Document Classification Using Non-content Features

馬場, 隆寛

Document Classification Using Non-content Features

Takahiro Baba

January 2021

Abstract

With the spread of computers and the Internet, a large amount of documents have become accessible as digital data, and advanced analysis technologies for those documents are required. It is expected that new knowledge will be discovered by performing statistical analysis on a large number of documents in addition to browsing each document. Information other than the content of each document can be extracted by a new knowledge extraction method obtained by statistical analysis. This can assist the understanding of the author's feelings and intentions and the estimation of the value of the document, and can be expected to be applied in various fields.

In this research, we work on the extraction of non-thematic information by classifying documents. In document classification, an efficient method by machine learning with appearing words has been established, and words that greatly contribute to classification can be regarded as characteristic of the set category. In general, document classification by machine learning requires a large amount of document data (training data) that has information (labels) of the categories to be classified. However, it is difficult to collect a large amount of documents labeled with non-content information compared to the content matter. Therefore, in this study, we create training data with labels related to the target non-thematic information by removing the features related to the content from the documents collected by the content.

In this paper, we propose document classification using non-content features. The content characteristics depend on the content. Therefore, it is necessary to

classify using features that do not depend on the content of the document. In addition to content features, documents should show non-content features such as writing style. By using this non-content feature, we can obtain statistics for learning the properties of interest that do not depend on the content of the document. This makes it possible to classify training data by non-content even if the training data assumes a label for a document set of a specific content.

In this paper, we conducted experiments on specific problems and verified the effectiveness of the proposed method. We examined three problems in which it was difficult to obtain a large amount of labeled training data.

First, we estimated the number of citations of research papers and verified that the attributes of documents can be classified. The content of the research treatise is the research content, and we attempted to extract academic impact, which is non-content information, based on the number of citations. A large number of abstracts of research papers labeled with the number of citations can be obtained. However, it was shown that when the classification is performed using the content characteristics, the tendency according to the research field is found rather than the paper itself. As a result of applying the proposed method to a large number of abstracts, we were able to estimate the number of citations with high accuracy using words and phrases unrelated to the research field.

Second, we estimated the author's native language from a document written in a specific language and verified that the author's attributes could be classified rather than the document itself. Although there are few documents in which the author's native language is clearly stated, the proposed method was applied by taking advantage of the fact that many research papers written in multiple

languages are accompanied by English abstracts. The content of the abstract of the research paper is the research content, and the author's native language is assumed as non-content information. By applying the proposed method, highly accurate native language detection was realized with features that do not depend on the research content.

Third, we estimated the mental state of the posters of comments on a social network system, and verified that even documents with a small number of characters can be classified. When each patient with a mental illness is identified, the amount of comments that can be obtained is limited. On the other hand, a large number of comments were obtained by using the comments from the group counseling site. The content of this comment is the content of counseling, and the non-content information is the mental state of the poster. By the proposed method, we have realized highly accurate estimation based on content-independent features from the classification of counseling site comments as positive examples and general SNS comments as negative examples.

From the above, it was found that the proposed method can be applied to a wide range of applied problems.

Contents

Chapter 1

Introduction

1.1 Background

The spread of computers and the Internet has enabled us to access a large amount of documents. Some tendencies in multiple documents indicate different kinds of knowledge from the content matter written directly in each document. For example, in a trial to analyze automatically more than 200,000 research papers related to COVID-19 [2], statistical information in the multiple research papers is expected to lead a useful discovery, such as unknown gene mutations related to a specific pathological condition. The information extracted from multiple documents by statistical analyses occurs in each document independently of the content of that document. Therefore, this "non-content" information can be used to understand implicit aspects of a document in various applications, such as understanding author's intention and emotion.

1.2 Motivation

In this paper, we extract non-content information from each document using results of document classifications. An example of non-content information of a document is the author's mental state. For document classification, efficient methods based on machine learning with word occurrence has been established [43], and words that greatly contribute to the classification are regarded as characteristics of the given category.

In general, document classification with machine learning requires a large amount of sample data (training data) that has information (labels) of the categories to be classified. However, it is more difficult to collect a large amount of documents labeled with non-content information than documents labeled with the content. For example, collecting documents that explain persons' mental states is possible by searching for a content, but collecting documents with information of the author's mental state requires additional processes. In this study, training data with labels of non-content information is generated from documents with labels of the content by hiding information related to the content in that documents. For example, we use comments on a group counseling site for mental illness as documents written by authors with mental illness. Although the content of the documents are related to counseling, hiding information related to the content enable us to compare those with other general documents written by general authors.

In document classification in this paper, we delete information related to the content of each document by generalizing words about content. The content of a

document can be recognized in a degree by analysing occurrences of words about content in that document. Therefore, a straightforward document classification based on word occurrence is highly affected by the contents of that documents. In this paper, we replace words related to content in documents to general tags and treat the occurrences of the tags and other words as features for classification. Concretely, the words related to content are technical terms for the target area, or the general content words consist of nouns, verbs, adjectives, and adverbs. Thus, a content-independent document classification is obtained using these "non-content" features.

We apply document classifications based on the previous idea to practical tasks. In each task, gathering training data is difficult, and this problem is solved using the content and non-content information of documents.

- **Citation count prediction:** We predict the number of citations of research papers using their abstracts. The content of each document is the research topic. We try to estimate an academic impact of a paper, which will not depend on the research topic, as non-content information.

- **Native language identification:** We estimate the author's native language from a document written in a specific language to verify that our method can extract the author's attributes in addition to that of a document itself. The author's native language is estimated as non-content information from documents whose contents are research topics.

- **Mental health prediction:** We estimate the mental state of persons from comments on social network systems (SNSs) to verify that our method is

applicable to short documents. The content and non-content information of the documents are the content of counseling and the author's mental state, respectively.

In the rest of this chapter, we describe each of the three tasks.

1.3 Citation count prediction

Researchers are required to efficiently determine previous literature that is related to their research and has scientific impact from among a large number of publications. The number of research papers available on-line is rapidly increasing. Ideally, researchers should survey all possible publications for their research, but it is difficult to read the main text of all paper carefully. Therefore, researchers are expected to choose papers relevant to their research from a huge amount of data, and papers with high impact should be chosen before papers with low impact.

We address the problem of predicting the scientific impact of research papers. The measure of this impact is the *citation count* of each paper, that is, the number of citations from other papers to that paper. Predicting citation count enables us to screen papers to determine papers that potentially have high impact. Citation count is a reasonable feature for formalizing scientific impact. The impact factor [29], which is often referred as a quality measure of journal titles, is defined for a journal using the citation counts of the articles published in that journal. The h-index [34], which is a measure of the contribution of a researcher to the society concerned, is also based on the citation counts of papers written by that researcher. Citation count will increase as time elapses, and hence is not ap-

propriate for measuring the impact of brand-new papers. Therefore, we need to predict the potential citation count of a paper using features that can be directly extracted from it.

A variety of features can be used for solving the problem of citation count prediction [51]. Previous literature on this problem claims that the textual data of a paper do not deeply affect the prediction compared with data about authors and venue of publication (see Section 3.1). However, textual data, especially abstracts, are worth analysing in detail for the following reason. Textual data should be directly related to the contents of a paper. In particular, abstracts are usually available as the metadata of respective papers. The other metadata, including authors, authors' institutions, and journal (or conference) titles can be features of multiple papers, while an abstract corresponds to the paper concerned. In the previous literature, we identified two common limitations in using only abstracts for prediction:

(1) The prediction accuracy is not high and detailed analyses are not conducted;

(2) The prediction is explained in terms of trivial findings regarding the trends in research topics.

For the first limitation, we address a binary classification of abstracts into high and low citation counts abstracts as the target task, instead of a regression which is addressed in most previous literature. We tackle an easier task to analyze the effects of abstracts on citation counts in detail. As for the second limitation, we investigate the effect of the technical terms that appear in abstracts on the prediction.

We investigated several types of classification of research paper abstracts to predict citation count. Our aim is to clarify the effects of (1) the abstracts of papers and (2) the technical and non-technical terms used in abstracts rather than to achieve high accuracy. We applied a standard classification method based on the bag-of-words model [43] to a set of abstracts of papers with high and low citation counts and investigated the accuracy and distinctive phrases. We obtained abstracts with citation counts from a database of research papers. Then, we defined the set of high and low citation counts by selecting top θ% and bottom θ% papers in order of citation count in the obtained abstracts. We also applied the same classification method to another set of modified abstracts in which the technical terms were replaced with a meaningless symbol. Additionally, we conducted a classification using only the technical terms that appear in the abstracts for comparison.

The results of our experiments indicate that the scientific impact of a research paper can be roughly predicted using only its abstract, and the effective features in the prediction are related to the trend of research topics. Papers with high and low citation counts can be accurately classified using their abstracts. However, the same classification of the modified abstracts with hidden technical terms had low accuracy. The accuracy of the classification using only technical terms was better than that of the modified abstracts. In other words, it was shown that if the classification is simply performed using the content characteristics, the tendency according to the research field is found rather than the paper itself.

1.4 Native language identification

Profiling the authors of documents is effective in advanced analyses of those documents. The author's attributes and any extra knowledge related to them can be used to understand the implicit meanings of documents; for example, the author's intentions and emotions. Understanding these implicit meanings is expected to yield novel methods for various document analysis tasks; for example, machine translation and dialogue generation.

This study aims to find differences in English writing styles between authors whose native languages are not English. Writing styles depending on a language should be affected by the history and culture of the society concerned; therefore, it can include information that we cannot directly extract from the documents. Thus, we conducted document classification and investigated the distinctive phrases of the authors' native languages. The classification is regarded as a task of native language identification (NLI), that is, identifying the native language (L1) of the author of a document written in the second language (L2). We assume that we could use only the textual data of documents as features for NLI, while other kinds of features can be used for NLI; for example, eye movements of subjects in reading documents written in L2 [16].

Our approach to NLI is based on machine learning. A difficulty in machine learning-based approaches to NLI is that we need a sufficient amount of supervised data. To remove this difficulty, we used research paper abstracts, which are available on a massive scale from the Internet, as training data for machine learning. We gathered English abstracts of papers written in one of five languages

other than English from PubMed [9]. On the assumption that the main text of a paper is written in the author's L1, the English abstract of that paper is regarded as a document written in the author's L2.

In this paper, we classified English abstracts of research papers written in Chinese, French, German, Japanese, or Spanish. We also conducted the classification with the abstracts modified by generalizing content words, to remove the effect of topics specialized in the data. Additionally, we considered to some distinctive phrases of each language found in the classification from the viewpoint of practices related to the language.

As a result of the experiments, we found some tendencies of writing styles. The classification accuracy was high for the normal and modified data. We have realized highly accurate native language detection with features that do not depend on the research content. Additionally, distinctive phrases used in the classification were related to some typical practices depending on the language concerned. Those results are expected to be used for estimating the authors' intentions in documents.

1.5 Mental health prediction

Mental illnesses have become a serious public problem. According to the statistics released by the World Health Organization [3], more than 350 million people suffer from one of the illnesses, depression. Many people are not aware of their psychological disorders, and therefore, they often run into serious conditions. As a solution for this situation, we need technologies for detecting symptoms of the

illnesses automatically and in early stages from changes in everyday behavior.

The problem we address is to detect persons who have mental health problems using their comments posted to SNSs. Becker et al. [14] summarized applications of information technologies into mental health care. Everyday life of each person is observed using mobile devices and networks, and a large amount of data can be analyzed using advanced technologies including machine learning. Especially, a number of studies use data posted to SNSs including Twitter [11] and Facebook [5] for detecting mental health problems [33]. Symptoms associated with psychosis are observable on SNSs, and automated methods can detect depression and other psychoses. Our approach is based on machine learning with a large amount of text data, especially, with statistics of word occurrences. A difficulty of this approach is in preparing a sufficient amount of supervised data, that is, comments of persons who have been diagnosed as having mental illnesses.

We used comments posted to a Web community for persons with mental health problems. Cocooru [1] is a Japanese Web site that provides users with counseling for mental health from clinical psychologists, and has a message board system for communication between users. Using the comments of the Web site enables us to obtain a large amount of positive samples in exchange for a degree of validity. Although each user has not been identified as a person with a mental illness, expressing own feeling in the community indicates that the person has a kind of problem in his or her mental health. It is confirmed by the administrator of the system that there is no comments posted by the counselors.

In this paper, we classified Japanese comments obtained from two SNSs to find the differences in their writing styles. We distinguished comments of Cocooru from

comments of Twitter. We applied a simple machine learning method to document features based on word occurrences to capture the characteristics of the SNS as phrases. We also conducted the classification with the comments modified by generalizing content words to remove the effect of topics specialized in the SNS.

As a result of the experiments, we found some characteristics in the writing styles of persons with mental health problems. The classification accuracy was high for the normal data. Although the accuracy was decreased by the generalization of content words, it was remarkably higher than a random prediction. Additionally, some of the distinctive phrases used in the classification were related to the results of earlier work with English comments. Those results can be used for detecting persons with mental illnesses from their comments in everyday lives. From the classification with the comments of the counseling site as the positive example and the comments of the general SNS as the negative example, highly accurate estimation was realized by the features independent of the subject.

Chapter 2

Preliminaries

In this chapter, we introduce the definitions of the terms and the tools used in this paper.

In this paper, document classification is performed using machine learning, especially supervised learning. Generally, supervised learning finds the function from the input vector to the output vector using given pairs (*training data*) of an input vector and a tag to construct the target vector (a *label*). The input vector is defined from each document using the occurrences of words. The function is obtained as a hyperplane that divides input vectors with different output labels using support vector machine (SVM) [17].

An n-gram of a sequence is a sequence of contiguous n elements of the sequence. For example, the 2-grams of the sentence "I am your father" are "I am", "am your", and "your father". The $(1, n)$-grams of a sequence is the union of the set of the i-grams of that sequence for $1 \leq i \leq n$. For example, the $(1, 2)$-grams of the previous sentence is the set of "I", "am", "your", "father", "I am", "am your",

and "your father".

The accuracy measures for our experiments are defined as follows. The *accuracy* of a classification is the ratio of the number of the correct predictions to the number of the total predictions examined in a validation. The *precision* and the *recall* for each class are the ratio of the number of the correct predictions to a class to the number of the predictions to the class, and the ratio of the number of the correct predictions to a class to the actual number of samples of the class, respectively. The *F-score* for a class is the harmonic mean of the precision and recall for the class.

The *coefficient of determination* for predicted values y_i and the corresponding true values $\hat{y}_i$ for $1 \leq i \leq n$ is defined to be

$$1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

In our Python code, we used the functions `SVC` and `TfidfVectorizer` defined in the library `scikit-learn` [32] for the classifier and transformer from a comment to a numerical vector. We used the library `NLTK` [7] to detect the POS of each word. We also used the analyzer `JUMAN++` [44] of Japanese documents to separate each comment to a list of words and to detect the POS of each word.

Chapter 3

Citation Count Prediction

Researchers are expected to find previous literature that is related to their research and potentially has a scientific impact from among a large number of publications. This section addresses the problem of predicting the citation count of each research paper, that is, the number of citations from other papers to that paper. Previous literature related to the problem claims that the textual data of papers do not deeply affect the prediction compared with data about the authors and venues of publication. In contrast, we detect the citation counts of papers using only the paper abstracts. Additionally, we investigate the effect of technical terms used in the abstracts on the detection. We classify abstracts of papers with high and low citation counts and apply the classification to the abstracts modified by hiding the technical terms used in them.

3.1 Related work

The novelty of our work is that we clarified the following:

- the effect of research paper abstracts on citation count prediction;
- the effect of the technical terms that appear in an abstract on the prediction.

Existing studies have concluded that textual data of papers (including abstracts) are not effective for citation count prediction when compared with data related to authors and publication venues. Additionally, the textual data are treated as topics extracted from raw text in existing work. Therefore, non-topic information included in the textual data, such as the writing style, has not been considered as a feature for prediction.

Yan et al. [51] treated textual data as topics and concluded that their effect on prediction is small. They formalized the problem of citation count prediction. They applied four kinds of regression to three types of features of each paper (content, author, and venue). The content feature category includes the topics, which were obtained using latent Dirichlet allocation [18] from the textual data of the papers. The categories of author and venue include attributes related to the authors of each paper and the conference or journal at which each paper was published, respectively. The results of their experiments indicate that the features in the content category have little effect on prediction accuracy.

Chen and Zhang [23] also treated textual data as topics, but their standalone effect on prediction is not clear. They predicted citation count using regression for features related to the contents and authors of papers. They conclude that

the features included in the content category are more effective than the author category. However, the effect of the textual data on the prediction was not clear, because both categories include topic information obtained from textual data using latent Dirichlet allocation, and the content category includes the information of past citation counts in addition to the topic information.

Li et al. [39] did not use textual data for prediction. They used the change in citation count over time for the prediction. Their method estimates this change using some paper features, but these features do not include textual data.

Dong et al. [27] treated textual data as topics for another type of prediction and concluded that their effect is small. They predicted h-index instead of citation count and investigated the effects of several features. They concluded that the relationship of the main author to the research topic and the venue are more effective for prediction than the trend in research topics and the co-authors. Textual data are used as topics for the prediction, and their standalone effect is small.

Yogatama et al. [52] used textual data as the main feature for prediction, but they treated them as topics. They used the change in topics over time for citation count prediction, which can be treated as research topics trends.

3.2 Methods

3.2.1 Data

We formalized the problem of an approximate citation count prediction as binary classification instead of regression. The positive and negative data were gener-

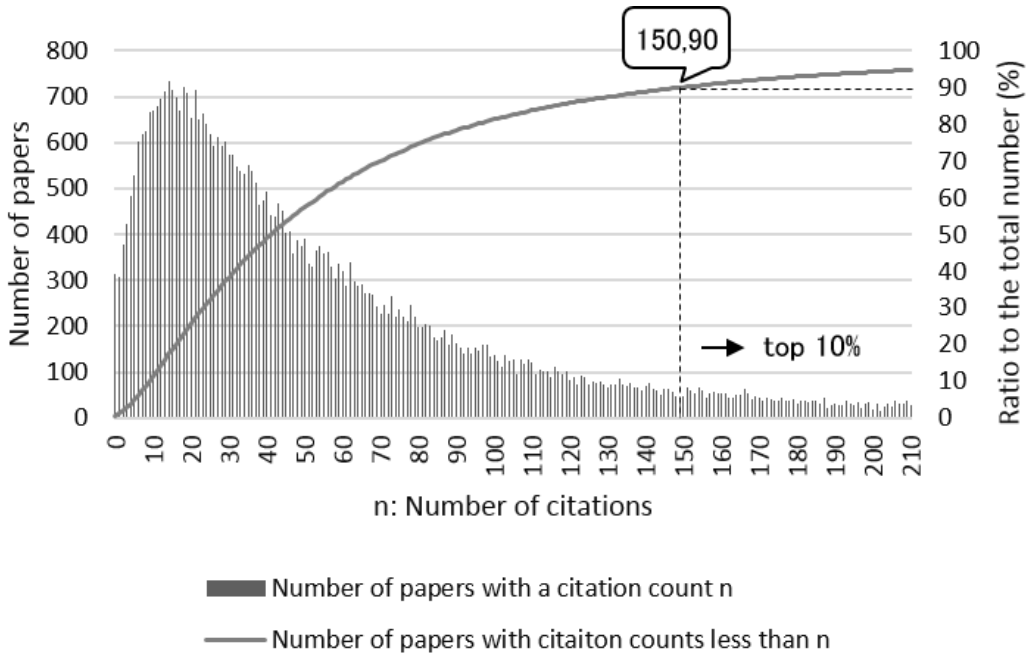

Figure 3.1: Distribution of the numbers of papers with a citation count.

ated from a set of papers using a threshold for the number of citation counts. Figure 3.1 shows the distribution of papers by citation count and the ratio of the accumulated number of papers to the number of the total papers we used in our experiments. For example, using a threshold of 10%, positive samples are defined as the abstracts of papers whose citation counts are more than 150. For the data sets described below, we conducted three classifications using a standard method using the occurrences of the total words, technical terms, or non-technical terms.

We used the abstracts and citation counts of papers published in the *Proceedings of the National Academy of Sciences* (PNAS) [8] for our experiments. PNAS is appropriate for our experiments for the reason that we can obtain a sufficient number of papers published in a single journal title and have citation counts. The

scope of PNAS includes any research area in general science. We obtained the metadata, which includes the abstract and citation count, from Europe PubMed Central (Europe PMC) [4].

We conducted a preliminary experiment and selected the data of papers published in PNAS from 1981 to 2003 (and available from Europe PMC) according to the result of the experiment as follows. Figure 3.2 shows the annual numbers of papers published from 1915 to 2017 and the average citation counts, where the citation counts are the values as of June 2017. As shown in the figure, the average citation count rapidly decreases after 2004, which indicates that the citation counts of papers published after 2004 could be potentially larger in the future. Therefore, we used the data of papers published before 2004. Additionally, there are two peaks in the graph of average citation counts in the 1970s, which can be attributed to some exceptional factors. Therefore, we used the data of papers published after 1980. Additionally, we used abstracts between 100 and 400 words in length. The number of abstracts in this *normal data set* is 49,171.

We also generated a *modified data set* of abstracts to clarify the effect of the occurrences of technical terms on the citation counts. As the corpus for defining technical terms, we used Medical Subject Headings (MeSH) [6], which is a medical thesaurus published by the National Library of Medicine. The indexes of MeSH includes a subject heading "Descriptor", a subheading "Qualifier", and a supplementary concept record "Concept". We used all the Descriptors, Qualifiers, and Concepts defined in the latest version as of February 2018 of MeSH as the technical terms for our experiment. The scope of MeSH is considered to be restricted to the life sciences while that of PNAS includes general science. The effect of

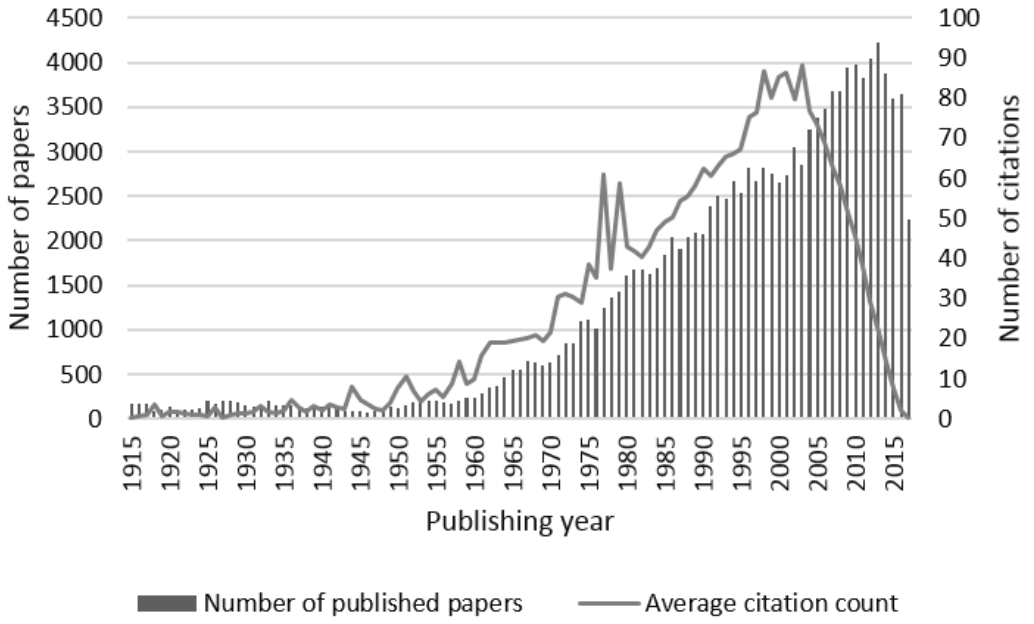

Figure 3.2: Annual number papers published in PNAS and available from Europe PMC as of June 2017, and their average citation counts.

this difference on the experimental results is examined in Section 3.4. In addition to the terms, we regarded phrases obtained by dividing the terms using commas as technical terms. We also used combinations of phrases divided by brackets in the terms as technical terms. For a technical term "A (B) C" for phrases A, B, and C, we used the phrases "A B C", "A C", "B C", A, B, and C as technical terms. The number of obtained technical terms is 187,573. Then, all technical terms appearing in the abstracts were replaced with a symbol "X". For example, the abstract of the paper [22]:

The genomic RNA of human rhinovirus type 14 was cloned in Escherichia

coli and the complete nucleotide sequence was determined. The RNA genome is 7212 nucleotides long. A single large open reading frame of 6536 nucleotides was identified, which starts at nucleotide 678 and ends 47 nucleotides from the 3' end of the RNA genome. Comparisons of the specified proteins with those of other picornaviruses showed a striking homology (44-65%) between rhinovirus and poliovirus. The rhinovirus genomic RNA is rich in adenosine (32.1%) and strongly favors an adenosine or uridine in the third position of codons. The predicted map locations of all the rhinovirus structural and non-structural proteins and their proposed proteolytic cleavage sites are described.

was modified by replacing technical terms with "X" as follows:

The X X of X X X 14 was cloned in X and the X X was determined. The X X is 7212 X X. A X X X of 6536 X was identified, which X at X 678 and ends 47 X from the 3' X of the X X. X of the specified X with those of other X showed a striking X (44-65%) between X and X. The X X X is rich in X (32.1%) and strongly favors an X or X in the X X of X. The predicted X X of X the X X and X and their proposed proteolytic X X are described.

All the abstracts in the normal data set included at least one technical term; hence, the size of the modified data set is equal to that of the normal data set.

3.2.2 Experiments

We predicted citation counts using the data sets. We conducted 5-fold cross-validation. For training and test data in each validation, we selected the top and bottom θ% papers in the order of citation counts as positive and negative samples, respectively, after normalizing the citation count of each paper by dividing the number by the average citation count of its publishing year. The threshold θ was set to be 2^i for $0 \leq i \leq 5$ and 50. Therefore, the size of the experimental data for each classification is 2θ% of the total data, because each data set is defined using the top θ% and the bottom θ% of the original data.

We conducted five classifications with the normal and modified data sets. We applied an SVM to the multisets of the words appearing in the abstracts in the two data sets. We also applied the SVM to the word-level $(1, 3)$-grams of the abstracts. We ignored all the single-character words except for the "X" used for technical terms, and did not use phrases that appeared in more than 50% of the training data for classification. Then, we predicted a positive or negative class using vectors obtained from the multisets. Additionally, we applied the SVM to the sets of technical terms appearing in the abstracts of the normal data set. Finally, we obtained the five data sets:

- *Normal*: the set of the multiset of the words appearing in each abstract in the normal data set;
- *Normal 1-3*: the set of the word-level $(1, 3)$-grams of each abstract in the normal data set;

- *Modified*: the set of the multiset of the words appearing in each abstract in the modified data set;
- *Modified 1-3*: the set of the word-level $(1, 3)$-grams of each abstract in the modified data set;
- *Technical term*: the set of the set of technical terms appearing in each abstract in the normal data set.

For comparison, we conducted a linear regression using the same data sets and features used in the binary classification. We investigated the coefficient of determination as a measure of accuracy for five regressions which correspond to the five classifications.

3.3 Results

Figure 3.3 shows the accuracy of the five classifications against threshold θ for generating positive and negative samples. Figure 3.4 shows the dimensionality of the vectors, that is, the number of the phrases, used in the classification for the cases. Tables 3.1, 3.2, and 3.3 show the confusion matrices of the classifications.

Table 3.4 shows the coefficient of determination for the five regressions that correspond to the five classifications. By the definition, a minus value of the coefficient means that the prediction is almost meaningless.

Tables 3.5, 3.6, and 3.7 show distinctive phrases of the abstracts of papers with high and low citation counts. The listed phrases correspond to the top and bottom five elements of the separating hyperplane used in the classifier in the order of the

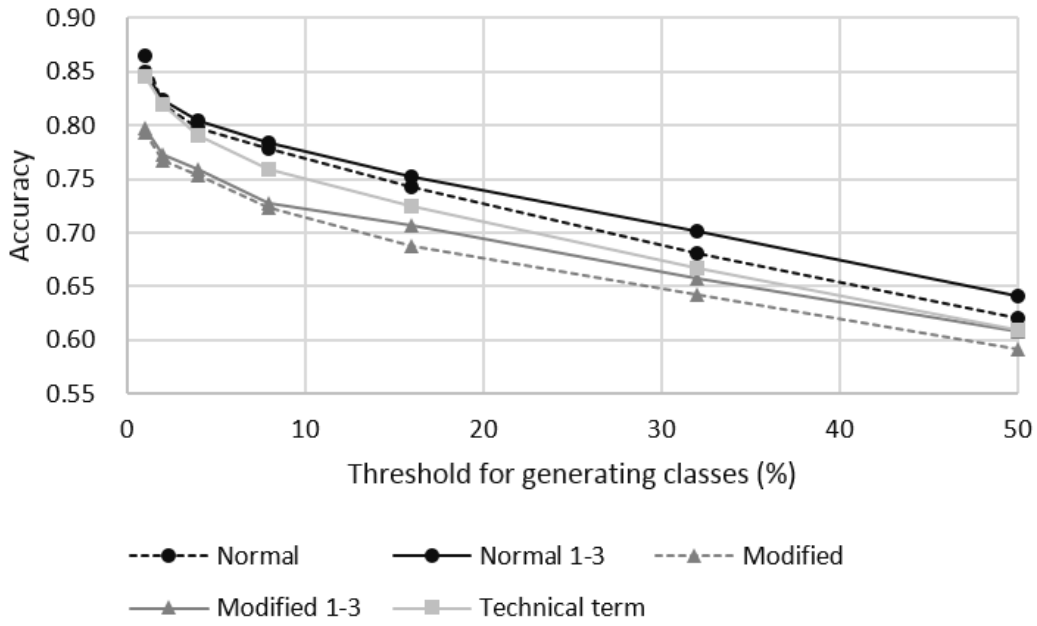

Figure 3.3: Classification accuracy of research papers with high and low citation counts.

coefficients, for the five classifications. Therefore, these phrases should be roughly distinctive of the positive or negative data.

3.4 Discussion

3.4.1 Main findings

We found that papers with high and low citation counts could be classified using only their abstracts. As shown in Figure 3.3, the accuracy of the five classifications using abstracts were better than the expected value 50% of that of random predictions. Therefore, by using this metrics, we can analyze the effects of con-

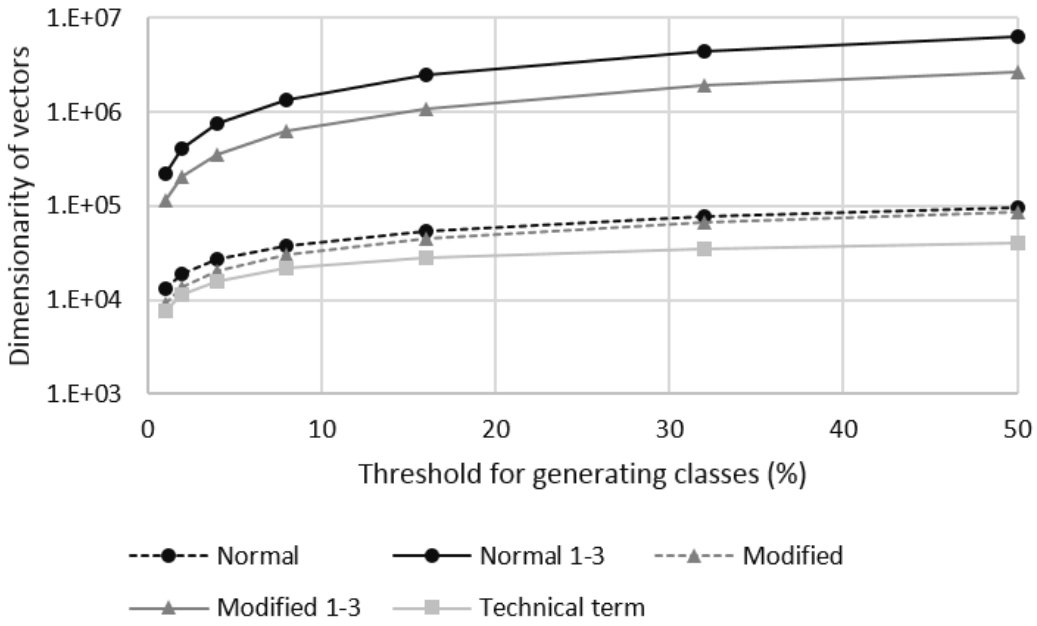

Figure 3.4: Dimensionality of the vectors used for classifications.

tiguous occurrences of words or technical terms on the classification, while the linear regression with the abstracts predicted meaningless values as shown in Table 3.4. Additionally, the accuracy increases in opposition to θ. In case where $\theta = 50$, which corresponds to the naive binary classification with the total data, the prediction accuracy is low for the five data sets. By using only restricted papers we can clarify the difference between papers with high and low citation counts.

The effect of contiguous occurrences of words on the classification is small. As shown in Figure 3.3, the classification accuracy generated by the two classifications using the 1-, 2-, and 3-grams were almost the same as that by the corresponding classifications using the 1-grams, although the dimensionalities of the vector spaces

Table 3.1: Confusion matrices generated by classifications of the normal data set.

		1-gram			(1-3)-gram		
		Positive	Negative	Precision	Positive	Negative	Precision
$\theta = 1$	Positive	439	81	0.84	456	113	0.8
	Negative	51	409	0.89	34	377	0.92
	Recall	0.9	0.83		0.93	0.77	
$\theta = 2$	Positive	839	212	0.8	890	255	0.78
	Negative	141	768	0.84	90	725	0.89
	Recall	0.86	0.78		0.91	0.74	
$\theta = 4$	Positive	1596	427	0.79	1697	498	0.77
	Negative	369	1538	0.81	268	1467	0.85
	Recall	0.81	0.78		0.86	0.75	
$\theta = 8$	Positive	3107	917	0.77	3293	1065	0.76
	Negative	823	3013	0.79	637	2865	0.82
	Recall	0.79	0.77		0.84	0.73	
$\theta = 16$	Positive	5907	2086	0.74	6234	2276	0.73
	Negative	1958	5779	0.75	1631	5589	0.77
	Recall	0.75	0.73		0.79	0.71	
$\theta = 32$	Positive	10861	5158	0.68	11645	5309	0.69
	Negative	4869	10572	0.68	4085	10421	0.72
	Recall	0.69	0.67		0.74	0.66	
$\theta = 50$	Positive	15482	9532	0.62	16873	9921	0.63
	Negative	9103	15053	0.62	7712	14664	0.66
	Recall	0.63	0.61		0.69	0.6	

increased by more than ten times by using (1,3)-grams as shown in Figure 3.4. Additionally, Tables 3.5 and 3.6 show that using (1,3)-grams does not find any phrase which represents a writing style or an idiom. Therefore, we suppose that the effect of the information obtained from textual data including the writing style and idioms is little on predicting high impact papers.

The effect of technical terms on the classification is large. The decrease in classification accuracy caused by using only technical terms from the normal one was smaller than that caused by hiding technical terms as shown in Figure 3.3. As

Table 3.2: Confusion matrices generated by classifications of the modified data set.

		1-gram			(1-3)-gram		
		Positive	Negative	Precision	Positive	Negative	Precision
$\theta = 1$	Positive	426	138	0.76	459	168	0.73
	Negative	64	352	0.85	31	322	0.91
	Recall	0.87	0.72		0.94	0.66	
$\theta = 2$	Positive	806	281	0.74	891	357	0.71
	Negative	174	699	0.8	89	623	0.88
	Recall	0.82	0.71		0.91	0.64	
$\theta = 4$	Positive	1533	535	0.74	1662	644	0.72
	Negative	432	1430	0.77	303	1321	0.81
	Recall	0.78	0.73		0.85	0.67	
$\theta = 8$	Positive	2915	1160	0.72	3117	1325	0.7
	Negative	1015	2770	0.73	813	2605	0.76
	Recall	0.74	0.7		0.79	0.66	
$\theta = 16$	Positive	5564	2602	0.68	6002	2757	0.69
	Negative	2301	5263	0.7	1863	5108	0.73
	Recall	0.71	0.67		0.76	0.65	
$\theta = 32$	Positive	10251	5792	0.64	11050	6100	0.64
	Negative	5479	9938	0.64	4680	9630	0.67
	Recall	0.65	0.63		0.7	0.61	
$\theta = 50$	Positive	14857	10372	0.59	15893	10610	0.6
	Negative	9728	14213	0.59	8692	13975	0.62
	Recall	0.6	0.58		0.65	0.57	

mentioned in Section 3.2, some technical terms used in the papers of PNAS are not included in the set of the technical terms defined using MeSH. The effect of technical terms was shown even using the insufficient set. As shown in Figure 3.4, the dimensionality of the vector space was small even compared with that of the classifications using the 1-grams. Therefore, we can estimate that most effective factors used in the classification were included in the set of technical terms.

Table 3.3: Confusion matrices generated by classifications of the normal data set using technical terms.

		Positive	Negative	Precision
$\theta = 1$	Positive	435	96	0.82
	Negative	55	394	0.88
	Recall	0.89	0.8	
$\theta = 2$	Positive	824	198	0.81
	Negative	156	782	0.83
	Recall	0.84	0.8	
$\theta = 4$	Positive	1570	430	0.78
	Negative	395	1535	0.8
	Recall	0.8	0.78	
$\theta = 8$	Positive	2987	950	0.76
	Negative	943	2980	0.76
	Recall	0.76	0.76	
$\theta = 16$	Positive	5741	2196	0.72
	Negative	2124	5669	0.73
	Recall	0.73	0.72	
$\theta = 32$	Positive	10494	5254	0.67
	Negative	5236	10476	0.67
	Recall	0.67	0.67	
$\theta = 50$	Positive	15009	9599	0.61
	Negative	9576	14986	0.61
	Recall	0.61	0.61	

3.4.2 Key findings

We can conclude that the effect of technical terms on the prediction is large, also from the distinctive phrases used in the classification. As shown in Tables 3.5 and 3.7, most distinctive phrases of positive samples in the classification with the normal data set are included in the technical terms. Additionally, most distinctive phrases of positive data in the classification with the modified data set using (1-

Table 3.4: Coefficients of determination in five regressions.

	Coefficient of determination
Normal	−3.18
Normal 1-3	0.03
Modified	−5.80
Modified 1-3	−0.04
Technical term	−14.51

3)-grams contain the symbol which means technical terms as shown in Table 3.6. Therefore, the effect of technical terms on the prediction is considered to be large. In the case of the modified data set, using the generalized technical term finds how technical terms are used in abstracts instead of what the terms are.

3.4.3 Future work

One of our future work is extending the analysis of distinctive phrases. We listed several phrases with large or small weights in the classifier, but the vocabulary size, that is, the dimensionality of the vector space for each classification was extremely large. Therefore, we need a method to investigate a large number of phrases efficiently.

Another direction is applying our results to actual problems. The results of the experiments is expected to be used for an automatic proofreading of research papers; the scores put on phrases can suggest a better phrase expected to lead high impact.

Table 3.5: Phrases with large and small coefficients in classification of the normal data set.

	θ	Phrases
Large	1	human; dna; brain; gene; neurons
	2	human; neurons; blood; genes; gene
	4	major; neurons; cortex; blood; genome
	8	cortex; bacterial; human; potent; fluorescent
	16	epithelial; bacterial; cortex; actions; endothelial
	32	consensus; cortex; mariner; variety; nitrocellulose
	50	confirming; latently; leaflet; women; overproducing
Large, (1-3)-gram	1	human; dna; gene; cells; expression
	2	gene; human; genes; dna; neurons
	4	gene; human; protein; brain; neurons
	8	human; gene; brain; cortex; sequence
	16	human; brain; bacterial; mice; cortex
	32	brain; human; cortex; common; oxygen
	50	human; bcl; cortex; brain; oxygen
Small	1	kinase; time; relax; idiotype; leukemia
	2	chicken; tcr; adenovirus; q10; epsilon
	4	chicken; yacs; oocytes; presence; fraction
	8	not; 14c; subunit; cd3; yacs
	16	not; homeodomain; respect; sea; material
	32	not; whether; yacs; question; example
	50	ria; xenografts; nonglycosylated; cytoskeleton; thermolysin
Small, (1-3)-gram	1	time; idiotype; to the; kinase; theory
	2	on; electron; that the; time; scale
	4	by the; subunit; on; oocytes; temperature
	8	by the; from the; globin; trna; subunit
	16	not; ii; chicken; subunit; that the
	32	from the; not; chicken; his; apob
	50	not; from the; intermediate; nb; induced

Table 3.6: Phrases with large and small coefficients in classification of the modified data set.

	θ	Phrases
Large	1	including; within; we; among; thus
	2	including; within; thus; identify; known
	4	encodes; including; known; base; amplified
	8	required; pathogenesis; potent; cloned; includes
	16	tasks; pathogenesis; here; cloned; glutamylcysteine
	32	constituent; glutamylcysteine; bpv; oxidatively; amyloid
	50	homopyrimidine; oxidatively; augment; pipet; suboptimal
Large, (1-3)-gram	1	in X X; that X; X by: X including; that X X
	2	in X X; that X; in X and; X including; by X
	4	in X X; that X; in X and; identified; hcv
	8	in X and; that X; identified; cloned; required
	16	in X and; cloned; the X and; isolates; of X and
	32	cloned; in X and; of X that; tasks; hcv
	50	of X that; ebna; are X of; thus X X; trx
Small	1	when; ca; atoms; if; gt
	2	na; if; hr; phosphorylated; scale
	4	yacs; phosphorylated; much; terms; na
	8	14c; repertoire; revertants; terms; authentic
	16	enzyme; interfere; arbitrary; pe; polymerase
	32	cd2; agrees; preincubated; carboxylase; reinitiation
	50	ria; exclusion; phosphatidylcholine; instructive; organized
Small, (1-3)-gram	1	on; on the; ca; atoms; when
	2	on; that the; if; na; on the
	4	on; na; mm; at the; phosphorylated
	8	mm; from the; by the; 14c; the results
	16	not; when the; ii; mm; injected
	32	kda; mm; not; when the; enzyme
	50	still; not; 65; mtx; X is not

Table 3.7: Phrases with large and small coefficients in classification of the normal data set using technical terms.

	θ	Phrases
Large	1	human; brain; dna; gene; sequence
	2	gene; brain; disease; dna; cell
	4	brain; gene; dna; cell; protein
	8	amyloid; extracellular; fluorescent; gene; endothelial
	16	amyloid; arabidopsis; fibronectin; gene; pathogens
	32	adhesive; p65; synonymous; arabidopsis; lysozyme
	50	bacteriophage p1; heat-shock proteins; adhesive; rna helicase; antiport
Small	1	late; problem; paper; origin; finite
	2	sensitivity; period; problem; fusion; isolated
	4	exposure; temperature; mutagenesis; problem; artificial
	8	probability; accessibility; paper; immunoblot; dehydrogenase
	16	homeodomain; shuttle vector; sepharose; radioimmunoassay; reporter
	32	mosquitoes; heavy-chain; retinol; yeast artificial; orangutan
	50	texas; orangutan; societies; schedules; disparity

Chapter 4

Native Language Identification

Profiling the author of a document is effective in estimating the document's implicit meanings. This study aims to find differences in English writing styles between authors whose native languages are not English. We conduct native language identification based on machine learning and investigated distinctive phrases from each language to determine their writing style. We classify English abstracts from 250,000 research papers written in one of five languages other than English. Additionally, we classify the abstracts modified by generalizing content words to remove the effect of topics specialized in the data.

4.1 Related work

The aim of this study is to find tendencies in writing styles depend on languages using results of NLI. We solved the problem of the lack of supervised data for NLI by using abstracts from research papers.

Preparing a sufficient amount of supervised data for NLI is difficult. According to the report of the NLI shared task 2017 [42], there was no submission to a competition that allows using external training data for NLI. Some content biases are estimated to exist from the results of attempts [19, 47] in cross-corpus NLI. Chen et al. [24] claim that their work is the first attempt to find writing styles related to languages using results of NLI with large-scale data. The novelty of their work is the point that they used Wikipedia [12] data with users' profiles as supervised data for NLI, rather than the point that they used NLI to find writing styles. Some of recent work try to use external data for NLI. Goldin et al. [31] used approximately 200,000,000 sentences posted to Reddit [10] and users' countries for NLI.

The data we used could generate high accuracy (0.91 and 0.75 for five languages) in NLI, which implies that the features obtained from the results of NLI can accurately explain the practical tendencies in the data, rather than that the data can realize an accurate NLI method for general data. From this viewpoint, it is meaningful to compare different data sets in terms of NLI accuracy. In the experiments with Wikipedia data conducted in [24], the accuracy was 0.50 for NLI with six languages chosen from 19 languages and 0.48 for five language families on 17 languages. In the experiments with Reddit data conducted in [31], the accuracy was 0.69 for 23 languages and 0.83 for four language families on the condition that allows using only textual data of comments.

We found some tendencies in writing styles as distinctive phrases used for NLI and found general characteristics by using part of speech (POS) tags, which follows earlier work in NLI. It is straightforward that features used for NLI are

connected to writing styles, especially on the situation that the features are defined based on string occurrences. The early work of NLI by Koppel et al. [36, 37] defined some features for NLI using known tendencies in the writing styles of the target languages. A number of work [35, 40, 31] show distinctive strings obtained from the results of NLI to explain a kind of writing style. As for attempts to find tendencies independent from the data contents, a number of work [50, 19, 20, 41, 40, 31] use occurrences of POS tags. Especially, Bykh and Meurers [20] investigated the case where only content words are converted to the corresponding POS tags in addition to the case where all words are converted.

4.2 Methods

4.2.1 Data

We conducted a classification of English research paper abstracts, where the main text of each paper is written in one of five languages other than English. On the assumption that the main text is written in L1 of the author, the classification into the five classes is regarded as an NLI.

The experimental data were obtained from PubMed [9]. We used research paper abstracts published from 1989 to 2018 available on PubMed. The metadata of each paper added in PubMed includes information about the languages of its abstract and main text. The number of obtained abstracts for the five languages, Chinese, French, German, Japanese, and Spanish, were 134,690, 71,096, 61,419, 53,143, and 58,006, respectively. We randomly selected 50,000 abstracts from each

language in the results for our experiments.

4.2.2 Experiments

First, we examined the accuracy of the classification. We conducted a 10-fold validation with the data set. We applied a linear SVM classifier to word-level n-grams of abstracts weighted using tf-idf. We investigated the classification accuracy for the $(1, n)$-grams of each abstract for some n's to find appropriate features. Then, we predicted the language in the five classes in the following conditions:

- We distinguished between upper- and lower-case letters;
- We deleted the words used for generating a format of abstracts, such as the caption "OBJECTIVES:";
- We didn't use the first and last sentences of each abstract;
- We didn't use phrases that appeared only once in the training data for the classification.

We also conducted the same experiment for modified data to find tendencies in writing styles that are not specialized to the data. We generalized each content word (a noun, a verb, an adjective, an adverb, or a digit) in the abstracts to be the name of its POS. For example, the sentence "I am your father" is modified to "I VBP your NN", where the words "VBP" and "NN" mean a verb (present, singular, non-3rd) and a noun (singular), respectively.

Next, we analyzed distinctive phrases of each class. We trained the classifier using the total data and listed phrases that corresponds to the elements of the

Table 4.1: Accuracy of the classification for $(1, n)$-grams of the normal and modified data.

Data set	$n = 1$	2	3	4	5	6	7
Normal	0.869	0.907	**0.908**	0.907			
Modified	0.576	0.690	0.728	0.742	0.748	**0.750**	**0.750**

Table 4.2: Accuracy of the classifications with the normal data for $(1, 3)$-grams and the modified data for $(1, 6)$-grams.

	Normal data			Modified data		
	Precision	Recall	F-score	Precision	Recall	F-score
Chinese	0.96	0.96	0.96	0.87	0.89	0.88
French	0.87	0.86	0.87	0.68	0.66	0.67
German	0.90	0.91	0.90	0.71	0.75	0.73
Japanese	0.92	0.92	0.92	0.76	0.74	0.75
Spanish	0.90	0.89	0.89	0.73	0.71	0.72
Accuracy	0.91			0.75		

separating hyperplane with large weight. Additionally, we considered to some distinctive phrases of each language.

4.3 Results

Tables 4.1 shows the accuracy of the classifications for the normal and modified data. The accuracy was optimal when $n = 3$ and $n = 6$ for the normal and modified data, respectively. Table 4.2 shows the precise accuracy of the two classifications for the optimum features. Tables 4.3 and 4.4 show the confusion matrices of the classifications.

Table 4.5 shows the top 20 phrases in the order of weights used in the classifier

Table 4.3: Confusion matrix of the classification with the normal data for $(1, 3)$-grams.

	Predicted language					
	CHI	FRE	GER	JPN	SPA	Total
Chinese	47,900	310	347	1,128	315	50,000
French	358	43,210	2,552	1,148	2,732	50,000
German	280	2,069	45,603	997	1,051	50,000
Japanese	1,210	939	1,033	46,061	757	50,000
Spanish	408	2,905	1,407	1,006	44,274	50,000

Table 4.4: Confusion matrix of the classification with the modified data for $(1, 6)$-grams.

	Predicted language					
	CHI	FRE	GER	JPN	SPA	Total
Chinese	44,291	921	1,116	2,661	1,011	50,000
French	1,077	33,118	6,556	3,150	6,099	50,000
German	923	5,470	37,278	3,194	3,135	50,000
Japanese	3,481	2,859	3,617	37,161	2,882	50,000
Spanish	1,313	6,330	3,636	3,068	35,653	50,000

trained with the total normal data. The size of the $(1, 3)$-grams of the normal data was 5,058,220. Table 4.6 shows the phrases for the modified data, where the generalized tags indicate POSs as Table 4.7. The size of the $(1, 6)$-grams of the modified data was 5,267,500.

Table 4.5: Phrases with large weights used in the classifier with the normal data for $(1, 3)$-grams (separated by ";"), where the bold phrases are related to regions.

Chinese	**Chinese**; **China**; P 0; showed that; **Taiwan**; P 0 05; obviously; **Beijing**; And; obvious; were; **in China**; could; respectively; paper; operation; TCM; literatures; data of; used to
French	**French**; **France**; Indeed; remains; **in France**; This; **Quebec**; **Tunisia**; concerned; considered as; were respectively; of cases; essentially; particularly; **Paris**; allowed; sex ratio; allowing; **Tunisian**; noted
German	**German**; **Germany**; und; further; In; **in Germany**; relevant; Additionally; diagnostics; e g; additional; Furthermore; report on; of the; **the German**; additionally; compared to; special; only; up to
Japanese	**Japan**; **Japanese**; **in Japan**; Although; We; However; Recently; revealed; using; **Tokyo**; we; because; examined; useful; The subjects; clarified; **Prefecture**; including; hr; each
Spanish	**Spanish**; **Spain**; **Mexico**; that; **Chile**; **Colombia**; studied; **Argentina**; although; **Mexican**; associated to; To; variables; greater; **in Spain**; diagnosed of; **Chilean**; and; related with; alterations

4.4 Discussion

4.4.1 Main findings

We found distinctive phrases as writing styles of non-native authors. The accuracy of the classification into the five languages was 0.91. Therefore, the phrases with large weights in the classifier were supposed to be distinctive of the corresponding language. The accuracy for the data modified by generalizing content words was still 0.75. In the case where content words are generalized, the distinctive phrases

Table 4.6: Phrases with large weights used in the classifier with the modified data for $(1, 6)$-grams (separated by ";").

Chinese	NNP NNP CD; And; The NNS VBD that; NNS VBD that; VBD that; while; etc; NN NN; The JJ NNS; than that; The; CD CD; Through; the NN NNS; The NNS; By; and VBN; VBN into; VBG NN; CD NN CD
French	This JJ; whatever; et; It; This NN; They; For each; This JJ NN; whatever the; CD NN of NNS; We VBP RB; Our; The NNS VBP; NNS VBG; CD JJ CD of; Their; RB CD NN and CD NN; in CD NN of NNS; in NN of; For each NN
German	RB; und; with NN to; In; VB VBN; Besides; could VB VBN; Within; We VBP on; RB RB; While; Furthermore; Therefore; whether; In NN to; up to; For; NNP und; an; VB VBN that
Japanese	Although; These NNS VBP that; The NNS VBD CD; as VBZ; We VBD; or JJR; We; because; Because; Since; On the JJ NN; we; As; toward; as VBZ CD; On the JJ; Of; those; VBD; As for
Spanish	although; To VB; it; To; that VBD; To VB the; NNS that; by NNS of; and; its; of them; de; NN that; that VBP; NNS VBN; We VBD the NN of a; this; NNP de NNP; Our NN VBD to; all of them

are not related to the content of documents (for example, research topics in this data set); therefore, the obtained characteristics are expected to explain general practices depend on the language.

4.4.2 Key findings

The results of the classifications reflect similarities between languages. As shown in Tables 4.3 and 4.4, the number of confusions was relatively large between

Table 4.7: POS tags used in the modified data.

Tag	POS
NN	Noun, singular
NNS	Noun, plural
NNP	Noun, proper, singular
VB	Verb, base
VBP	Verb, present, singular, non-3rd
VBZ	Verb, present, singular, 3rd
VBD	Verb, past
VBN	Verb, past participle
VBG	Verb, gerund or present participle
JJ	Adjective
JJR	Adjective, comparative
RB	Adverb
CD	Digit

French and German and between French and Spanish. From a linguistic point of view, French and Spanish should be classified into a single class. Geographically, French, German, and Spanish comprise a different class to Chinese and Japanese.

In the rest of this subsection, we give detailed considerations to some distinctive phrases used in the classifications from the viewpoint of practices related to languages.

Some characteristics are found from the result for the normal data as shown in Table 4.5.

- A number of phrases trivially indicate a language; for example, the phrase "Chinese" for the language Chinese. Note that the word "TMC" in the data for Chinese means Traditional Chinese medicine, and that the word "Prefecture" in the data for Japanese means an administrative subdivision

of Japan and is often used with a proper noun.

- The word "Indeed" is distinctive of French. This phenomenon corresponds to the observation by Koppel et al. [37] that "indeed" is frequently used by authors whose L1 is French, and to the result shown by Ionescu et al. [35] that the most discriminating overuse sequence for French is "indeed" in one of their classifiers.

Some characteristics can be found using the result for the modified data shown in Table 4.6.

- Contiguous occurrences of nouns are distinctive of Chinese. The phrases "NNP NNP CD", "NN NN", and "the NN NNS" have large weights for Chinese. This phenomenon is supposed to be related to the fact that nouns can be used for qualifying a noun in Chinese.

- The phrase "On the other hand" is distinctive of Japanese. The 88% of the occurrences of "On the JJ NN" in the modified data for Japanese are caused by the occurrences of "On the other hand" in the corresponding normal data. Paquette [46] point out that the phrase is often used wrongly in research papers written by Japanese and a reason is that the word is conventionally translated to a Japanese word which has more general meanings.

- The use of the relative pronoun "that" is distinctive of Spanish. The word "that" itself has a large weight as shown in Table 4.5. In Table 4.6, "that" is listed as "NNS that", "NN that", "that VBD", or "that VBP", which means that in most occurrences of "that" it is used as a relative pronoun.

A reason is supposed that the relative pronoun "que" in Spanish cannot be omitted and the word tends to be translated into "that" in English.

4.4.3 Future work

Improving the classifier is one of our future work. We used the simple classifier with SVM and the straightforward features based on word occurrences in documents. Ionescu et al. [35] point out that surface-level features are effective for NLI. A number of improvements for classification are proposed, even if the scope is restricted to NLI: Wong and Dras [50] used parse structures for NLI; Bykh and Meurers [20] used n-grams with a simple restriction on their occurrences; Li and Zou [40] used a multilayer perceptron. Another promising approach is the use of well-trained vector representations of words (or phrases), which can represent the similarity of words and be adapted to simple classifiers. Additionally, we need technologies to classify documents using unbalanced training data. For some languages, the amount of available papers, whose abstract is written in English and main text is written in the target language, is small.

Another future work is applying the resulting knowledge of the classification to practical systems; for example, automatic proofreading and e-learning for second languages. Our results can help to find and correct typical errors related to the author's native language. A difficulty is that we must examine the relation between the statistical tendencies obtained from results and practical writing styles of target languages. We need to continue interdisciplinary research with proofreaders, linguists, or native speakers of each language.

Chapter 5

Mental Health Prediction

Mental illnesses should be detected automatically and in their early stages from changes in everyday behavior. This study aims to detect persons who have mental health problems using their comments posted to social network systems. A difficulty of machine learning-based approaches to the detection is the lack of supervised data. To remove the difficulty, we used comments posted to a Web community for persons with mental health problems. In this section, we classify approximately 240,000 Japanese comments obtained from the community and a general social network system, and investigated the distinctive phrases used in the classification. We also conduct the experiment with the comments modified by generalizing content words to remove the effect of topics specialized in the community.

5.1 Related work

The novelty of our study is that we used comments posted to the SNS for persons with mental health problems as supervised data for machine learning. We conducted classification with the comments on the condition that content words were generalized, to investigate the versatility of characteristics obtained from the result of the classification.

Preparing a sufficient amount of supervised data for detecting mental illnesses is difficult. Guntuku et al. [33] review recent research aimed at predicting psychiatric disorder using SNSs. They summarized related studies from the perspective of assessment methods of users' information about their mental health. The methods are roughly classified into two cases: that subjects took a survey about the target illness and that information to identify the target illness is obtained from online sources. In the first case, the number of subjects for positive samples is at most hundreds [25, 49, 48, 28]. The methods in the second case require lower cost for correcting data than the first one. Guntuku et al. mention three assessment criteria for the second case: self-declared status [45, 15], fora or communities users belong [26, 30, 13, 21], and related keywords used in comments [38]. Our study is included in the second criterion. The earlier studies in this approach use data posted to an SNS Reddit [10] to assess tens of thousands of users with mental health problems using sub-communities, called "subreddits", the users belong. Our assessment method using a community for persons with mental health problems is expected to have a better degree of validity than the methods using sub-communities.

A problem in using data obtained from different SNSs is that classifying the comments can sink into a trivial detection of topics specialized in the SNSs. We tried to find general characteristics by converting each content word in comments into the part of speech (POS) tag. It is straightforward to use occurrences of POS tags to find tendencies independent from the data contents. For example, in the earlier studies of detecting mental health problems, Leis et al. [38] use POS tags. We used POS tags for restricted words in two levels. Additionally, our target language is Japanese and therefore the definition of content words should be configured to the situation.

5.2 Methods

5.2.1 Data

We generated two data sets of Japanese comments for our experiments. We used the total comments posted to Cocooru as of October, 2019, as positive data. The number of comments was 120,327 and the number of users was 3,283. As for negative data, we randomly gathered the same number of Japanese comments from Twitter. The number of users was 112,452. The average length (number of words) of the comments in the positive and negative data are 26.0 and 19.1, respectively. The accuracy of a naive classification with the two data sets using the lengths of comments and the optimized threshold was 0.57.

5.2.2 Experiments

We investigated the classification accuracy with the data by conducting a 10-fold cross validation. We applied an SVM to the word-level n-grams of the comments weighted using tf-idf. We didn't use phrases that appeared in training data only once. Additionally, we deleted URLs and email addresses. To find characteristics independent from the data contents, we conducted the classification with the following three kinds of data:

- *Normal* data: the original data, where each numeral is converted to a tag;
- *Generalized* data: obtained from the normal data by converting each content word (a noun, a verb, an adjective, or an adverb) to its POS tag, where modal verbs are not included in verbs;
- *Semi-generalized* data: obtained from the normal data by converting each content word to its POS tag except for formal, adverbial, and temporal norms and adverbs.

5.3 Results

Tables 5.1 shows the accuracy of the classifications for $(1, n)$-grams of the normal and modified data, where the $(1, n)$-grams is the union of the set of the i-grams for $1 \leq i \leq n$. The accuracy was optimal when $n = 2$, $n = 4$, and $n = 3$ for the normal, generalized, and semi-generalized data, respectively. Tables 5.2, 5.3, and 5.4 show the precise accuracy and confusion matrices of the three classifications for the optimal features.

Table 5.1: Accuracy of the classification for $(1, n)$-grams of the normal, generalized, and semi-generalized data.

Data \ n	1	2	3	4	5
Normal	0.898	**0.901**	0.899		
Generalized	0.796	0.819	0.825	**0.826**	0.823
Semi-generalized	0.815	0.831	**0.837**	0.836	

Table 5.2: Confusion matrix of the classification with the normal data for $(1, 2)$-grams.

	Positive	Negative	Recall	F-score
Positive	109,160	11,160	0.91	0.90
Negative	11,936	108,384	0.90	0.90
Precision	0.90	0.91		

Tables 5.5 and 5.6 show the top 10 phrases in the order of weights used in the classifier trained with the total data for the three kinds of data sets. The vocabulary size was 413,358 for the $(1, 2)$-grams of the normal data, 376,262 for the $(1, 4)$-grams of the generalized data, and 230,027 for the $(1, 3)$-grams of the semi-generalized data. Note that basically Japanese nouns are not inflected for grammatical number and that the POSs in Japanese don't completely correspond to that in English.

Table 5.3: Confusion matrix of the classification with the generalized data for $(1, 4)$-grams.

	Positive	Negative	Recall	F-score
Positive	100,246	20,074	0.83	0.83
Negative	21,184	99,136	0.82	0.83
Precision	0.83	0.83		

Table 5.4: Confusion matrix of the classification with the semi-generalized data for $(1, 3)$-grams.

	Positive	Negative	Recall	F-score
Positive	101,495	18,825	0.84	0.84
Negative	19,826	100,494	0.84	0.84
Precision	0.84	0.84		

5.4 Discussion

5.4.1 Main findings

We found the distinctive phrases as characteristics in the writing styles of persons who have mental health problems. The classification accuracy with comments posted to Twitter and Cocooru was 0.90. Therefore, the phrases with large weights of the classifier are distinctive of the comments in Cocooru. As shown in Table 5.5, most phrases with large weights were related to problems in mental health. The accuracy for the data modified by generalizing content words were 0.83 and 0.84. In the case of the generalized data, the distinctive phrases are not related to the content of documents; therefore, the characteristics explain general tendencies depending on the user's mental health.

Table 5.5: Phrases with large weights used in the classifier for the $(1, 2)$-grams of the normal data.

Phrase	Note
親	a noun. *parent*
自殺	a noun. *suicide*
リスカ	a noun. *wrist-cutting*
鬱	a noun. *depression*
カウンセリング	a noun. *counseling*
坊	a noun. *boy*
自分	a noun. *myself*
カウンセラー	a noun. *counselor*
自傷	a noun. *self-mutilation*
過食	a noun. *overeating*

5.4.2 Key findings

The result for the normal data shows that a number of phrases related to mental health problems are used in the classifier. In Table 5.5, the distinctive phrases except for "親 (parent)", "坊 (boy)", and "自分 (myself)" mean actions or matters directly related to mental health problems. The three words are related to users themselves or their families. In the classifier, also the word "毒親" which means "toxic parent" has a large weight. The word "坊" is supposed to be used for referring to his or her son. The result of the other word corresponds to the observation in English comments by Eichstaedt el al. [28] that users with mental health problems tend to use first-person singular pronouns.

The result for the modified data indicates characteristics in the writing styles rather than the contents. In Table 5.6, the word "のに" is an adversative conjunction which contains author's complaint about the result. The phrase "の か な"

means an interrogation. Although the words “んだろう” and “のだろう” mean expectations, the words mean interrogations when using with a word corresponds to “why”. In the result for the semi-generalized data, we found some adverbs which are distinctive of the positive data. The word “何もかも” means “everything” which is relatively exaggerated. The word “何で” corresponds to “why” which is supposed to be used with the previous phrases of interrogations. The word “所詮” means “in the end” which intimates a situation that the result is not affected by the process.

5.4.3 Future work

Improving the classifier is one of our future work. We used the simple classifier with SVM and the straightforward features based on word occurrences. Although applying other machine learning methods to the classification is expected to improve the accuracy, it become difficult to explain the classifier in terms of practical writing styles. A promising approach is using a well-trained vector representation of words (or phrases), which can represent the similarity of words and be adapted to the simple classifier.

Predicting mental illnesses in their early stages is also one of our future work. The classifier we generated can put a continuous degree of mental health to each comment. Predicting some kinds of changes in mental health is possible by analyzing multiple comments ordered as a time series with the scores.

Table 5.6: Phrases with large weights used in the classifiers for $(1, 4)$-grams of the generalized data and the $(1, 3)$-grams of the semi-generalized data. The tags "N" and "V" denote a noun and a verb, respectively. The word "joshi" is a POS used in Japanese which means a postpositional word functioning as an auxiliary to a main word.

	Phrase	Note
Generalized	のに	a conjunction, adversative
	んだろう	a modal verb, expectation
	の か な	three joshi, interrogation
	のだろう	a modal verb, expectation
	お N さん に	a prefix, a noun, a suffix, and a joshi. *visitor* or *customer* with N:"客" (68.6%)
	V N N 時	a verb, two nouns, and a suffix means *time*
	たい	a modal verb, intention
	ない	a modal verb, negation
	ばかり	a joshi, limitation
	ばっか	a joshi, limitation
Semi-generalized	んだろう	-
	のに	-
	のだろう	-
	何もかも	an adverb. *everything*
	不 N 者	a prefix, a noun, and a formal noun. *sociopath* with N:"適合" (92.3%)
	イライラ V	an adverb and a verb. *get irritated* with V:"する" (97.3%)
	何で	an adverb. *why*
	ない	-
	たい	-
	所詮	an adverb. *in the end*

Chapter 6

Conclusion

We proposed a novel method for analysing documents. The proposed method can extract implicit knowledge by classifying documents using non-content features. We applied the method to the three tasks: citation count prediction, native language identification, and mental health prediction, for evaluation. For the task of citation count prediction, we could estimate the future impact of a research paper using only its abstract. The impact was discovered independently of the research topic. For the task of native language identification, we achieved high accuracy and the characteristics in the English writing styles of non-native writers. The general results were achieved from specific documents, research papers. For the task of mental health prediction, we found the characteristics in comments of persons with mental health problems. The obtained classifier with only non-content features generated higher accuracy than that of a classifier with the total features.

Bibliography

[1] Cocooru. https://cocooru.com. Accessed Nov. 19, 2019.

[2] COVID-19 Open Research Dataset Challenge. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed Nov. 24, 2020.

[3] Depression, World Health Organization. http://www.who.int/en/news-room/fact-sheets/detail/depression. Accessed Nov. 19, 2019.

[4] Europe PMC: Europe PubMed Central. https://europepmc.org/. Accessed Feb. 5, 2018.

[5] Facebook. https://www.facebook.com. Accessed Nov. 19, 2019.

[6] MeSH: Medical Subject Headings. https://www.nlm.nih.gov/mesh/. Accessed Feb. 5, 2018.

[7] NLTK. https://www.nltk.org/. Accessed Oct. 25, 2019.

[8] PNAS: Proceedings of the National Academy of Sciences. http://www.pnas.org/. Accessed Feb. 5, 2018.

[9] PubMed. https://www.ncbi.nlm.nih.gov/pubmed/. Accessed Oct. 25, 2019.

[10] Reddit. https://www.reddit.com/. Accessed Nov. 8, 2019.

[11] Twitter. https://twitter.com. Accessed Nov. 19, 2019.

[12] Wikipedia. https://www.wikipedia.org/. Accessed Nov. 8, 2019.

[13] S. Bagroy, P. Kumaraguru, and M. De Choudhury. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 1634–1646, New York, NY, USA, 2017. ACM.

[14] D. Becker, W. van Breda, B. Funk, M. Hoogendoorn, J. Ruwaard, and H. Riper. Predictive modeling in e-mental health: A common language framework. *Internet Interventions*, 12:57–67, 2018.

[15] A. Benton, M. Mitchell, and D. Hovy. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

[16] Y. Berzak, C. Nakamura, S. Flynn, and B. Katz. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 541–551, 2017.

[17] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[19] J. Brooke and G. Hirst. Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, pages 391–408. The COLING 2012 Organizing Committee, 2012.

[20] S. Bykh and D. Meurers. Native language identification using recurring n-grams – investigating abstraction and domain dependence. In *Proceedings of COLING 2012*, pages 425–440. The COLING 2012 Organizing Committee, 2012.

[21] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro. Early detection of depression: Social network analysis and random forest techniques. *J Med Internet Res*, 21(6), Jun 2019.

[22] P. L. Callahan, S. Mizutani, and R. J. Colonno. Molecular cloning and complete sequence determination of rna genome of human rhinovirus type 14. *Proceedings of the National Academy of Sciences*, 82:732–736, 1985.

[23] J. Chen and C. Zhang. Predicting citation counts of papers. In *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, pages 434–440, July 2015.

[24] Y. Chen, R. Al-Rfou, and Y. Choi. Detecting english writing styles for non native speakers. *CoRR*, abs/1704.07441, 2017.

[25] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. AAAI, July 2013.

[26] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, New York, NY, USA, 2016. ACM.

[27] Y. Dong, R. A. Johnson, and N. V. Chawla. Can scientific impact be predicted? *IEEE Transactions on Big Data*, 2(1):18–30, March 2016.

[28] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, and H. A. Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[29] E. Garfield. The history and meaning of the journal impact factor. *JAMA*, 295(1):90–93, 2006.

[30] G. Gkotsis, A. Oellrich, T. Hubbard, R. Dobson, M. Liakata, S. Velupillai, and R. Dutta. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 63–73, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[31] G. Goldin, E. Rabinovich, and S. Wintner. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empir-*

ical Methods in Natural Language Processing, pages 3591–3601. Association for Computational Linguistics, 2018.

[32] S. Guido and A. C. Müller. *Introduction to Machine Learning with Python*. O'Reilly Media, Inc., Oct. 2016.

[33] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

[34] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46):16569–16572, November 2005.

[35] R. T. Ionescu, M. Popescu, and A. Cahill. String kernels for native language identification: Insights from behind the curtains. *Comput. Linguist.*, 42(3):491–525, Sept. 2016.

[36] M. Koppel, J. Schler, and K. Zigdon. Automatically determining an anonymous author's native language. In *Proceedings of the 2005 IEEE International Conference on Intelligence and Security Informatics*, pages 209–217, Berlin, Heidelberg, 2005. Springer-Verlag.

[37] M. Koppel, J. Schler, and K. Zigdon. Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 624–628, New York, NY, USA, 2005. ACM.

[38] A. Leis, F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz. Detecting signs of depression in tweets in spanish: Behavioral and linguistic analysis. *Journal of medical Internet research*, 21, 2019.

[39] C.-T. Li, Y.-J. Lin, R. Yan, and M.-Y. Yeh. Trend-based citation count prediction for research articles. In T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, and H. Motoda, editors, *Advances in Knowledge Discovery and Data Mining*, pages 659–671, Cham, 2015. Springer International Publishing.

[40] W. Li and L. Zou. Classifier stacking for native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 390–397. Association for Computational Linguistics, 2017.

[41] S. Malmasi and M. Dras. Multilingual native language identification. *Natural Language Engineering*, 23(2):163–215, 2017.

[42] S. Malmasi, K. Evanini, A. Cahill, J. Tetreault, R. Pugh, C. Hamill, D. Napolitano, and Y. Qian. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[43] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[44] H. Morita, D. Kawahara, and S. Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Pro-*

ceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2292–2297, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[45] M. Nadeem. Identifying depression on twitter. *CoRR*, abs/1607.07384, 2016.

[46] G. Paquette. *English Composition for Scholarly Works (in Japanese)*. Kyoto University Press, 2004.

[47] F. Rangel, P. Rosso, J. Brooke, and A. Uitdenbogerd. Cross-corpus native language identification via statistical embedding. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 39–43. Association for Computational Linguistics, 2018.

[48] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*, 7(1), 2017.

[49] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3187–3196, New York, NY, USA, 2015. ACM.

[50] S.-M. J. Wong and M. Dras. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610. Association for Computational Linguistics, 2011.

[51] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li. Citation count prediction: Learning to estimate future citations for literature. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1247–1252, New York, NY, USA, 2011. ACM.

[52] D. Yogatama, M. Heilman, B. O'Connor, C. Dyer, B. R. Routledge, and N. A. Smith. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 594–604. Association for Computational Linguistics, 2011.

Acknowledgement

This paper summarizes the research results of the author while enrolled in the doctoral program at the Department of Informatics, Graduate School and Factory of Information Science and Electrical Engineering, Kyushu University.

Associate professor Ikeda of the same department, was given the opportunity to carry out this research as an academic advisor, and he gave me guidance from beginning to end. I would like to express my deepest gratitude here.

Professor Tomiura of the same department, Professor Takeda of the same department, and Professor Hirokawa of the Department of Advanced Infomation Technology provided advice as a deputy examiner and provided guidance in every detail of this paper. I would like to express my gratitude.

I would like to express our deep gratitude to Professor Baba belongs to Cyber-physical Engineering Informatics Research Core, Okayama University, for his useful discussions and advice.

Finally, I received a great deal of support from my wife Natsumi through this research and the preparation of this paper. I would like to express our deepest gratitude here.