EasyChair Preprint

№ 6692

Identifying Legality of Japanese Online Advertisements Using Complex-Valued Support Vector Machine with DFT-Based Document Features

Satoshi Kawamoto, Toshio Akimitsu and Kikuo Asai

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

September 25, 2021

Identifying Legality of Japanese Online Advertisements using Complex-valued Support Vector Machine with DFT-based Document Features

Satoshi Kawamoto[1,2], Toshio Akimitsu[1], and Kikuo Asai[1]

[1] The Graduate School of Arts and Sciences, The Open University of Japan, Japan
[2] Engineering Div. i-mobile Co.,Ltd., Japan kawamoto@i-mobile.co.jp
https://www.i-mobile.co.jp/

Abstract. As Internet advertising market expands, the number of advertisements containing inappropriate language is increasing. Advertisements that exaggerate the efficacy of products may contravene the Pharmaceutical Affairs Law and the Act against Unjustifiable Premiums and Misleading Representations. Therefore, a system that can detect problematic expressions is required. Some advertisements cannot be classified using only the statistics of words. Therefore, embedding other information, such as word order and word period in the features is effective to categorize documents. However, the number of labeled data in advertising documents is limited; consequently, models with complex structures tend to overlearn. In addition, features and discriminant models with high generalization performance must be found even if the number of data is small. To address these severe issues, we propose a document feature based on the discrete Fourier transform(DFT) of word vectors weighted using an index previously proposed in a study that attempted to categorize Chinese online advertisements. We also propose a document discriminant model based on a complex-valued support vector machine. We demonstrate that the proposed model outperforms previous models in terms of discriminative performance of F-measure. We found that the proposed index emphasizes word vectors of specific nouns and verbs in Japanese advertisements. In addition, we found that DFT significantly increased the norms of document vectors of illegal documents. These factors contributed to the better performance of the proposed model.

Keywords: Discrete Fourier Transform · Natural Language Processing · Internet Advertisement · Complex-valued Support Vector Machine

1 Introduction

To enhance their appeal, online advertisements often contain text as well as images and videos. Textual information makes it easier to convey the appeal of a product. However, although text can increase the effectiveness of an advertisement, the text may include legally or ethically inappropriate expressions. Advertising service providers exclude inappropriate advertisements through manual screening; however, as the Internet advertising market expands, the cost of the screening process is increasing. Therefore, to reduce the workload, a method that automatically identifies inappropriate advertising expressions is required. Such an automatic identification system can reduce the risk of unintentional delivery of inappropriate advertisements.

As shown in Table 1, there is a finite number of advertising documents that are labeled legal or illegal. In addition, as Huang et al. [2] pointed out, determininig the legality of advertisements requires legal training. Consequently, it is impractical to prepare annotation data using methods that do not involve legally trained annotators, such as crowdsourcing. In addition, as shown in Section 4.3, some documents cannot be identified only based on simple statistics of the words that appear in the document. Therefore, it is necessary to maintain the simplicity of the discriminant models and features while embedding word order and other information into the features.

In this study, we first define inappropriate advertising expressions and describe the characteristics of problematic documents. Then, we propose a document embedding method based on the index proposed by Tang et al. [1] and discrete Fourier transform(DFT) to identify illegal advertisements effectively. We conducted simulations using complex-valued support vector machines to obtain accuracy, precision, recall, and F-measure values. The results demonstrate that, in terms of discriminative performance(evaluated by F-measure), the proposed model outperformed models proposed in previous studies [1] [2].

2 Relatated Work

Since 2014, many studies have investigated document identification in web content, such as determining whether an advertisement is legal or whether a news article is fake news.

Tang et al. [1] proposed a method to determine the legality of Chinese advertisements using unigram and support vector machines. They showed that word weightings using Equation (1)(Section 4.1) improved accuracy.

Huang et al. [2] proposed a model to discriminate the legality of Chinese advertisements using a dependency-based CNN [3]. They showed that additional inputs of syntactic structures into the CNN improves the discriminative performance compared to only inputting the word vectors. The overall structure of the CNN is based on a previous model [4]. In their study, accuracy, precision, recall, and F-measure values were evaluated. Their proposed model showed overall high discriminative performance.

Zhang et al. [7] proposed a model based on neural networks to detect fake news, and Kaurr et al. [8] proposed a method to detect fake news by majority vote using multiple features, such as TF-IDF and BOW, and multiple discriminative models, such as support vector machine and logistic regression.

Mahajan et al. [12] proposed using wavelet coefficients to reduce the dimensionality of a document vector represented as a bag of words. In their model, a document vector is considered a one-dimensional sequence of signals, and its dimensionality reduction is performed by wavelet transform. Mahajan et al. showed that detection performance does not degrade in the SMS spam detection task.

Wieting et al. [13] devised BOREP, which multiplies a sequence of word vectors by a random matrix and creates a document vector using a pooling function. Despite its simplicity, BOREP exhibited high performance.

Devlin et al. [14] employed a neural network-based technique, i.e. BERT, and demonstrated that their proposed method achieved high performance on various tasks related to natural language processing, suggesting the effectiveness of the attention mechanism.

3 Legality of Advertising Documents

3.1 Definition of problematic documents

Occasionally, Internet advertisements contain inappropriate materials from legal and ethical perspectives. It is necessary to clearly define inappropriate advertisements to create a system to detect such documents.

In this study, we defined problematic advertisements based on the Pharmaceutical Affairs Law. Advertising expressions for cosmetics are regulated by Article 66 of the Pharmaceutical Affairs Law, which prohibits false and exaggerated advertising. In addition, the Ministry of Health, Labour and Welfare's Standards for Proper Advertising of Drugs and Other Products [5] provides specific standards. In the following, we define problematic expressions and present concrete examples.

Restrictions on expressions related to efficacy and safety The possible range of expressions regarding efficacy for cosmetics is given in the Pharmaceutical Affairs Law No. 0721-1. Expressions such as "eliminates fine lines and wrinkles," "has an anti-aging effect," and "improves wrinkles and sagging skin," are prohibited in cosmetics advertisements. In addition, there are strong restrictions on the use of efficacy and safety claims for pharmaceuticals and quasi-drugs, including cosmetics. Specifically, it is prohibited to use historical expressions, e.g., "effective based on evidence from the past 100 years" and give examples of clinical or experimental data. Expressions that guarantee efficacy, e.g., "few side effects," are also not permitted. Note that testimonials about the impressions of using a product are permitted; however, testimonials regarding efficacy and safety are not permitted. Relative to efficacy and safety, statements that claim a maximum level of efficacy or productivity, e.g., "the best efficacy" or "the ace of gastrointestinal drugs" are also not permitted.

Restrictions on expressions about ingredients and raw materials Restrictions on special labeling for cosmetics are outlined in the Standards for Proper Advertising of Drugs and Other Products. In the case of special labeling of raw materials, the purpose of their inclusion (within the range of efficacy approved for cosmetics) should be stated clearly.

Restrictions on slanderous advertising of other companies' products Defamatory expressions, e.g., "this works better than other companies' products" are not permitted.

Recommendations from pharmaceutical professionals, etc. Advertisements that contain expressions that convey endorsements or recommendations by pharmaceutical professionals, clinics, universities, or other institutions are prohibited. This type of expression is not permitted even if they are true, which means that strong restrictions are placed on advertisements that may have substantial impacts on people's decisions. In addition, expressions regarding patents are also inappropriate even if true.

4 Features of Inappropriate Advertising Documents

4.1 Frequency features of words

In the previous section, we described the definitions of inappropriate advertising documents. In Section 4.1 and 4.2, we discuss the statistical characteristics of problematic documents. Tang et al. [1] identified that there were differences in the frequency of word occurrence between normal and inappropriate advertisements. In addition, they proposed to use Equation (1) to weight word vectors. Their simulation using an SVM demonstrated that the weighting of word vectors improves discrimination accuracy.

$$U_w = \log \left(\frac{\left(\frac{l_w}{L}\right)}{\left(\frac{k_w}{K}\right)} \right) \tag{1}$$

Here, l_w is the number of words w that appear in problematic advertisements, k_w is the number of occurrences of w in nonproblematic advertisements. L is the total number of words (i.e., tokens) in problematic advertisements, and K is the number of tokens in the nonproblematic advertisements.

In this section, we describe the features of the top-level words of U_w in advertisements provided by i-mobile Co.,Ltd. As shown in Table 1, the advertisements include documents about cosmetics, health foods, and other products. In addition, the advertisements for cosmetics and health food are labeled to identify whether there are problems relative to the Pharmaceutical Affairs Law. Here, positive and negative labels are applied by the holder of a pharmaceutical law administrator license.

As shown in Tables 2, and 3, words related to medicine, e.g., "medicine" and "pharmaceutical" appear more frequently in problematic advertisements. As described in Section 3.1, recommendation expressions in advertisements by pharmaceutical professionals are not permitted; therefore, U_w tends to be higher for words related to pharmaceuticals.

In this study, we used MeCab (version 0.996) for morphological analysis, and the default IPA dictionary was used.

Table 1. Number of advertisements

Total number of advertisements	78581
Cosmetics (nonproblematic documents)	8103
Cosmetics (problematic documents)	3008
Health Foods(nonproblematic documents)	12999
Health Foods(problematic documents)	1487

Table 2. High U_w words(cosmetics)

Word	U_w	POS
極限 (limit)	4.309	Noun
ウチ (inner)	4.053	Noun
綿棒 (cotton swab)	3.871	Noun
大学 (university)	3.697	Noun
(company name)	3.648	Noun
誌 (magazine)	3.471	Noun
医学 (medical science)	3.401	Noun
放っ(leave)	3.360	Verb
医薬品 (pharmaceuticals)	3.332	Noun
地肌 (skin)	3.273	Noun

Table 3. High U_w words(health foods)

Word	U_w	POS
医学 (medical science)	4.893	Noun
誌 (magazine)	4.794	Noun
すすめる (recommend)	4.519	Verb
作り方 (how to make)	4.519	Noun
排便 (bowel movement)	4.505	Noun
医師 (medical doctor)	4.359	Noun
掲載 (publication)	4.118	Noun
歯医者 (dentist)	3.949	Noun
? !?	3.949	Noun
断言 (affirm)	3.949	Noun

4.2 Part-of-speech features

It is necessary to identify effective features to determine the legality of advertisements. Table 4 shows the percentage of occurrence of the parts of speech in each document type. Unfortunately, there is no characteristic that a particular part of speech is more likely to appear in problematic documents. In other words, the expressions in problematic advertisements do not deviate from the Japanese grammar.

However, if we plot the distribution of U_w in parts-of-speech units, we can observe large differences in their distribution. Fig. 1 plots the distribution of U_w for each part of speech in cosmetic advertisements. As shown in Fig. 1, the variance of U_w is large for nouns and verbs. In particular, U_w for nouns shows many outliers, which means that there are nouns and verbs that are likely to appear in illegal advertisements. For prefixes, particles, auxiliaries, and symbols, the variance of U_w is small (although there are some outliers), and the influence of the legality of documents is small.

4.3 Documents that cannot be discriminated by words only

As described in Sections 4.1, and 4.2, the frequency of nouns related to medical field (e.g., “medicine” and “drug”) is high in problematic advertisements. In addition, the frequency of verbs that appear in contexts where medical professionals recommend products (e.g., “can” and “recommend”) is high. However, we cannot judge a document as problematic simply because it contains words with large U_w. In the following, we present examples where legality cannot be determined using only the occurrence of words.

Table 4. Occurrence of each part of speech in advertising documents

	Cosmetics (illegal)	Cosmetics (legal)	Foods (illegal)	Foods (legal)
particle	23.2720%	23.2153%	22.6153%	21.3215%
auxiliary verb	3.8632%	4.3906%	4.0897%	4.9718%
adjective	1.2968%	1.6053%	0.9073%	1.2065%
symbol	12.2399%	12.9804%	12.0993%	13.0527%
interjection	0.0972%	0.1196%	0.0228%	0.0601%
filler	0.0145%	0.0378%	0.0105%	0.0252%
conjunction	0.1235%	0.1446%	0.1033%	0.1258%
prefix	1.1261%	1.3446%	0.9949%	1.0684%
verb	9.5028%	10.3234%	10.8663%	11.7316%
adverb	1.7118%	2.6067%	1.8969%	3.1471%
adnominal adjective	0.3324%	0.5450%	0.1769%	0.3245%
noun	46.4197%	42.6848%	46.2168%	42.9647%
others	0.0000%	0.0019%	0.0000%	0.0000%

Documents in which subjects are not medical professionals As discussed in Section 3.1, texts in which medical professionals recommend products are not permitted; however, there are cases where the subjects are not medical professional, as in the following example.

> 皮膚科医の妻「毛穴汚れはこれ」簡単すぎて話題に (in Japanese)
> Dermatologist's wife said, "Here's how to clean your pores." It's too easy and went viral.

In this case, the subject is the "dermatologist's wife," which does not correspond to the expression of the doctor's personal recommendation.

Items that do not express efficacy explicitly There are strong restrictions on advertising expressions about efficacy. For example, the statement, "the effects of the cosmetics will make your skin beautiful" is not permitted. However, some advertising expressions do not explicitly state the existence of efficacy, as in the following example.

> 20代に見える40代女医さんの透明感の秘密！(in Japanese)
> The secret to the beautiful skin of a forty-something female doctor who looks like she's in her twenties!

5 Features to Discriminate Advertising Documents

5.1 Properties of effective features for discrimination

As discussing in Section 4.3, we cannot detect all illegal advertisement documents accurately using only U_w. However, if the positional information of words is

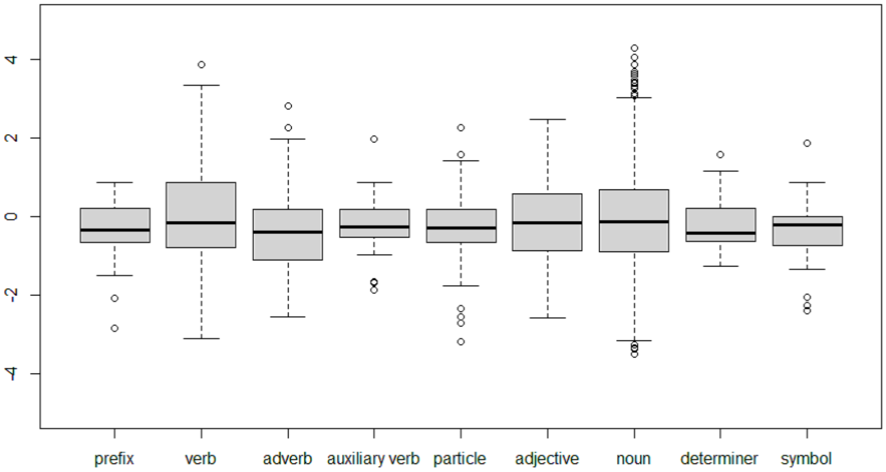

Fig. 1. Distribution of U_w of words in each part of speech

embedded in the document vector, it is possible to determine whether the subject is a medical professional. In addition, document vectors embedded with the periodic features of words are effective relative to discriminating contrastive expressions, e.g., “40-something who looks like a 20-something” or “I lost 60kg to 45kg.”

In addition, as shown in Fig. 1, certain nouns and verbs are used at high frequency in problematic documents, and it is expected that weighting the word vectors will improve the discrimination performance of advertisement documents.

Threfore, we propose a document vector that combines weighting of word vectors by the log frequency ratio(U_w) and DFT.

5.2 Document vector combining U_w and DFT

When a document D comprises a sequence of words $(l_0, l_1, ..., l_{N-1})$, D has a sequence of word vectors $(\mathbf{v}_{l_0}, \mathbf{v}_{l_1}, ..., \mathbf{v}_{l_{N-1}})$. Here, $\mathbf{v}_{l_t}(t = 0, 1, 2, ., N-1)$ are 200-dimensional word vectors created by word2vec (skip-gram; window size: 10). Then, we define $(\mathbf{u}_{l_0}, \mathbf{u}_{l_1}, ..., \mathbf{u}_{l_{N-1}})$ as follows.

$$(\mathbf{u}_{l_0}, \mathbf{u}_{l_1}, ..., \mathbf{u}_{l_{N-1}}) = (U_{l_0}\mathbf{v}_{l_0}, U_{l_1}\mathbf{v}_{l_1}, ..., U_{l_{N-1}}\mathbf{v}_{l_{N-1}}) \tag{2}$$

The DFT of this sequence is expressed as follows.

$$\mathbf{F}(\theta) = \sum_{t=0}^{N-1} \mathbf{u}_{l_t} \exp\left(-i\frac{2\pi(\theta-1)}{N}t\right) \tag{3}$$

Here, t is the position at which the word appears. The document vector $\mathbf{x}_D$ is obtained by multiplying $\mathbf{F}(\theta)$ by a random matrix $\mathbf{W}_\theta$ as follows.

$$\mathbf{x}_D = \sum_{\theta=1}^{\Theta} \mathbf{F}(\theta)\mathbf{W}_\theta \tag{4}$$

Here, $\Theta \in \{1, 2, 3, 4, 5\}$. We also evaluate the performance of F-measure when the word vectors are not weighted, which is discussed in Section 7.3(see Equation (3), where $\mathbf{v}_{l_t}$ is used rather than $\mathbf{u}_{l_t}$). $\mathbf{W}_\theta$ is a random matrix obtained via sparse random projection [15] as follows.

$$W_{\theta_{kl}} = \begin{cases} -1 & (\text{with probability } \frac{1}{6}) \\ 0 & (\text{with probability } \frac{2}{3}) \\ 1 & (\text{with probability } \frac{1}{6}) \end{cases} \tag{5}$$

Here, $W_{\theta_{kl}}$ is the kth row and lth column element of the matrix (where $1 \leq k, l \leq 200$).

6 Discriminating Documents using Complex-valued Support Vector Machine

As shown in Table 1, the number of positive examples of the data used in this study is in the order of thousands. Therefore, to determine whether advertising document D is problematic document, we must use a discriminant model with high generalization performance. Thus, we employed the complex-valued support vector machine(CV-SVM) [9] as a discriminant model with high generalization performance. The discriminant function of CV-SVM is expressed as $f(\mathbf{x}_D) = \mathbf{w}\phi(\mathbf{x}_D^*) - b$. Here, $\mathbf{w}$ is a complex-valued weight vector, and $\mathbf{x}_D^*$ is the vector in which each element of $\mathbf{x}_D$ is conjugated. In addition, $\phi(\mathbf{x})$ is the basis function, and b is the bias term of the complex number.

The objective function E is expressed as follows, where the problem is to minimize E. Here, Γ is the document set and α_D, β_D are the Lagrange coefficients. If D is a problematic advertising document, y_D is labeled $y_D = 1$; otherwise, $y_D = -1$. Note that ξ_D, ζ_D are the relaxation parameters of the constraints.

$$\begin{aligned} E = \frac{1}{2}|\mathbf{w}|^2 &- \sum_{D \in \Gamma} \alpha_D \left(\mathrm{Re}\left(y_D(\mathbf{w}\mathbf{x}_D^* - b)\right) - 1 + \xi_D\right) \\ &- \sum_{D \in \Gamma} \beta_D \left(\mathrm{Im}\left(y_D(\mathbf{w}\mathbf{x}_D^* - b)\right) - 1 + \zeta_D\right) \\ &+ C \sum_{D \in \Gamma} \xi_D + C \sum_{D \in \Gamma} \zeta_D \end{aligned} \tag{6}$$

However, it is easier to solve the dual problem than solve Equation(6). The dual problem was proved to be derived using the Wiltinger derivative by Bouboulis [10]. Specifically, $\frac{\partial E}{\partial \mathbf{w}^*}, \frac{\partial E}{\partial b^*}, \frac{\partial E}{\partial \xi_D}, \frac{\partial E}{\partial \zeta_D}$ are calculated and the dual problem is expressed as follows.

$$E = -\frac{1}{2} \sum_{D_1 \in \Gamma} \sum_{D_2 \in \Gamma} \psi_{D_1} \cdot \psi^*_{D_2} \cdot y_{D_1} \cdot y_{D_2} \cdot K\left(\mathbf{x}_{D_1}, \mathbf{x}_{D_2}\right) + \sum_{D \in \Gamma} (\alpha_D + \beta_D) \quad (7)$$

where $\psi_D = \alpha_D + i\beta_D$. In addition, the following conditions must be satisfied as constraints.

$$\sum_{D \in \Gamma} \alpha_D \cdot y_D = 0, \sum_{D \in \Gamma} \beta_D \cdot y_D = 0, 0 \leq \alpha_D, \beta_D \leq C \quad (8)$$

The discriminant function is obtained by finding α_D, β_D that maximizes E while satisfying the constraints. When $\mathrm{Re}(f(\mathbf{x}_D)) + \mathrm{Im}(f(\mathbf{x}_D)) \geq 0$, D is considered a problematic document; otherwise, D is a nonproblematic document. In this study, we use the RBF kernel function $K(\mathbf{x}_1, \mathbf{x}_2)$, which is defined as follows.

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{(\mathbf{x}_1 - \mathbf{x}_2)(\mathbf{x}_1 - \mathbf{x}_2)^*}{\sigma^2}\right) \quad (9)$$

7 Discrimination Simulation of Cosmetic Advertisements

7.1 Performance indicators of discriminant model

We conducted a numerical evaluation of a model to discriminate the legality of cosmetic advertisements. We found relatively few positive examples, as shown in Table 1. Therefore, it is desirable to evaluate the performance of the discriminant model using a metric other than accuracy.

It is desirable to have high recall and precision with a discrimination model. Therefore, we evaluated model performance using the F-measure as an index.

7.2 Simulation using holdout method

Here, we used the holdout method in the simulation to compare F-measure. Specifically, we split the data in Table 1 into training data, validation data, and test data at ratio of 2:1:1, respectively. In other words, the model was trained using the training data, the parameters with high F-measure were searched using the validation data, and the actual performance of the model was evaluated using the test data.

The C parameters of the SVM and CV-SVM were fixed at $C = 256$, and the σ^2 parameters were selected from $\sigma^2 \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$ using the grid search method.

7.3 Discriminant models and document vectors to compare

The word vectors in this simulation involve two patterns, i.e., word2vec word vectors weighted by U_w and and unweighted word vectors.

In addition, SWEM-Aver [11] and the document vector defined by Equation (4) were compared in the simulation, which was performed with five patterns of $\Theta(\in \{1, 2, 3, 4, 5\})$.

Here, we compared Tang's SVM method [1], Huang's CNN method [2], and the proposed CV-SVM method. Huang showed that adding word vectors and clause structures to the input vectors improved the discriminative performance of the CNN method; however, the improvement was limited. Therefore, to simplify implementation, in this simulation, the comparison was performed without adding the clause structure to the input vector.

The patterns of the simulation are shown in Table 5. Here, the weighted word vectors are denoted word2vec(U_w), and the unweighted word vectors are denoted as word2vec.

Table 5. Simulation Patterns

Discriminant model	Word vector	Document vector
SVM	word2vec	SWEM-Aver
CNN	word2vec	-
CV-SVM	word2vec	DFT($\Theta \in \{1, 2, 3, 4, 5\}$)
SVM	word2vec(U_w)	SWEM-Aver
CNN	word2vec(U_w)	-
CV-SVM	word2vec(U_w)	DFT($\Theta \in \{1, 2, 3, 4, 5\}$)

The configuration of the CNN used for comparison is shown in Table 6. Here, the dropout rate was set to 0%. N is the total number of words in the document, and the labels used for training are $y = 1$ for problematic advertisements and $y = -1$ for nonproblematic advertisements.

Table 6. CNN configuration

Unit	Detail
input layer	200(Dimensionality of word vectors)$\times$ N
convolutional layer(ReLU)	200 $\times$ 3 : 100channels
	200 $\times$ 4 : 100channels
	200 $\times$ 5 : 100channels
pooling layer	Max Pooling
fully-connected layer	activation function : ReLU
output layer	activation function : y=x

7.4 Simulation results

We simulated the discrimination of problematic advertisements using the patterns given in Table 5, and the results are shown in Table 7.

Here, when $\Theta = 1$, the phase is zero even if the word position t changes, as shown in Equation (3) and (4). In addition, the document vector $\mathbf{x}_D$ is the same as BOREP [13] using average pooling. The difference between SWEM-Aver [11] and $\mathbf{x}_D$ is the presence of a random matrix. In this simulation, when weighted by U_w, multiplying SWEM-Aver by a random matrix greatly improved the accuracy and F-measure values. One reasons of this result might be that multiplying by a random matrix embeds a small amount of noise in the document vector, which helps to suppress overtraining; however, further numerical simulation and investigation are needed to clarify what this result means. When the word vectors are not weighted, multiplying by a random matrix is ineffective. Determining the conditions by which a random matrix works effectively for the discriminant model is left to a future work.

When $\Theta = 2$, the document vector $\mathbf{x}_D$ is BOREP (average pooling) with additional word order information. In addition, for $\Theta = n$, words that occur $n - 1$ times in the document are highlighted and embedded in the document vector. As a result, as Θ increases, the word statistics, word order, and period information are embedded in sequence.

In this simulation, precision was improved from 0.7607 to 0.8543 by increasing Θ when weighting by U_w was involved. However, recall was lower when $\Theta \geq 4$, resulting in the highest F-measure and accuracy values when $\Theta = 3$. The fact that precision increased as Θ increased means that proposed CV-SVM was able to accurately determine the legality of documents in which the same word (or similar words) appears multiple times. However, as Θ increased, it became increasingly difficult to discriminate documents without occurrences of the same word due to the noise caused by the random matrix.

Table 7. Simulation results

	σ^2	WordVector	DocumentVector	Accuracy	Precision	Recall	F-value
SVM	1.0×10	word2vec	SWEM-Aver	0.8646	0.7401	0.7696	0.7546
CNN	-	word2vec	-	0.6485	0.4070	0.6516	0.5010
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 1$)	0.8545	0.7430	0.7074	0.7248
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 2$)	0.8836	0.8060	0.7513	0.7777
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 3$)	0.8527	0.7308	0.7221	0.7264
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 4$)	0.8509	0.7012	0.7832	0.7399
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 5$)	0.7950	0.5956	0.7550	0.6659
SVM	1.0	word2vec(U_w)	SWEM-Aver	0.8509	0.6891	**0.8176**	0.7479
CNN	-	word2vec(U_w)	-	0.6262	0.3918	0.6888	0.4995
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 1$)	0.8790	0.7607	**0.8072**	**0.7832**
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 2$)	**0.8848**	0.8130	0.7460	0.7781
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 3$)	**0.8891**	0.8274	0.7460	**0.7846**
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 4$)	0.8797	**0.8519**	0.6729	0.7519
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 5$)	0.8818	**0.8543**	0.6791	0.7567

The combination of weighting word vectors and DFT improved precision, which can be partially explained by the norm of the document vector.

Figures 2,3, and 4 show histograms of the norms of the vectors of inappropriate advertisements for $\Theta = 1, 3, 5$, respectively. When $\Theta > 1$, the distribution of the norm was bimodal, as shown in Fig. 3 and Fig. 4. This phenomenon occured because words with large U_w (particularly nouns and verbs) were emphasized by DFT when they occurred more than once in a document. For example, for $\Theta = 3$, if a word with large U_w, e.g., "pharmaceutical," appears two or three times in a document or if words with similar word vectors, e.g., "doctor" and "physician" appear in a document, the norm of the document vector becomes large. We found that some of the problematic advertising documents have this feature.

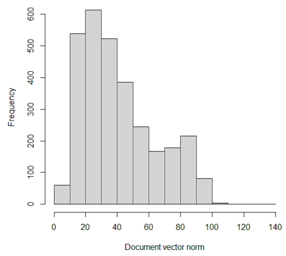

Fig. 2. Histogram of $|\mathbf{x}_D|$ ($\Theta = 1$, illegal documents)

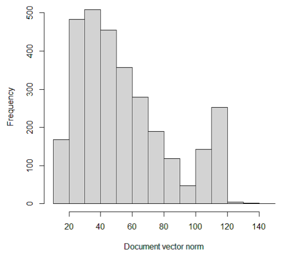

Fig. 3. Histogram of $|\mathbf{x}_D|$ ($\Theta = 3$, illegal documents)

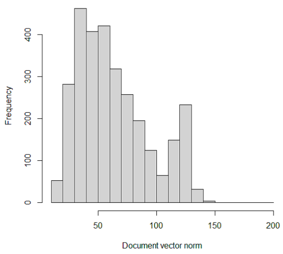

Fig. 4. Histogram of $|\mathbf{x}_D|$ ($\Theta = 5$, illegal documents)

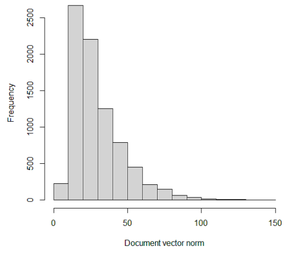

Fig. 5. Histogram of $|\mathbf{x}_D|$ ($\Theta = 1$, legal documents)

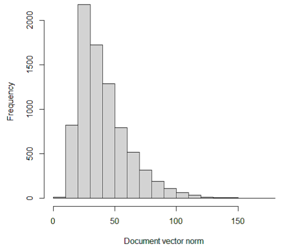

Fig. 6. Histogram of $|\mathbf{x}_D|$ ($\Theta = 3$, legal documents)

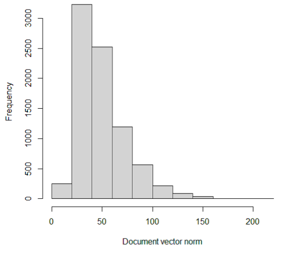

Fig. 7. Histogram of $|\mathbf{x}_D|$ ($\Theta = 5$, legal documents)

As shown in Figs. 5, 6, and 7, the bimodal distribution property was not observed with legal advertisements. Ordinary advertisements do not have word repetitions with large U_w. Thus, the norms of the document vectors do not become large. In addition, when the word vectors were not weighted by U_w, we found that the bimodality property did not appear regardless of the legality of the documents.

As a result of combining weighting by U_w and DFT, the sizes of the norm of the document vectors of the problematic and the normal documents tended to differ. Note that this property is highly effective for discriminant models.

8 Conclusion

In this paper, we have proposed a document vector and discriminant model to discriminate the legality of Japanese advertisement documents for cosmetics. In addition, we evaluated and compared their performance to exsisting models.

In the proposed model, word vectors are weighted by the index of Tang [1], and DFT is embedded in the document vectors. Such document features are utilized effectively in the proposed CV-SVM. The experimental results denmostrate the proposed CV-SVM can provide high generalizability even with limited data; the F-measure value has improved from 0.7479 to 0.7846 compared to the model of Tang et al.

In addition, we have demonstrated that Tang's index has the effect of highlighting nouns and verbs that are likely to appear in problematic Japanese advertisements. By combining this effect with DFT, we can obtain features that are effective for discrimination, and by embedding DFT, the norms of some problematic documents increased. It is likely that the increased norms have some relations with discrimination performance; however, it is necessary to clarify what type of documents have larger norm. In addition, it is also necessary to clarify the causal relationship between the size of the norm and discrimination performance.

Few studies on natural language processing have extended the feature set to complex values. To the best of our knowledge, studies that consider complex numbers are limited to information compression; however, as demonstrated in this paper, by extending document features to complex-valued vectors, it is possible to embed word order information and word periodic occurrence features into document vectors in a simple manner. The result of this study demonstrate that features combining word vector weighting and DFT are effective in terms of discriminating advertisements. We also believe that the proposed method may be useful for general natural language processing tasks.

The ability to create flexible document features with a small computational load is a key feature of complex-valued models, and clarifying tasks for which complex-valued features are effective compared to language models, e.g., BERT, is also a future task.

References

1. Y.Tang, and H.Chen, FAdR: A System for Recognizing False Online Advertisements, *In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL*, pp. 103–108(2014)

2. H.Huang, Y.Wen, and H.Chen, Detection of False Online Advertisements with DCNN, *in Proceedings of the International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee*, pp. 795–796(2017)
3. M.Ma, L.Huang, B.Xiang, and B.Zhou, Dependency-based convolutional neural networks for sentence embedding, *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2, pp. 174–179(2015)
4. Yoon Kim, Convolutional Neural Networks for Sentence Classification, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751
5. Ministry of Health, Labour and Welfare, [Standard for Adequate Advertisement of Pharmaceutical Products]Iyakuhin tou tekisei koukoku kijun(in Japanese), https://www.mhlw.go.jp/file/06-Seisakujouhou-11120000-Iyakushokuhinkyoku/0000179263.pdf
6. Y.Lu, P.Lio, and S.Hand, On low dimensional random projections and similarity search, *In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM ’08*, pp. 749–758(2008)
7. J.Zhang, B.Dong, and S.Philip, Fakedetector: Effective fake news detection with deep diffusive neural network,*In 2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp 1826–1829(2020)
8. S.Kaur, P.Kumar, and P.Kumaraguru, Automating fake news detection system using multilevel voting model, *Soft Computing 24*, pp. 9049–9069 (2020)
9. H.Shinoda, M.Hattori, and M.Kobayashi, [Complex-Valued Support Vector Machine]Hukuso Support Vector Machine(in Japanese), The 73rd national Convention of IPSJ, pp.315 - 316(2011)
10. P.Bouboulis, S.Theodoridis, C.Mavroforakis, and L.Evaggelatou-Dalla, Complex support vector machines for regression and quaternary classification, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, Issue. 6, pp. 1260–1274(2014)
11. D.Shen, G.Wang, W.Wang, M.Min, Q.Su, Y.Zhang, C.Li, R.Henao,and L.Carin, Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 440–450(2018)
12. A.Mahajan, S.Jat, and S.Roy, Feature Selection for Short Text Classification using Wavelet Packet Transform, *Proceedings of the 19th Conference on Computational Language Learning*, pp. 321–326(2015)
13. J.Wieting and D.Kiela, No training required: Exploring random encoders for sentence classification, *arXiv preprint*, arXiv:1901.10444
14. J.Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint*, arXiv:1810.04805, 2018.
15. D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with bi-nary coins, *Journal of Computer and Systems Science*, Special issue of invited papers from PODS ’01, pp. 671–687(2003)