Machine Learning Approach for Depression Detection in Japanese

Yuka Niimi
Graduate School of Social Informatics
Aoyama Gakuin University
c8121005@aoyama.jp

Yutaka Miyaji
Graduate School of Social Informatics
Aoyama Gakuin University
uta@si.aoyama.ac.jp

Abstract

In this paper, we describe the results of our research aimed at detecting depression in Japanese sentences. In the United States and elsewhere, machine learning approaches to detect depression from language have been demonstrated. However, in Japanese text, there are only two studies that have addressed the detection of depression. In this study, to detect depression based on linguistic features, even in documents that do not explicitly mention the topic of depression, we build a machine learning model that detects depression in Japanese by eliminating topics that suggest depression or depression. We also examine the accuracy when the parts of speech are limited and when the number of labels is aligned. In the performances of our models with 5-fold cross-validation, we were able to obtain a high evaluation.

1 Introduction

Depression is a common illness that affects over 264 million people worldwide. Depression and other mental health conditions are becoming more common around the world. At its worst, depression can lead to suicide. A resolution passed by the World Health Assembly in 2013 advocated for a comprehensive, coordinated approach to mental illness at the national level[1].

According to a patient survey conducted by the Ministry of Health, Labor and Welfare (MHLW), an official organization in Japan, the total number of patients with mood disorders such as depression in Japan has increased 2.9 times in 21 years, from 433,000 in 1996 to 1,276,000 in 2009. Depression also affects the economy, and the annual social loss due to depression in Japan is estimated to be 2 trillion yen (Sato, 2014). Therefore, "stress check system"[2] has been introduced into the public system as a mandatory requirement for business establishments with 50 or more employees since 2015 to prevent mental health problems such as depression. As can be seen from these examples, depression is an issue of high importance in Japan.

In psycholinguistics, one of the behavioral indicators of depression is linguistic features. (Fine, 2008) reported that the idea that language use reflects the speaker's mind. Psychiatrists also have used it to assess mental health conditions. Many studies focus on the linguistic features of depression and analyze natural language. (Yip, 2018) manifested the communicative patterns of anxiety and depression communication using the framework of discourse analysis. Unfortunately, lack of resources, a shortage of trained healthcare providers, and the social stigma associated with mental illnesses are all obstacles to the effective treatment of depression. Therefore, They sometimes miss the appropriate time for taking care of depression. Detecting depression with machine learning will become increasingly important in the future.

In this study, we use blog data to limit topics with LDA, vectorize documents with TF-IDF, reduce dimensions with SVD, and build classifiers with SVM.

[1] https://www.who.int/news-room/fact-sheets/detail/depression

[2] https://www.mhlw.go.jp/stf/houdou/0000082587.html

Our contributions are summarized as follows:

- We demonstrated the usefulness of depression detection in Japanese.
- We focused on eliminating not only depression but also topics associated with depression, which has not received much attention in the research community.
- After erasing the topics, we found that adjusting the number of labels in the control group to the same number as in the depressed group produced more convincing results.
- We found that the linguistic features of depression require more strength than minimal preprocessing.

2 Related Work

A myriad of studies also has been conducted to detect depression from natural language (Rinaldi et al., 2020; Morales et al., 2018). The most widely used feature engineering method is to extract lexical features from the Linguistic inquiry word count (LIWC) lexicon, which contains over 32 psychological construct categories (Pennebaker et al., 2007). Many studies have applied depression detection from natural language, such as predicting depression from social media (Choudhury et al., 2013; Song et al., 2018; Orabi et al., 2018), analyzing the impact of COVID-19 using a depression predictor (Wolohan, 2020) and detecting suicidal attempt (Coppersmith et al., 2016; Gao et al., 2017). In the United States and elsewhere, machine learning approaches to detect depression from language have been demonstrated. However, only two studies (Hiraga, 2017; Tsugawa et al., 2015) have addressed the detection of depression from natural language in Japanese. Even though most foreign studies can provide publicly available datasets, there are currently no datasets on depression in Japanese. (Hiraga, 2017) obtained blog data from 111 people and used a total of 7,358 articles to build a corpus. After eliminating topics related to depression and comparing the accuracy of several algorithms, the best-performing model had an accuracy of 95.5. (Tsugawa et al., 2015) obtained Twitter data from 209 volunteers and built a corpus of up to 3,200 Tweets per person. They also used the user's behavioral history as a feature, resulting in an accuracy of 69 for the model they built.

In this paper, we present a machine learning approach for depression detection in Japanese which departs from this previous work in four main ways:

First, in this study, we prepare more training data than in previous studies. (Tsugawa et al., 2015) uses data from Twitter. Although depression detection from multimedia such as Twitter is imperative, this study aims to improve the accuracy by using blog data and increasing the number of characters. (Hiraga, 2017) uses the same data from blogs as in this study. In this study, we increase the data from (Hiraga, 2017) to verify accuracy. We obtain the data of approximately 900 people from blogs and build a corpus using approximately 300,000 articles.

Second, in this study, we compare the use of limited parts of speech with the use of all parts of speech. Many related studies have limited the number of parts of speech, believing that it is possible to capture the characteristics with a small number of parts of speech, and others have maximized the use of parts of speech, believing that all parts of speech, such as prepositions and conjunctions, have essential functions. In this study, we compare the use of limited parts of speech with the use of all parts of speech.

Third, our approach erases not only depression but also topics that are associated with depression. It is indispensable to detect depression from linguistic features, even if they are not directly related to depression. Hence, some studies address this issue by eliminating topics related to depression. In many studies, topics related to "depression" are only eliminated. However, topics that include words that are far from the control group, such as "sleeping pills," "hospital," and "dying," are also considered to be topics that indirectly suggest depression. To detect depression from topics that are similar to those of the control group, in this study, we eliminate topics that are associated with depression as well as depression itself.

Fourth, in this study, we adjust the number of the control group and the number of the depression group to be the same when the above topics are eliminated. If we obtain valuable results, it could be due

to an imbalance in the ratio of the number of data in the depressed group to the number of data in the control group. Numerous studies have not corrected this point. Thus, in this study, we adjust the number of labels when we eliminate the topics related to depression.

3 Data

This study, obtained data on depression patients from TOBYO : Depression Fighting Blog[3], one of the largest disease-fighting portals to access patients' valuable experiences and knowledge. This site compiles articles on depression from several blog platforms. We obtained 441 users' information from Ameba blog[4], which does not prohibit scraping, has not terminated its service, and has a large number of users. All the articles of the users were retrieved, and 149,997 articles were retrieved. For data on people who are not depressed, we randomly selected 460 users from several Ameba blog genres like Marriage/Pregnancy/Child Care, Lifestyle, Married Couples, Pets, Entertainment/Hobbies, Travel/Regional, and Fashion/Cosmetics to get a total of 166,312 articles.

4 Method

The purpose of this study is to detect depression in Japanese. We construct a model that detects depression based on linguistic features, even in articles that do not explicitly refer to depression. In this study, we perform the following three tasks.

1. When we limit the parts of speech, we examine our models.

2. When we remove topics related to depression and topics associated with depression, we examine our models.

3. When the number of data in the depressed and control groups is the same, we compare our models.

4.1 Preprocessing

To avoid removing necessary information from the analysis target, we perform the minimum necessary preprocessing. First, we remove pictograms, URLs, stop words, convert numbers, and normalize words.

Unlike English, which has clear word boundaries in most text, Japanese does not have clear word boundaries. In Japanese, the process of word segmentation is more complex than in other languages. The merit of the N-gram method used in the previous research (Hiraga, 2017) is that it does not cause retrieval omissions because the documents are mechanically divided and written. On the other hand, the disadvantage is that it generates much noise. For example, if we search for a document with the keyword "Kyoto," we mistakenly get a hit for "Kyoto," which is the last two letters of "Tokyo-to." Because POS-based models have the highest accuracy in previous studies, we use the way separated by parts of speech in this study. We use a Japanese tokenizer, called sudachi[5], which is resistant to word shaking.

For Task 1, "When we limit the parts of speech, we examine our models.", we prepare data using all parts of speech and data limited to only nouns, verbs, adjectives, adjectival verbs, and adverbs. We use articles with more than 50 words. For the data with all parts of speech, the number of depressive labels is 33,806, and the number of control labels is 47,275. For the data with limited parts of speech, the number of depression labels was 25,704, and the number of control labels was 38,442.

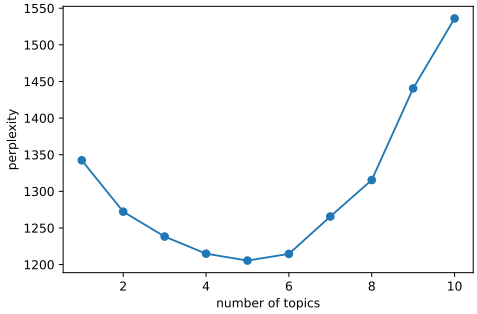
4.2 Depression topic withheld by LDA

For Task 2, "When we remove topics related to depression and topics associated with depression, we examine our models.", we eliminate articles by Latent Dirichlet Allocation (LDA) that explicitly mention depression or topics that suggest depression. LDA is a probabilistic generative model that assumes that a document consists of multiple topics. (Resnik et al., 2015) uses LDA for depression detection. As a clue to determine the number of topics, we use perplexity to measure the accuracy of the probabilistic model. Figure 1 shows the change in perplexity when the number of topics = k is varied. In both the case of using all parts of speech and the case of limiting the number of parts of speech, the minimum value was obtained when k=5. For this reason, we set the number of topics to 5 in both cases. The

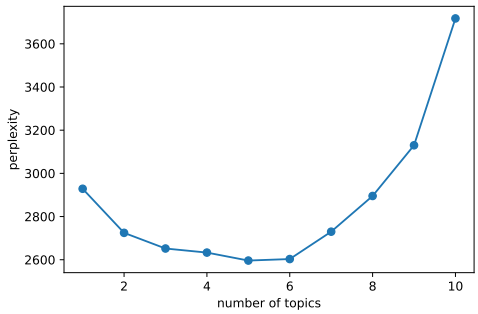
[3]https://www.tobyo.jp/

[4]https://www.ameba.jp/

[5]https://github.com/WorksApplications/Sudachi



(a) all parts of speech



(b) limited parts of speech

Figure 1: Relationship between number of topics and perplexity.

topic with the highest percentage of the five topics is considered the blog post's topic, and we classify the blog articles by topic.

Table 1 shows the top 10 words for each topic using all parts of speech. When the top 100 word probabilities are displayed, Topic2 is the only topic containing "depression." In addition to "depression," words such as "hospital," "medicine," "symptom," "disease," "spirit," "live," "work," and "anxiety" were also observed. Although it did not contain the word "depression," Topic 1 contained words that were associated with depression, such as "medicine," "sleeping pills," and "prescription. On top of that, Topic 1 included words such as "diet" and "teacher. Thus, we eliminate topic 1 and topic 2.

Table 2 shows the top 10 words for each topic when the parts of speech are limited to nouns, verbs, adjectives, adjectival verbs, and adverbs. When the top 100 word probabilities are displayed, Topic1 is the only topic containing "depression." In addition to "depression," we observed other words such as "company," "medicine," "symptom," "disease," "spirit," "die," "work," "understanding," "family," "work hard," "stress," and "disorder." On top of that, although it did not contain the word "depression," Topic 5 contained words associated with depression such as "medicine," "sleeping pills," "hospital," "mental," "examination," and "prescription. Moreover, Topic 5 included words such as "teacher," "work hard," "weight," "tired," and "work. Thus, we eliminate topic 1 and topic 5.

4.3 Classification

For Task 3, "When the number of data in the depressed group and the control group is the same, we compare our models.", the first step is to align the number of labels for all the data to be compared. This study vectorizes sentences using a count-based method, TF-IDF, the inverse of Term Frequency and Document Frequency. This method evaluates the importance of a word by multiplying its Term Frequency by its Inverse Document Frequency. Because TF-IDF has as many features as there are morphemes, it is likely to be a sparse matrix, and the processing time will belong. To reduce the amount of computation, we use singular value decomposition (SVD) to reduce the dimensionality to a low level. SVD is a method of matrix factorization in linear algebra for matrices with complex or real components. This SVD brings the dimensionality to 250 dimensions. We then use the Support Vector Machine (SVM) classification method to perform binary classification tasks between the depressed and control groups. SVM is a pattern recognition method that uses supervised learning based on margin maximization.

5 Results

The evaluation indices are accuracy and F1 value. For each evaluation index, we took the average value of five parts by cross-validation. Table 3 shows the results of all the models.

Task 1, "When we limit the parts of speech, we examine our models." is to examine the accuracy of model1(all topic / all parts of speech) and model2(all

topic1 n=6808	**topic2 n=15879**	**topic3 n=8303**	**topic4 n=18077**	**topic5 n=4801**
kusuri(drugs)	omou(think)	hito(people)	o(o)	o(o)
demo(but)	watashi(I)	watashi(I)	kyou(today)	burogu(blog)
toki(time)	jibun(myself)	omou(think)	hi(dat)	ne(ne)
nomu(drink)	nu(not)	kedo(but)	iku(go)	he(he)
taberu(eat)	iu(say)	iu(say)	toki(time)	wa(wa)
neru(sleep)	hito(people)	ne(ne)kuru	(come)	tuki(month)
nu(not)	kuru(come)	muru(see)	nu(not)	mura(village)
ne(ne)	nani(what)	yo(yo)ne	(ne)	watashi(I)
hi(day)	ne(ne)	nu(not)	omou(think)	toshi(year)
omou(think)	ima(now)	ya(ya)	taberu(eat)	sama(sama)

Table 1: Using all parts of speech Top 10 probabilities of a word occurring in a topic.

topic1 n=13130	**topic2 n=3863**	**topic3 n=5078**	**topic4 n=18328**	**topic5 n=13469**
omou(think)	burogu(blog)	iu(say)	kyou(today)	toki(time)
jibun(myself)	honjitsu(today)	miru(see)	iku(go)	nomu(drink)
hito(people)	mura(village)	omou(think)	hi(day)	neru(sleep)
iu(say)	en(yen)	sensei(teacher)	omou(think)	taberu(eat)
ima(now)	gozaru(gozaru)	hito(people)	iu(say)	kyou(today)
kuru(come)	rankingu(ranking)	kiku(listen)	kaeru(go home)	hi(day)
kangaeru(consider)	itasu(itasu)	hanashi(story)	kau(buy)	kusuri(drug)
sigoto(work)	ouen(support)	nihon(japan)	toki(time)	omou(think)
dou(how)	itadaku(itadaku)	jyosei(woman)	ie(house)	jikan(time)
koto(koto)	koto(koto)	kuru(come)	deru(get out)	okiru(wake up)

Table 2: Using limited parts of speech Top 10 probabilities of a word occurring in a topic.

topic / limited parts of speech). When model1 and model2 are compared, model2 has higher accuracy, which indicates that higher accuracy can be obtained by limiting the parts of speech.

In Task 2, "When we remove topics related to depression and topics associated with depression, we examine our models.", we compared model1(all topic / all parts of speech) with model3(depression topic withheld / all parts of speech), and model2(all topic / limited parts of speech) with model4(depression topic withheld / limited parts of speech). When comparing model1(all topic / all parts of speech) and model3(depression topic withheld / all parts of speech), the model's accuracy using the data of all topics was higher than the model eliminating depression and depression-related topics. This result indicates that topics related to depression provide cues to recognize characteristics of depressed groups using all parts of speech. On the other hand, when comparing model2(all topic / limited parts of speech) and model4(depression topic withheld / limited parts of speech), the accuracy of the model that eliminated the topics related to depression and depression was higher than the model that used the data of all topics. When we limited the part-of-speech, the accuracy was lower when we deleted topics related to depression. However, this may be because the percentage of data showing the control group became larger by eliminating topics related to depression. We will demonstrate this problem in Task 3.

In Task 3, "When the number of data in the depressed group and the control group is the same, we compare our models.", we compare model1(all topic / all parts of speech) and model5(all topic / all parts of speech / adjustment), model2(all topic / limited parts of speech) and model6(all topic / limited parts of speech / adjustment), model3(depression topic withheld / all parts of speech) and model7(depression topic withheld / all parts of speech / adjustment), and model4(depression topic withheld / limited parts of speech) and model8(depression topic withheld / limited parts of speech / adjustment). The scores for model1 and model5, model2 and model6, and model3 and model7 varied slightly. On the other hand, significant changes were observed in model4(depression topic withheld / limited parts of speech) and model8(depression topic withheld / limited parts of speech / adjustment), indicating that a large number of the control group improved the accuracy. The model with the highest accuracy in this study is Model 4, but the most convincing model is Model 8. Several studies eliminate topics related to depression, but no studies have subsequently adjusted the number of labels. We found that adjusting the number of labels when eliminating topics related to depression produced more valuable results.

The most valuable and accurate model was the model trained by eliminating depression and depression-related topics, keeping the depressed group and the control group equal, and limiting the parts of speech to nouns, verbs, adjectives, adjectival verbs, and adverbs, which resulted in its accuracy of 95.6 and its F1 value of 95.9. These results are higher than those of previous studies (Hiraga, 2017; Tsugawa et al., 2015).

6 Conclusion

In this study, we constructed a machine learning model for detecting depression from natural language to detect depression in Japanese. In Task 1, "When we limit the parts of speech, we examine our models.", we found that by limiting the parts of speech, the features of the depressed group can be better recognized. In Task 2, "When we remove topics related to depression and topics associated with depression, we examine our models.", we confirmed the accuracy by eliminating topics related to depression and depression, and found that we could detect depression from linguistic features even in materials where the topic of depression is not explicitly mentioned. In Task 3, "When the number of data in the depressed group and the control group is the same, we compare our models.", we found that we can build a more convincing model by balancing the labels.

The most valuable and accurate model was trained by eliminating depression and depression-related topics, keeping the number of depressed and the control group equal, and limiting the parts of speech to nouns, verbs, adjectives, adjectival verbs, and adverbs. Its accuracy was 95.6, and its F1 value was

Model	Accuracy	F1
model1(all topic / all parts of speech)	.922	.921
model2(all topic / limited parts of speech)	.923	.924
model3(depression topic withheld / all parts of speech)	.899	.876
model4(depression topic withheld / limited parts of speech)	.971	.962
model5(all topic / all parts of speech / adjustment)	.921	.921
model6(all topic / limited parts of speech / adjustment)	.925	.925
model7(depression topic withheld / all parts of speech / adjustment)	.884	.884
model8(depression topic withheld / limited parts of speech / adjustment)	.956	.959

Table 3: Performance of our model with 5-fold cross-validation.

95.9. This study shows that it is possible to extract the linguistic features of depression from Japanese documents, even if there are no topics related to depression. This study will help detect depression in people who are not aware of it and establish a system to prevent depression.

References

Mitsuhiro Sado Frontiers of Psychiatry. 2014. *Journal of Neuropsychiatry*, 16(2):107–115. (in japanese)

Fine, Jonathan. Language in psychiatry: A handbook of clinical practice. 2006. *Equinox Publishing*

Yip, Jesse Wai Chi Communicating Social Support in Online Self-help Groups for Anxiety and Depression: A Qualitative Discourse Analysis. 2018. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*

Rinaldi, Alex and Tree, Jean E Fox and Chaturvedi, Snigdha. Predicting Depression in Screening Interviews from Latent Categorization of Interview Prompts. 2020. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 7–18

Morales, Michelle and Scherer, Stefan and Levitan, Rivka. A linguistically-informed fusion approach for multimodal depression detection. 2018. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* 13–24

Pennebaker, James W and Chung, CK and Ireland, M and Gonzales, A and Booth, RJ The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin. 2007. *Technical Report 2.*

De Choudhury, Munmun and Gamon, Michael and Counts, Scott and Horvitz, Eric. Predicting depression via social media. 2013. *Seventh international AAAI conference on weblogs and social media*

Song, Hoyun and You, Jinseon and Chung, Jin-Woo and Park, Jong C Feature Attention Network: Interpretable Depression Detection from Social Media. 2018. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*

Orabi, Ahmed Husseini and Buddhitha, Prasadith and Orabi, Mahmoud Husseini and Inkpen, Diana Deep learning for depression detection of twitter users. 2018. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* 88–97

Wolohan, JT Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. 2020. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*

Coppersmith, Glen and Ngo, Kim and Leary, Ryan and Wood, Anthony Exploratory analysis of social media prior to a suicide attempt. 2016. *PProceedings of the third workshop on computational linguistics and clinical psychology* 106–117

Gao, Yuanbo and Li, Baobin and Wang, Xuefei and Wang, Jingying and Zhou, Yang and Bai, Shuotian and Zhu, Tingshao Detecting suicide ideation from Sina microblog. 2017. IEEE *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 182–187

Hiraga Misato Predicting depression for japanese blog text. 2017. *Proceedings of ACL 2017, Student Research Workshop*, 107–113.

Tsugawa, Sho and Kikuchi, Yusuke and Kishino, Fumio and Nakajima, Kosuke and Itoh, Yuichi and Ohsaki, Hiroyuki Recognizing depression from twitter activity. 2015. *Proceedings of the 33rd annual ACM conference on human factors in computing systems* 3187–3196

Resnik, Philip and Armstrong, William and Claudino, Leonardo and Nguyen, Thang and Nguyen, Viet-An and Boyd-Graber, Jordan Beyond LDA: exploring supervised topic modeling for depression-related lan-

guage in Twitter. 2015. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* 99–107