Predicting Depression for Japanese Blog Text

Misato Hiraga
Indiana University
mhiraga@indiana.edu

Abstract

This study aims to predict clinical depression, a prevalent mental disorder, from blog posts written in Japanese by using machine learning approaches. The study focuses on how data quality and various types of linguistic features (characters, tokens, and lemmas) affect prediction outcome. Depression prediction achieved 95.5% accuracy using selected lemmas as features.

1 Introduction

The World Health Organization (WHO) recognizes that depression is a leading cause of ill health and disability (2017). In Japan, it is also the most frequent reason for sick leave from work (Kitanaka, 2012). However, many people with depression may not be aware that their mood change and fatigue are due to depression. In order to offer help for those who need it, we first need to identify them. This study examines whether linguistic features in written texts can help predict whether the author is depressed by using a supervised machine learning approach. Specifically, we examine the effectiveness of morphological (character *n*-grams), syntactic (token *n*-grams), and (syntactic-)semantic (lemmas of selected POS categories) features. In addition, we remove the topic bias so that the methods can be used to predict depression in people who do not know they are depressed and thus do not write about depression. The results show that lemmas from verb and adverb categories improve performance in classifying authors. Additionally, the selected words include words not typically thought of as related to depression. Thus, the study suggests that feature engineering should not be constrained by our notion of what would be related to certain mental conditions or personalities as changes in people's language use may be very subtle.

Section 2 discusses previous work on author profiling and depression detection. In Section 3, we describe data acquisition, topic modeling, and classifications with different features. Section 4 summarizes the results, and Section 5 discusses the results. Finally, Section 6 concludes the paper with a summary and an outlook.

2 Related Work

Language and social media activities have been utilized for author profiling including personality prediction (Bachrach et al., 2012; Golbeck et al., 2011). Although depression is not a personality, the studies in personality prediction can be extended to predicting depression since one's mental state is often reflected in his or her language and social activities.

Character *n*-grams are reported to do well in gender prediction in English blog text (Sarawgi et al., 2011) and personality prediction in Dutch (Noecker et al., 2013). However, the PAN 2014 challenge (Rangel et al., 2014) reports that character *n*-grams are not useful in author profiling (age and gender) in English and Spanish social media, Twitter, blogs, and hotel reviews. Japanese is different from Germanic or Romance languages in terms of how much information can be encoded in one character. Japanese basic characters, *hiragana* and *katakana*, represent one mora, which is a sound unit similar to a syllable, and *kanji* (Chinese characters) can encode more than one mora. Moreover, there are many characters: 50 each for *hiragana* and *katakata*, and approximately 2000 *kanji* for everyday writing. The current study thus examines the effectiveness of character *n*-grams as features in Japanese text classification.

Matsumoto et al. (2012) built a classifier to

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 107–113
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-3018

	Class	# of Author	Word/Author	# of Document	Word/Doc
All	Depressed	51	3666	842	222
	Non-Depressed	60	3692	1020	220
Topic	Depressed	49	2630	739	192
	Non-Depressed	59	2890	904	191

Table 1: Number of authors and documents and average word count before and after topic modeling

predict whether a blog text is written by a depressed author or a non-depressed author, using approximately 1800 blog texts written by 30 depressed authors and 30 non-depressed authors. They experimented with bag-of-word (BOW) features and also examined whether words that are associated with emotions from the JAppraisal dictionary (Sato, 2011) were useful in classification. Their classifier using Naive Bayes and BOW performed the best, resulting in average of 88.1% accuracy with 10-fold cross validation. One would question whether their prediction with BOW was based on topic, however. By browsing blogs written by depressed authors, we find many blogs are about depression. Therefore, texts written by depressed authors are biased towards the topic of depression. Unless Matsumoto et al. (2012) controlled topic-bias, their system may be heavily affected by topic. The goal of this study is to predict depression from topic-general texts. Thus, the current study controls topic by performing topic modeling (Section 3.2) before classification and examines its effect. As their model using emotion-related words did worse (<70%) than BOW, the current study examines various features that are not obviously related to depression.

3 Method

3.1 Data

Blog texts are collected from a blog ranking site[1] and by searching on blog provider websites (Yahoo Japan, Livedoor, Hatena, FC2, Seesaa, Nifty, Muragon, and Ameblo)[2]. The blog ranking site ranks registered blogs by the number of votes from readers who visit the blogs. The site is divided into categories which include "depression". In this category, most authors state that they are diagnosed with or have been suffering from depression. Blogs are chosen to be included in the study if the authors report that they themselves are suffering from depression and have written their blogs for at least three months. Some authors write only a little in a month, but most write at least 10 entries within three months. Thus, for each author, three months' worth of blog posts are collected. This "depressed" group consists of 51 authors. Three months' worth of texts for 60 "non-depressed" authors are also collected. They are randomly chosen among those who have a similar profile as the depressed authors. For example, if a depressed author is a male in his 40's, a blog author who had the similar profile is chosen. Moreover, non-depressed authors with the same interest as the depressed authors are collected. Their interests include pets, food, and sports.[3] As many blogs written by depressed authors are retrieved from a blog ranking site, they include fixed phrases, such as "Please vote", to encourage their readers to vote for their blogs. These fixed phrases that appear repeatedly are removed as they do not appear in blogs written by the non-depressed authors. Moreover, a document whose file size is smaller than 100 bytes is removed as it contained very few words. The average number of words and characters per author and per document are in Table 1.

3.2 Removing Blog entries on depression

Before performing classification, topic modeling is performed to divide documents into topic classes. The purpose of topic modeling is to remove blog entries that are biased towards the topic of depression so that classification will not classify documents based on topics (see Section 2). MALLET (McCallum, 2002) is used for topic-modeling. Dividing documents into 5 topics with a hyperparameter optimization value of 50 is found to work the best by manually testing different values. Two of the five topics included topic keys that are related to depression, and thus the documents in those two topics are excluded in the study. The numbers are summarized in Table 1.

[1] http://mental.blogmura.com/utsu/

[2] Data by Matsumoto et al. (2012) was not available

[3] We are aware that we cannot guarantee that non-depressed authors are not depressed.

	Class	# of Author	Word/Author	# of Document	Word/Doc
Train	Depressed	39	2404	590	193
	Non-Depressed	47	2466	722	195
Test	Depressed	10	3512	149	186
	Non-Depressed	12	4233	182	177

Table 2: Number of authors and documents and average word count in training and test sets

The number of authors is reduced to 49 depressed and 59 non-depressed authors. The reason for the reduction in non-depressed authors seemed to be because documents that are related to companies or work are classified together with the topic of depression. For example, one topic class which include the topic key "depression" also includes keys, such as "company", and "investment".

3.3 Classification

We perform classification of texts into whether the author is depressed or not. Our system learns from texts written by a group of both depressed and non-depressed authors, and classifies unseen texts written by a different group of depressed and non-depressed authors. We perform classification of texts per author (henceforth, *author-level* classification) and per document (*document-level* classification). In author-level classification, one document contains all the blog entries written by one author. In document-level classification, one document contains one blog entry. For both experiments, data are divided into a training and a testing set. In document-level classification, documents written by the authors in the training set do not appear in the test set, and thus none of the authors have their documents in both training and testing sets. The number of documents are summarized in Table 2.

Classification is performed using Multinomial NaiveBayes (NB), Linear Support Vector Machines (SVM), and Logistic Regression (LR) in scikit-learn (Pedregosa et al., 2011). Multinominal NB is used with the default alpha value (alpha=1.0), and SVM and LR classifiers are both used with the default regularization value (C=1). Univariate and model-based feature selection is performed for each experiment. In univariate feature selection, features that are above the 75 percentile are chosen. For model-based feature selection, SVM and LR, both with penalty of L1 and C=1 are used to select features with non-zero coefficients.

3.4 Features

3.4.1 Character *n*-grams

The first feature set is character *n*-grams. Character n-grams worked well in other languages as discussed in Section 2, and they provide a generic cross-linguistic way of getting at morphological units. To test the effects of the writing system, two types of character *n*-grams, Japanese character *n*-grams and Romanized Japanese *n*-grams, are used. Japanese has three types symbols (hiragana, katakana, and kanji (Chinese characters)), and thus one word could be written in several ways. For example, the word *kawaii* "cute" could be written all in hiragana or katakana, or combination of hiragana and katakana, or combination of kanji and hiragana. Without Romanization, all of them would be treated differently despite their shared meaning. However, Romanization can also collapse words that are pronounced the same but written differently with different meanings. *hashi* can mean "chopsticks", "bridge", and "edge" depending on which Chinese characters are used and in what context they are used. To experiment with Romanized Japanese character *n*-grams, Japanese characters are converted to Romaji (Roman alphabets) using jConverter[4]. The value of *n* ranges between 1 and 10, and only features of one *n* value are used as features in one experiment, and features of different values of *n* are not combined together, to avoid the number of features from becoming too large. For instance, an experiment with trigrams only uses trigram features.

3.4.2 Token *n*-grams

The next feature set is token *n*-grams with varying values of n (1-10). Tokens retain inflections and conjugation, so token *n*-grams represent syntactic properties of written text. The Japanese text is tokenized with Cabocha (Kudo and Matsumoto, 2002), as Japanese does not use a space to indicate word boundaries.

[4] http://jprocessing.readthedocs.io/en/latest/#id2

	Author			Document		
Features	NB	SVM	LR	NB	SVM	LR
CharUni	68.2	54.5	54.5	65.3	63.8	67.2
CharUni+FS	63.6 (U)	68.2 (U)	68.2 (U)	65.6 (U)	65.9 (L)	67.2 (L)
CharN(n)	81.8 (3)	59.1 (5)	54.5 (3)	70.0 (3)	69.7 (10)	70.3 (7)
CharN+FS	**86.4 (U)**	59.1 (U)	72.7 (U)	**75.5 (U)**	73.4 (U)	75.2 (U)
RomUni	56.5	52.5	69.6	50.2	56.8	56.8
RomUni+FS	65.2 (L)	60.9 (S)	69.6 (L)	52.6 (L)	57.1 (L)	57.1 (S)
RomN(n)	82.6 (5)	60.9 (3)	78.3 (2)	71.6 (5)	68.3 (8)	70.4 (8)
RomN+FS	**82.6 (U)**	60.9 (L,U)	65.2 (L,S)	**71.9 (U)**	68.9 (U)	70.7 (U)

Table 3: Accuracies (%) for character *n*-grams. CharUni and CharN: Japanese character uni- and *n*-grams. RomUni and RomN: Romanized character uni- and *n*-grams. FS:Feature Selection. The value in parentheses indicates the best value of *n* for *n*-gram and a method for feature selection (L:Logistic Regression, S:SVM, U:Univariate)

	Author			Document		
	NB	SVM	LR	NB	SVM	LR
TokenUni	**86.4**	54.5	59.1	63.2	58.9	62.0
TokenUni+FS	**86.4 (U)**	72.7 (U)	72.7 (U)	64.8 (L)	62.6 (S)	65.1 (S)
TokenN (n)	77.3 (2)	54.5 (8)	59.1 (2)	66.0 (2)	63.2 (3)	66.7 (3)
TokenN+FS	81.8 (U)	72.7 (S)	81.8 (U)	65.7 (U)	**68.8 (U)**	67.3 (U)

Table 4: Accuracies (%) for Token and Token *n*-grams. The value in parentheses indicates the value of n for *n*-grams or model of feature selection (L:Logistic Regression, S:SVM, U:Univariate)

3.4.3 Lemmas and selected lemmas

The next feature set is lemmas. As lemmatization suppresses inflections, lemmas represent use of words regardless of their form in a sentence. Given this semantic nature and the results of token *n*-grams (Section 4.2), we only examine lemma unigrams. In addition, certain types of words may convey more relevant information than others, and thus in order to find whether certain categories of words are more informative in classification, POS categories are used to extract groups of words. First, words are POS-tagged with Cabocha (Kudo and Matsumoto, 2002), and words from each POS category (e.g. Noun) are used as features. Then, all possible combinations of 13 POS categories (Noun, Verb, Auxiliary, Adverb, Adjective, Particle, Symbol, Filler, Rentaishi[5], Conjunction, Affix, Interjection, Other) are created and a feature set containing words from each set of combined POS categories is evaluated.

[5] *Rentaishi* is a category of words that are not adjectives but modify nouns.

4 Results

4.1 Character *n*-grams

The accuracy scores for Japanese character *n*-grams and Romanized character *n*-grams are summarized in Table 3. Selected trigrams (97,992 features) achieved an accuracy of 86.4% with NB in author-level classification. Trigrams selected by univariate feature selection (114,303 features) worked best for the document-level classification, resulting in 75.5% accuracy. Romaji *n*-grams achieved similar accuracies, but they were below Japanese character *n*-grams.

4.2 Token *n*-grams

Table 4 shows the results of classification with token *n*-grams as features. Token unigrams with the NB classifier yielded 86.4% with (14,656 features) or without feature selection (10,992 features) in author-level classification, which was the same accuracy as the model with selected character trigrams. For the document-level classification, SVM with selected token trigrams (124,528 features) worked the best (68.8%) though it was not as good as the accuracy obtained from character trigrams.

	Author			Document		
	NB	SVM	LR	NB	SVM	LR
Lemma	81.8	50.0	59.1	66.8	57.8	60.7
Lemma+FS	81.8 (U)	68.2 (L)	77.3 (U)	68.1 (U)	63.6 (U)	64.2 (L)
POS feature	V,Adv	N,Adv, Ren	N,Adv,Ren, Sym,Fill	Adj,Aux, V,Ren	Aux,V, Ren	Aux,V, Ren,Fill
POS	**95.5**	77.3	81.8	**69.0**	64.5	63.6
POS+FS	**95.5 (U)**	90.9 (L)	86.4 (S)	67.4 (U)	67.1 (L)	67.4 (L)

Table 5: Accuracies (%) for lemmas and lemmas of POS categories with the highest accuracy. The character in parentheses shows a model of feature selection (L:Logistic Regression, S:SVM, U:Univariate)

	Depressed		Non-Depressed	
Verb	*iru*	"there is (someone)"	*agaru*	"go up"
	naru	"become"	*aru*	"there is (something)"
	dekiru	"can do"	*wakaru*	"understand"
	suru	"do"	*yaru*	"do"
	kangaeru	"think"	*motsu*	"have"
	tsukareru	"get tired"	*yomu*	"read"
	shinu	"die"	*yaru*	"do"
Adv	*nandaka*	"somehow"	*itsumo*	"always"
	sukoshi	"a little"	*maa*	"relatively"

Table 6: Some selected verbs and adverbs.

4.3 Lemmas and Selected Lemmas

The accuracy score for each classifier with the lemma feature set is shown in Table 5. The NB classifier resulted in the highest accuracy of 95.5% with verbs and adverbs as features (2,627 features). After feature selection within the set of verbs and adverbs (2,007 features), the accuracy stayed the same. With different set of features, the SVM achieved 90.9% accuracy after further feature selection. For document-level classification, the highest accuracy was 69.0% with selected lemmas of four POS categories (2,579 features). Even though the accuracy improved from the baseline, it did worse than the character *n*-grams. Some of the selected lemmas from the best resulting author-level classification are shown in Table 6. Words that appear more frequently in one class are listed under that class.

5 Discussion

5.1 Classification and features

In all the experiments, author-level classification is better than document-level classification. This may be because each document contains around 200 words in document-level classification, and many features may not appear in one document, leaving feature vectors sparse.

Lemmas of verb and adverb categories give the best accuracy for the author-level classification. This suggests that frequency of words, regardless of their inflection or the surrounding context, is most useful when provided with sufficient amount of text. Although some of the selected lemmas are related to symptoms of depression (fatigue, suicidal thoughts), lemmas appearing frequently in depressed authors' documents are not necessarily related to emotions or mood (e.g. somehow, always). This suggests that there are subtle differences in choice of words by depressed authors which we may not immediately associate with depression.

Morphological and syntactic information such as inflection and word order, may be useful, but they do not provide accuracies that are as good as lemmas in the author-level classification. However, the experiment with character trigrams results into having the best accuracies for the document-level classification. This is likely because within a limited amount of text, character *n*-grams appear more often than lemmas. Romanizing characters do not improve the performance. Representing Japanese language with Romaji suppresses homonyms and *kanji* that may otherwise

	Author		Document	
	Before	After	Before	After
CharUni	81.8 (NB)	68.2 (NB)	75.3 (LR)	67.2 (LR)
TokenUni	72.7 (NB)	86.4 (NB)	71.5 (LR)	63.2 (NB)
LemmaUni	72.7 (NB)	81.8 (NB)	71.5 (LR)	66.8 (NB)

Table 7: Accuracies before and after topic modeling

be informative (see Section 3.4.1).

5.2 Topic bias

We now take a closer look at the effect of topic bias, i.e., we compare the results when the entries on depression have been removed (see Section 3.2 for details) to the condition when these entries are kept in the training and the test set. The latter condition corresponds to the settings that have been used by Matsumoto et al. (2012). The classification on the document-level with the full data set does worse than the classification with the cleaned data (see Table 7). This is expected because topic bias is factored out after topic-modeling. However, on the author-level, it is a more complex picture: for word-based units (token and lemma), accuracy actually goes up once topic bias is removed. As a blog entry tends to focus on one topic, and depressed authors' documents contain more words about depression, the document-level classification seems to be affected by the topic. We will investigate why the author-level classification improves with the cleaned data in future work.

6 Conclusion and Future work

This study showed that selected lemmas can predict whether authors of written texts are depressed or not with an accuracy of 95.5%. This is higher than Matsumoto et al. (2012) though it is difficult to compare because of different data sets. The better performance of author-level classification suggests that documents should contain enough text to be classified correctly. The next step will involve finding out how much text per document is necessary to achieve such high accuracy.

As the current study only tested default parameters for SVM and LR in classification and feature selection, different parameter settings will be tested in the future work.

As the study is small-scale, it is necessary to examine how the results extend to larger data. Moreover, expanding the scope of study to other mental conditions may reveal the nature of language use in relation to mental health. Further investigation of selected lemmas in connection with clinical studies may provide us insights on why these words work well as features.

Finally, the methods of the current study can easily be adapted to other languages with different character systems if a language can be tokenized and POS-tagged. It would be worth exploring how depression can be detected from texts in different languages and performing a cross-linguistic comparison of characteristics found in depressed authors' writings.

References

Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. 2012. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, pages 24–32.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, pages 149–156.

Junko Kitanaka. 2012. *Depression in Japan: Psychiatric cures for a society in distress*. Princeton University Press.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*. pages 63–69.

Kazuyuki Matsumoto, Nobuhiro Yoshioka, Kenji Kita, and Fuji Ren. 2012. Utsu key phrase to kanjo hyogen hendo ni motozuku blog kara no utsu kensyutsu syuho. *The Association for Natural Language Processing 18th Conference Proceedings* pages 1126–1129.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

John Noecker, Michael Ryan, and Patrick Juola. 2013. Psychological profiling through textual analysis. *Literary and Linguistic Computing* 28(3):382–387.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Francisco Rangel, Paolo Rosso, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, Walter Daeleman, et al. 2014. Overview of the 2nd author profiling task at pan 2014. In *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, volume 1180, pages 898–927.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pages 78–86.

Hiroki Sato. 2011. Japanese dictionary of appraisal - attitude- (jappraisal dictionary). *JAppraisal Dictionary ver1* .

World Health Organization. 2017. Depression. http://www.who.int/mediacentre/factsheets/fs369/en/.