including topics as a feature.

The experimental results of this paper suggest that the length of the selected activity history is important. De Choudhury *et al.* [14] set the observation period at one year, but our experimental results indicate that accuracy is likely to be improved

[3] `http://semiocast.com/en/publications/2011_11_24_Arabic_highest_growth_on_Twitter`

by adjusting this period appropriately. However, the period of tweets that is informative may differ by user. A more detailed analysis is necessary to isolate per-user differences.

This study has some limitations. First, there is still room for further improvement of the features used. De Choudhury *et al.* [14] reduced the dimensions of high-dimensional features by using principal component analysis and treated the results as SVM features. They also applied data abstraction techniques such as using the average, variance, and entropy of the features for each day. It is important to verify how effective these techniques would be on the data of this study.

It may also be important to demonstrate what kind of machine learning method is most effective for the purpose of recognizing depression. Rather than using only SVM, the use of the newest machine learning frameworks, such as deep learning [3, 4] and ensemble learning [8, 9], is expected to have the potential to further increase estimation accuracy. In particular, deep learning is known to be able to extract useful features from within a large quantity of features. The bag of words and other features that could not be applied effectively in this study may be more effective if deep learning is used in place of SVM.

Finally, this experiment was carried out using information from only 209 users; it is important to determine the degree to which the accuracy increases (if at all) by increasing the number of users to a larger number, such as 1,000 or 10,000 people. Collecting a large quantity of learning data is difficult to do because of the data characteristics, but the collection of larger-scale learning data may be important to obtain the best possible results from the approach described here.

CONCLUSION

This study investigated how useful the various features extracted from Twitter user history are for recognizing depression, and the degree of accuracy with which the presence of active depression could be detected by using these features. Our aim was to establish a method by which to recognize depression by analyzing the large-scale records of users' activities in social media. The following specific results were obtained: depression can be recognized in users with an accuracy of approximately 69%; topics extracted by a topic model are useful features; around two-months observation is sufficient for the tweets used in learning and prediction; and long observation periods may decrease accuracy.

The estimation of user depression over time is suggested for future research. The ability to estimate daily variations in depression may be a useful tool for self-diagnosis as well as for diagnosis by a medical professional. Determining techniques that could be used in a medical context to identify depression from social media users' activities is an important future task.

ACKNOWLEDGMENTS

The authors would like to thank participants of the experiments and ethical committee members. This work was partly supported by JSPS KAKENHI Grant Number 26870076 and Leave a Nest research grant Fuji Xerox award.

REFERENCES

1. Barry, D. T. Cultural and demographic correlates of self-reported guardedness among East Asian immigrants in the US. *International Journal of Psychology 38*, 3 (2003), 150–159.
2. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. An inventory for measuring depression. *Archives of General Psychiatry 4*, 6 (June 1961), 561–571.
3. Bengio, Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning 2*, 1 (Jan. 2009), 1–127.
4. Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 8 (Aug. 2013), 1798–1828.
5. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research 3* (Jan. 2003), 993–1022.
6. Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (Mar. 2011), 1–8.
7. Bradley, M., Lang, P., and Cuthbert, B. Affective norms for english words (ANEW). *Gainsville, Fla, NIMH Centre for the Study of Emotion and Attention, University of Florida* (1999).
8. Breiman, L. Bagging predictors. *Machine Learning 24*, 2 (Aug. 1996), 123–140.
9. Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct. 2001), 5–32.
10. Culotta, A. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14)* (Apr. 2014), 1335–1344.
11. De Choudhury, M., Counts, S., and Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)* (Apr. 2013), 3267–3276.
12. De Choudhury, M., Counts, S., and Horvitz, E. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci'13)* (May 2013), 47–56.
13. De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'14)* (Feb. 2014), 626–638.
14. De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)* (July 2013), 128–137.

15. Government of Japan. White paper on goverment measures for suicide, June 2014. Also available as `http://www8.cao.go.jp/jisatsutaisaku/whitepaper/index-w.html` (in Japanese).

16. Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101*, Suppl 1 (Apr. 2004), 5228–5235.

17. Hamilton, M. Development of a rating scale for primary depressive illness. *British Journal of Social & Clinical Psychology 6*, 4 (Dec. 1967), 278–296.

18. Houston, T. K., Cooper, L. A., Vu, H. T., Kahn, J., Toser, J., and Ford, D. E. Screening the public for depression through the Internet. *Psychiatric Services 52*, 3 (Mar. 2001), 362–367.

19. Inui, T., Umezawa, Y., and Yamamoto, M. Complaint sentence detection via automatic training data generation using sentiment lexicons and context coherence. *Journal of Natural Language Processing 20*, 5 (Dec. 2013), 683–705. (in Japanese).

20. Kudo, T. Mecab: Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net/`.

21. Marcus, M., Yasamy, M. T., van Ommeren, M., Chisholm, D., and Saxena, S. Depression: A global public health concern. Tech. rep., WHO Department of Mental Health and Substance Abuse, Oct. 2012. Also available as `http://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf`.

22. Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., and Becker, T. Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depression and Anxiety 28*, 6 (June 2011), 447–455.

23. O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)* (May 2010), 122–129.

24. OECD. Health at a glance 2013, Nov. 2013.

25. Ono, T., and Thompson, S. A. Japanese (w)atashi/ore/boku 'I': They're not just pronouns. *Cognitive Linguistics 14*, 4 (2003), 321–348.

26. Pak, A., and Paroubek, P. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)* (May 2010), 1320–1326.

27. Park, M., Cha, C., and Cha, M. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD'12)* (Aug. 2012), 1–8.

28. Park, S., Lee, S. W., Kwak, J., Cha, M., and Jeong, B. Activities on Facebook reveal the depressive state of users. *Journal of Medical Internet Research 15*, 10 (Oct. 2013), e217.

29. Paul, M. J., and Dredze, M. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)* (July 2011), 265–272.

30. Radloff, L. S. The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement 1*, 3 (1977), 385–401.

31. Resnik, P., Garron, A., and Resnik, R. Using topic modeling to improve prediction of neuroticism and depression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)* (Oct. 2013), 1348–1353.

32. Rihmer, Z. Can better recognition and treatment of depression reduce suicide rates? A brief review. *European Psychiatry 16*, 7 (Nov. 2001), 406–409.

33. Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)* (Aug. 2013), 24–29.

34. Simon, G., Goldberg, S., Tiemens, B., and Ustun, T. Outcomes of recognized and unrecognized depression in an international primary care study. *General Hospital Psychiatry 21*, 2 (Mar. 1999), 97–105.

35. Takamura, D., Inui, T., and Okumura, M. Extracting semantic orientations using spin model. *IPSJ Journal 47*, 2 (Feb. 2006), 627–637. (in Japanese).

36. Tiemens, B., Ormel, J., Jenner, J., Van der Meer, K., Van Os, T., Van Den Brink, R., Smit, A., and Van den Brink, W. Training primary-care physicians to recognize, diagnose and manage depression: does it improve patient outcomes? *Psychological Medicine 29*, 04 (July 1999), 833–845.

37. Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., and Ohsaki, H. On estimating depressive tendency of twitter users from their tweet data. In *Proceedings of the 2nd International Workshop on Ambient Information Technologies (AMBIT'12)* (Mar. 2013), 29–32.

38. Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)* (May 2010), 178–185.

39. Zhang, X., Fuehres, H., and Gloor, P. Predicting stock market indicators through twitter " I hope it is not as bad as I fear ". *Procedia-Social and Behavioral Sciences 26* (2011), 55–62.

40. Zhung, W. W. K. A self-rating depression scale. *Archives of General Psychiatry 12*, 1 (1965), 63–70.