原著論文 Original Paper

離散フーリエ変換文書特徴を用いた，複素SVMによる日本語Web広告文書の適法性判別

Legality Discrimination of Japanese Web Advertisements by Complex-valued SVM using Document Features based on Discrete Fourier Transform

河本 哲 Satoshi Kawamoto
放送大学大学院文化科学研究科, 株式会社アイモバイル
The Graduate School of Arts and Sciences, The Open University of Japan, i-mobile Co.,Ltd.
kawamoto@i-mobile.co.jp, https://www.ouj.ac.jp/, https://www.i-mobile.co.jp/

秋光 淳生 Toshio Akimitsu
放送大学大学院文化科学研究科
The Graduate School of Arts and Sciences, The Open University of Japan
https://www.ouj.ac.jp/

浅井 紀久夫 Kikuo Asai
（同　上）
https://www.ouj.ac.jp/

keywords: discrete fourier transform, natural language processing, internet advertisement, complex-valued SVM

Summary

In Internet advertising, text information is added to increase the appeal of the ad to the viewers. However, some of the advertising documents contain inappropriate expressions. Wording or expressions that exaggerate the efficacy of a product or that recommend a product by a medical professional may violate the Pharmaceutical Affairs Law and the Act against Unjustifiable Premiums and Misleading Representations. Therefore, a system that can effectively and quickly detect problematic advertisements is required. Some advertisements cannot be properly classified based on word statistics alone. Therefore, information other than word statistics must be embedded in the document vector. The advertising documents targeted in this study have characteristics such as “biases in the word positions of specific words” and “periodic occurrence of specific words.” Frequently appearing words in problematic documents (especially in cosmetics advertisements) have strong biases in their word positions, resulting in a complex multimodal distribution of position of occurrence. Therefore, embedding word order information and word period information in document vectors is considered very effective for identifying problematic advertising documents.

In recent years, the effectiveness of the BERT model has been recognized in various natural language processing tasks. However, it is also true that faster models are required for application on the Internet advertising. Therefore, as a means of achieving both inference speed and discrimination performance, we propose a document feature based on the discrete Fourier transform(DFT) of word vectors weighted by an index previously proposed in a study that attempted to categorize Chinese Internet advertisements. In addition, we employed the Complex-valued Support Vector Machines as discriminative models that can handle complex numbers and have high generalization performance even with small amounts of data.

Although the discrimination performance of the proposed model is inferior to that of ALBERT and BERT to some extent, it is higher than that of DistilBERT, XGBoost, and LightGBM. The inference speed of the proposed model is somewhat slower than XGBoost and LightGBM and needs improvement, but is faster than DistilBERT. Those results indicate that the proposed model is promising when applied on the Internet. In addition, we found that when the index proposed in the previous study (which attempted to categorize Chinese advertisements) was applied to Japanese advertisements, that index emphasized the word vectors of specific nouns and verbs.

1. は　じ　め　に

1·1　インターネット広告を取り巻く状況

1996 年からの推計以来，日本におけるインターネット広告費は成長を続け，EC（E コマース）広告の規模も拡大している [電通 21]．テキスト情報が付与されている広告は，商品の魅力が伝わりやすいという優れた側面があるが，広告効果を追い求めるあまり，法律や倫理上不適切な表現を含んだ広告が配信されてしまう危険性がある．広告配信事業者は，広告の審査工程で不適切な広告を除去しているが，市場規模の拡大に伴う審査工程の負担を低減させることが求められている．

また，問題のある文書がインターネット上に配信されてしまったことが契機となり，行政指導・処分などの措置がなされ，広告主に大きな負担が発生する可能性もあ

る．問題のある広告文書を判別できるシステムが提供されることにより，広告主に対する法律上のリスクを低減させることも可能になる．

問題のある文書の判別システムをインターネット広告で応用するには，文書を高速に判別可能であることも重要な要素となる．例えば，広告文書を高速に判断するシステムをインターネット媒体の運営者が保持することにより，不適切な広告露出によるブランド棄損を防ぐことが可能になる．

このように，広告文書の適法性を判別するシステムには高い需要があるが，日本語広告を対象とした研究は盛んであるとはいえず，発展が望まれている．また，人が文書を作成することをシステムが支援する場合，最終的には「なぜこの文書だと問題があるのか」を，出現単語だけではなく，問題のある文章構造の面から説明可能であることも望まれる．

1·2 広告文書データの性質と考案モデル

本研究が対象にする広告文書は，適法・違法のラベルが付与されたデータ数に限りがある．また，Huang[Huang 17] が指摘する通り，文書の適法性を判断するためには法律上の訓練を必要とするため，クラウドソーシングのような手段でアノテーションデータを拡張することも実用的ではない．そのため，少量の学習データであっても未知の文書を判別する能力の高いモデルが求められる．

シンプルでありながら文書の判別に有効な文書特徴量のひとつに，SWEM-Aver [Shen 18] がある．これは文書中の出現単語ベクトルの算術平均を文書特徴量とした，非常に簡単なものであるが，文書の判別タスクにおいて優れた特徴量であることが Shen らによって示されている．しかしながら，本研究の対象とする広告文書は，3·2 節で例示しているように，出現単語の単純な統計情報のみでは判別できない文書も存在する．よって，出現単語の統計情報のみでなく，広告文書の特徴を捉えた特徴量を定義し，かつ適切な判別モデルを用いる必要がある．

BERT[Devlin 18] は自然言語処理の様々なタスクで高い性能を示しており，ファインチューニングすることで，広告文書の特徴を捉えた高性能な判別モデルを作ることができる．但し，インターネット広告への応用を鑑みると，判別性能を維持した上で，より軽量なモデルが望まれる．

本研究では，広告文書において，特定単語の出現位置や出現周期に特徴があることを定量的な議論により明らかにしている．このような特徴を離散フーリエ変換により文書ベクトルに埋め込み，複素 SVM により適法性判別を行うモデルを提案している．また，提案モデルの判別能力は BERT に劣るものの，推論速度では大きく上回ることを示している．

本論文は河本ら [Kawamoto] [Kawamoto 21] による化粧品広告を対象とした広告文書の特徴の調査および適法性判別の研究に加え，健康食品の広告文書についても追加調査を行い，BERT モデルおよび勾配ブースティングモデルとも比較し，定量的な議論を深めたものである．

2. 関 連 研 究

広告文書が適法であるか，あるいはニュース記事の真偽の判別といった，Web コンテンツの文書判別に関する研究は 2010 年代前半から盛んに行われている．

中国語のインターネット広告文書の適法性を判別するモデルとして Tang ら [Tang 14] は，unigram を用いて，SVM にて適法性判別を行う方法を提案した．その際，Tang らは問題のある広告文書内で相対的に出現頻度の高い単語の重みを大きくした文書ベクトルを作ることで Accuracy が向上することを示した．

Zhang ら [Zhang 20] はニューラルネットを用いた特徴抽出および判別モデルを用いたフェイクニュースの検出モデルを提案している．また，Kaur ら [Kaur 20] は TF-IDF, Bag of Words（BOW）など複数の特徴量と SVM, ロジスティック回帰など複数の判別モデルを用いて多数決でニュースの真偽判定を行う方法を提案している．

Huang ら [Huang 17] は，Dependency-based CNN[Ma 15] を用いることで，中国語広告の適法性を判別するモデルを提案している．構文構造を CNN に追加入力することで，単語ベクトルを CNN に入力しただけのものよりも判別性能が向上することを示している．その際，CNN の全体的な構成は Kim のモデル [Kim 14] を用いている．このモデルは，文書データを画像と見立てて文書の特徴を捉えるモデルである．例えば，ある文書の単語数が n であり，単語ベクトルの次元数が m であるとき，Kim のモデルでは文書を $(m \times n)$ のサイズの画像とみなし，畳み込み処理やプーリング処理を行い，特徴抽出をする．Huang らは，Accuracy を評価するだけではなく，Precision, Recall, F 値も評価することで，2017 年時点における CNN モデルの総合的な判別能力の高さを示している．

しかしながら，2018 年に Devlin ら [Devlin 18] が提案した BERT モデルは自然言語処理に関する様々なタスクで高い性能を示しており，現在における自然言語処理の主流のモデルとなっている．BERT は学習済みモデルをファインチューニングすることにより，学習データ量が少量であっても汎化性能の高いモデルを作成することができる．BERT の軽量モデルとして ALBERT[Lan 19] や DistilBERT[Sanh 19] などが提案されており，BERT よりも短い学習時間や推論時間を達成している．

また，BERT の学習コストを低減する手段として，Lee-Thorp ら [Lee-Thorp 21] は，BERT の Self-Attention 層を離散フーリエ変換層に置き換え，実数部分を利用した FNet を提案した．FNet は，Large モデルにおける学習の安定性が BERT よりも高く，BERT と比較して，97% 程の Accuracy を示したことが報告されている．フーリ

エ変換によって，周期的に出現する単語が強調されるため，同単語が繰り返し出現する文書の特徴を効果的に抽出できる．

Mahajan ら [Mahajan 15] は，BOW 表現された文書ベクトルの次元を削減する方法として，ウェーブレット係数を用いることを提案している．文書ベクトルを 1 次元の信号の列とみなしてウェーブレット変換により次元削減を行い，SMS（Short Message Service）のスパム検出タスクにおいて，検出性能が低下しないことを示している．

Chen ら [Chen 16] が考案した XGBoost は，勾配ブースティング決定木を用いることで推論の高速性と高い判別能力を両立させている．

また，Ke ら [Ke 17] は Gradient-based One-Side Sampling (GOSS) と Exclusive Feature Bundling (EFB) を用いた LightGBM を提案し，非常に高速な学習と高い判別能力を実現した．

Wieting[Wieting 19] は，単語ベクトルのシーケンスに乱数行列を掛け，プーリング関数により文書ベクトルを作成するというシンプルな文書ベクトル (BOREP) の性能の高さを示している．

篠田ら [篠田 11] は，SVM を複素数の領域に拡張した複素 SVM（以降 CV-SVM とする）を提案し，UCI machine learning repository[UCI] を用いた複数のタスクで CV-SVM の汎化性能の高さを示した．

Bouboulis ら [Bouboulis 14] は，ウィルティンガーの微分を用いて CV-SVM の双対問題を理論的に導出し，CV-SVM の最適化問題を容易に解けることを示した．

河本ら [Kawamoto] [Kawamoto 21] は，単語ベクトルの重み付けと離散フーリエ変換を組み合わせて広告文書の複素特徴量を作る方法を考案した．化粧品広告文書の複素特徴量を CV-SVM で分類することにより，薬機法上問題のある文書と通常文書を効果的に分類できることを，化粧品広告文書の判別シミュレーションで示した．

3. 広告の適法性の定義および判別の難しさ

3·1 不適切な文書の定義

不適切な広告文書を検出するためには，まず問題のある文書の定義を明確にしておく必要がある．本研究では，薬機法上問題があるかどうかを基準として，広告文書の適切性を判断している．化粧品広告の表現は，薬機法 66 条により規制されており，虚偽・誇大広告に該当する表現を禁止している．また，健康食品の表現範囲は薬機法 68 条にて制限されており，医薬品や医療機器と誤認させるような表現を禁止している．具体的な判断基準は，厚生労働省の医薬品等適正広告基準 [厚生 17] により示されている．以下の項では，問題となる表現を具体例を示しつつ定義する．

§1 効果効能や安全性に関する表現の制限

化粧品の効能の表現可能な範囲については，薬食発（厚生労働省医薬食品局の局長から各都道府県関係部署への通知）0721 号第 1 号にて詳細な具体例が示されている．いくつか例示すると，化粧品の効能により「アンチエイジング効果が得られる」「シワ・たるみの改善」「シミ・そばかすが除去される」「美白・美肌効果が得られる」といった表現を禁止している．化粧品のみならず，医薬品や医薬部外品についても効能や安全性に関する表現には強い制限が課されている．具体的には,「○○年における裏付けから効果がある」などといった歴史的な表現や，臨床データや実験例等を例示することは禁止されている．また「副作用が少ない」といった，効果を保証するような表現も認められない．また，商品の使用感を個人の体験談として表示することは問題が無いが，効果効能や安全性に関する体験談は認められていない．効果効能および安全性に関して,「最高の効き目」「胃腸薬のエース」等の最大級の表現および類する表現も認められていない．

§2 他社製品に対する誹謗広告の制限

化粧品を含む医薬品等の品質，効果効能，安全性等について，他社製品を誹謗する表現は認められていない．但し，誹謗表現の解釈範囲は広く,「他社製品よりもよく効きます」といった比較表現も認められていない．

§3 医薬関係者等の推薦

化粧品を含む医薬品等の広告について，医薬関係者や診療所，大学などの機関が商品を公認したり推薦したりするような広告は禁止されている．これは，事実であっても認められておらず，世人の認識に相当の影響を与えるような広告に関して強い制限が加えられていることに他ならない．また特許に関する表現も，仮に事実であっても認められていない．

§4 化粧品の成分および原材料に関する表現の制限

化粧品の特記表示（商品に配合されている成分中，特に訴求したい成分を目立つように表示すること）に関する制限は，医薬品等適正広告基準に示されている．原材料などを特記表示する場合は,（化粧品に認められた効能の範囲内での）配合目的を併記することが求められている．

§5 健康食品の表現規制

一般的に健康に良いとされる食品を健康食品と呼ぶが，法律上は健康食品の定義は存在せず，商品の特性により「未承認医薬品」あるいは「食品」に分類される．具体的には「医薬品専用の成分を含むか」「治療・予防効果，改善効果等が存在するか」「アンプル，舌下錠などの特殊な形状をしているか」「時間や服用量の指定があるか」といった 4 点を総合的に判断して，医薬品か食品かの判断がなされる．未承認医薬品と判断された場合，薬機法 68 条により「商品の名称」「製造方法」「効果効能および性能」に関する広告が禁止される．食品と判断された場合，広告文書は健康増進法による表現範囲の制限を受ける．

3·2 適法性判別の難しい文書の例

§1 効能を謳っているとまではいえない文書

化粧品広告において，美肌効果を謳うことは認められていないが，実際の広告には判別が難しい文書も存在する．例えば下記の 3 つの文書例は，いずれも美肌効果を明確には主張していない．但し，文書 3) は暗示的な表現ではあるが化粧品の効果により肌がツルツルになることを表現しており，問題のある広告文書である．

1) モチ肌女子の必須アイテム 違いが分かる 1ヵ月。化粧液 1 本の値段でぜんぶ試せる！初回限定〇〇円
2) 化粧かぶれが気になる？ 敏感肌に悩む女性が選んだ、〇〇円からできるスキンケア方法は？
3) 敏感肌専用のコスメが話題 !? 敏感肌でも肌がツルツルになる !? 敏感肌専用のエイジングケア方法が話題

§2 医療関係者が推薦しているとまではいえない文書

化粧品広告において，医療関係者等が商品を推薦するような文章は認められていない．しかし，例えば，文書 4) は，主語が「皮膚科医の妻」であり，皮膚科医が直接的な商品推薦をしているわけではない．また，文書 5) は，医師が「仕事なくなる」と発言しているが，商品を推薦しているとまではいえない．しかし，文書 6) では，医師が効果効能の存在を認めていると解釈できる．よって，文書 6) は不適切な広告文書である．

4) 皮膚科医の妻「毛穴汚れはこれ」簡単すぎて話題に
5) 整形外科医「仕事なくなる」10 秒でぴちぴち肌の裏技とは!?
6) 医師「毛穴レスになりたい人は必見！洗顔前にミストして」

4. 広告文書の特徴および判別に適したモデル

4·1 広告文書の定性的な特徴と単語の重み付けの課題

詳細な議論は 4·3 節にて後述するが，広告文書には，大まかに以下の 3 点の特徴を持つ可能性がある．

A) 問題のある文書内で出現しやすい単語が存在する
B) 文書内の特定単語の出現位置には偏りがある
C) 周期的な出現特性を持つ単語が存在する

3·2 節で示した文書例も，これらの特徴を持つ．例えば文書 3) は「敏感肌」という表現が周期的に出現する．また，文書 6) では，問題のある文書に頻出する「医師」という単語が出現している．

特徴 A) を持つ文書を検出しやすくする方法として，Tang ら [Tang 14] は，式 (1) で示される対数頻度比を用いて，単語の重み付けをすることを提案した．

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

ここで，式 (1) における l_w は，問題のある広告文書セット内で出現した単語 w の数であり，k_w は問題の無い広告文書セット内における w の出現数である．また L は問題のある広告文書セットの延べ単語数（トークン数）であり，K は問題の無い広告文書セットのトークン数である．$l_w = 0$ あるいは $k_w = 0$ のときは $U_w = 0$ とする．

しかし，単語の重み付けだけでは，特定単語の出現位置の偏りや，周期的な出現特性を特徴量として表現できない．本研究では，その課題を解決するため，離散フーリエ変換による複素特徴量を活用し，単語の出現位置情報や周期情報をシンプルに表現した文書ベクトルを考案した．

4·2 特徴量の複素数への拡張

§1 離散フーリエ変換の言語モデルへの適用

ある文書 D が n 個の単語のシーケンスで構成されているとし，単語ベクトルのシーケンスを $(\mathbf{u}_0, \mathbf{u}_1, ..., \mathbf{u}_{n-1})$ とする．ここで，単語ベクトルの次元数を m とすると，図 1 に示すように，文書データを $(m \times n)$ のサイズの画像データと見立てることができる．

このとき，式 (2) で定義される離散フーリエ変換を用いて単語ベクトルのシーケンスを変換する．但し θ は $1 \leq \theta \leq n$ を満たす自然数である．

$$\mathbf{F}(\theta) = \frac{1}{n}\sum_{k=0}^{n-1} \mathbf{u}_k \exp\left(-i\frac{2\pi(\theta-1)k}{n}\right) \tag{2}$$

$\theta = 1$ のときは，単語の出現位置 $k(k = 0, 1, ..., n-1)$ の値によらず位相はゼロとなり，$\mathbf{F}(1)$ は文書中に出現する単語ベクトルの算術平均となる．

また，$\theta = 2$ のときは，文頭 $(k = 0)$ から文末 $(k = n-1)$ に掛けて位相が 1 回転する．すると，文書の前半に出現する単語の位相の範囲は $(-\pi, 0]$ となり，後半に出現する単語の位相は $(-2\pi, -\pi]$ となる．そのため $\mathbf{F}(2)$ は，単語の語順情報が埋め込まれたベクトルとなる．

$\theta = 3$ のときは，文頭から文末に掛けて位相が 2 回転するため，周期 $n/2$ で 2 回出現する単語が強調される．同様に，$\theta = l(l = 4, 5, ..., n-1)$ のときは，文書内で $l-1$ 回出現する単語が強調される．つまり，$\mathbf{F}(3), \mathbf{F}(4), ..., \mathbf{F}(n)$ は，単語の出現周期情報が埋め込まれたベクトルとなる．

よって，$\mathbf{F}(1), \mathbf{F}(2), ..., \mathbf{F}(n)$ を用いることで単語の統計情報，語順情報，周期情報を埋め込んだ文書ベクトルを作成することができる．また，$\mathbf{F}(1), \mathbf{F}(2), ..., \mathbf{F}(n)$ に対してフーリエ逆変換を行うことで，元の単語ベクトルのシーケンスを復元することもできる．また，高周波成分を除去することで，文書判別に必要な本質的な特徴のみを抽出することもできる．図 1 の例では，$\mathbf{F}(4), \mathbf{F}(5), ..., \mathbf{F}(n)$ を除去している．

また，文書ベクトル $\mathbf{x}_D$ は $\mathbf{F}(1), \mathbf{F}(2), ..., \mathbf{F}(\Theta)$ を利用して，式 (3) で得られるものとする．

$$\mathbf{x}_D = \sum_{\theta=1}^{\Theta} \mathbf{F}(\theta)\mathbf{W}_\theta \tag{3}$$

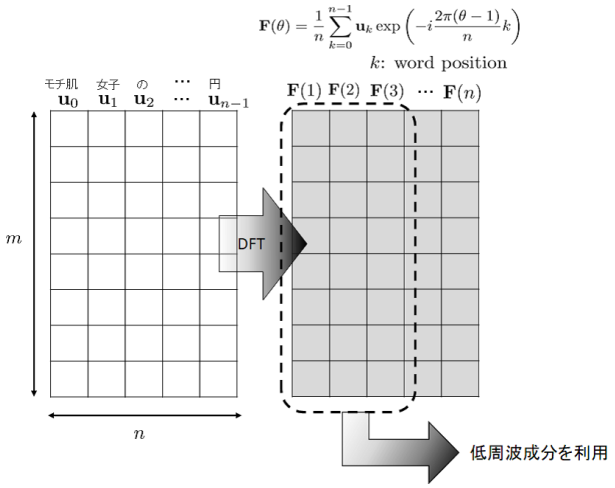


図 1: 単語シーケンスに対する 1-D DFT の適用

但し Θ は解くべきタスクに応じて適切な値を探索する必要がある．また，広告文書のトークン数は 30〜40 語程度のものが多いが，トークン数の少ない短文の広告は $\theta > n$ となる場合がある．このような θ をとる場合は式 (2) を用いる代わりに $\mathbf{F}(\theta) = \mathbf{0}$ とみなして文書ベクトルを作成する．$\mathbf{W}_\theta$ は Sparse Random Projection [Achlioptas 03] によって得られた $(m \times m)$ のサイズの乱数行列である．$\mathbf{W}_\theta$ の a 行 b 列の要素を $W_\theta^{(a,b)}$ とすると，$W_\theta^{(a,b)}$ は式 (4) の通りとなる．

$$W_\theta^{(a,b)} = \begin{cases} -1 & (\text{with probability } \frac{1}{6}) \\ 0 & (\text{with probability } \frac{2}{3}) \\ 1 & (\text{with probability } \frac{1}{6}) \end{cases} \tag{4}$$

Sparse Random Projection は行列の要素の $\frac{2}{3}$ がゼロであるため計算上の負荷が小さく，次元数変換の際の歪みも小さいというメリットを持つ．

本研究では，word2vec（skip-gram）に式 (1) の重み付けをしたベクトルを単語ベクトルとして用いた．その際，単語ベクトルの次元数およびウィンドウサイズを決める必要があるが，最適な数値は一般的に自明でなく，解くタスクにより異なる．Melamud ら [Melamud 16] により行われた，種々のタスクの性能評価によると，単語ベクトルの次元数は 200〜300 次元程度，ウィンドウサイズが 10 程度の値であるとき，良好な結果が得られている．そこで本研究では，ウィンドウサイズを 10 とし，単語ベクトルの次元数を 200 と設定した．単語ベクトルは表 1 に示す 78 581 件の広告文書を用いて作成した．また，動詞や形容詞の活用形を原形に戻す処理は行わず，同じ単語の活用変化は異なる単語ベクトルとした．

§2 複素特徴量を取り扱う判別モデルとしての CV-SVM

離散フーリエ変換によって得られるベクトル $\mathbf{F}(\theta)$ は複素数のベクトルであるため，$\mathbf{F}(\theta)$ を文書判別に用いるためには，複素数を取り扱える判別モデルが必要になる．

そこで，複素数を取り扱うことができ，かつ少量のデータでも高い汎化性能を得られるモデルとして，CV-SVM を用いた判別を行う．CV-SVM の識別関数は $f(\mathbf{x}_D) = \mathbf{w}\phi(\mathbf{x}_D^*) - b$ と表現される．$\mathbf{w}$ は複素数の重みベクトルであり，$\mathbf{x}_D^*$ は文書ベクトル $\mathbf{x}_D$ の各成分が共役になったベクトルである．また $\phi(\mathbf{x})$ は基底関数であり b は複素数のバイアス項である．D が問題のある文書であるときは $\mathrm{Re}(\mathbf{w}\phi(\mathbf{x}_D^*) - b) \geq 1$ および $\mathrm{Im}(\mathbf{w}\phi(\mathbf{x}_D^*) - b) \geq 1$ が満たされるように学習し，問題の無い広告文書であるときは $\mathrm{Re}(\mathbf{w}\phi(\mathbf{x}_D^*) - b) \leq -1$ および $\mathrm{Im}(\mathbf{w}\phi(\mathbf{x}_D^*) - b) \leq -1$ が満たされるように学習する．目的関数 E は式 (5) のように表現され，これを最小化する問題になる．但し，文書セットを Γ とし，α_D, β_D をラグランジュ係数とする．また，文書 D が問題のある広告文書であれば $y_D = 1$ とし，問題の無い文書であれば $y_D = -1$ とする．ξ_D, ζ_D は制約条件の緩和パラメータである．

$$\begin{aligned} E = \frac{1}{2}|\mathbf{w}|^2 &- \sum_{D\in\Gamma} \alpha_D \left\{\mathrm{Re}\left(y_D(\mathbf{w}\phi(\mathbf{x}_D^*) - b)\right) - 1 + \xi_D\right\} \\ &- \sum_{D\in\Gamma} \beta_D \left\{\mathrm{Im}\left(y_D(\mathbf{w}\phi(\mathbf{x}_D^*) - b)\right) - 1 + \zeta_D\right\} \\ &+ C\sum_{D\in\Gamma} \xi_D + C\sum_{D\in\Gamma} \zeta_D \end{aligned} \tag{5}$$

但し，式 (5) を直接的に解くよりも，双対問題を解く方が容易である．CV-SVM の双対問題はウィルティンガーの微分を用いて，$\frac{\partial E}{\partial \mathbf{w}^*}, \frac{\partial E}{\partial b^*}, \frac{\partial E}{\partial \xi_D}, \frac{\partial E}{\partial \zeta_D}$ を求めることで導出可能であることが Bouboulis ら [Bouboulis 14] により示されており，式 (6) のように変形される．

$$\begin{aligned} E = -\frac{1}{2}\sum_{D_1\in\Gamma}\sum_{D_2\in\Gamma} &\psi_{D_1} \cdot \psi_{D_2}^* \cdot y_{D_1} \cdot y_{D_2} \cdot K\left(\mathbf{x}_{D_1}, \mathbf{x}_{D_2}\right) \\ &+ \sum_{D\in\Gamma} (\alpha_D + \beta_D) \end{aligned} \tag{6}$$

但し，D_1, D_2 は広告文書であり $(D_1 \in \Gamma, D_2 \in \Gamma)$，$\psi_{D_1} = \alpha_{D_1} + i\beta_{D_1}, \psi_{D_2} = \alpha_{D_2} + i\beta_{D_2}$ とする．また，制約条件として

$$\sum_{D\in\Gamma} \alpha_D \cdot y_D = \sum_{D\in\Gamma} \beta_D \cdot y_D = 0, 0 \leq \alpha_D, \beta_D \leq C \tag{7}$$

を満たす必要がある．制約条件を満たした上で E を最大化する α_D, β_D を求めることで，識別関数が得られる．最終的な文書の判別であるが，$\mathrm{Re}(f(\mathbf{x}_D))$ と $\mathrm{Im}(f(\mathbf{x}_D))$ の符号が異なっているケースも想定される．そのため予測時は $\mathrm{Re}(f(\mathbf{x}_D)) + \mathrm{Im}(f(\mathbf{x}_D)) \geq 0$ であれば D は問題のある文書であると判定する．また，本研究では，カーネル関数 $K(\mathbf{x}_{D_1}, \mathbf{x}_{D_2})$ は式 (8) の RBF カーネルを用い

表 1: 広告の文書数

総広告文書数	78 581
化粧品 (通常文書)	8 103
化粧品 (薬機法上問題のある文書)	3 008
健康食品 (通常文書)	12 999
健康食品 (薬機法上問題のある文書)	1 487

表 2: 化粧品広告における U_w の大きな単語

単語	U_w	品詞
極限	4.309	名詞
認める	3.884	動詞
ウチ	4.053	名詞
綿棒	3.871	名詞
大学	3.697	名詞
（会社名）	3.648	名詞
家庭	3.583	名詞
誌	3.471	名詞
監修	3.438	名詞
医学	3.401	名詞

表 3: 健康食品広告における U_w の大きな単語

単語	U_w	品詞
医学	4.893	名詞
誌	4.794	名詞
すすめる	4.519	動詞
作り方	4.519	名詞
排便	4.505	名詞
医師	4.359	名詞
掲載	4.118	名詞
歯医者	3.949	名詞
? !?	3.949	名詞
断言	3.949	名詞

ている．

$$K(\mathbf{x}_{D_1},\mathbf{x}_{D_2}) = \exp\left(-\frac{(\mathbf{x}_{D_1}-\mathbf{x}_{D_2})\cdot(\mathbf{x}_{D_1}-\mathbf{x}_{D_2})^*}{\sigma^2}\right) \tag{8}$$

4·3 広告文書の持つ定量的な特性

§1 出現単語の頻度特性

本節では，株式会社アイモバイルから提供された広告文書について，式 (1) における U_w の大きな単語の特徴について述べる．対象となる広告文書は表 1 に示すように，化粧品および健康食品に関する文書データとその他の商材の文書データから構成される．また，化粧品および健康食品広告の文書については，薬機法上問題があるかどうかのラベルが付与されている．なお，正例および負例は薬事法管理者資格の保持者により分類されている．

表 2 は，化粧品広告における U_w の大きな単語であり，表 3 は，健康食品広告における U_w の大きな単語である．化粧品広告文書においては医薬関係者等の推薦表現が認められていないため，医薬に関連する単語は U_w が大きくなる傾向にある．また，未承認医薬品と判断された健康食品は薬機法 68 条の表現規制を受けるため，化粧品広告と同様に，医薬に関連する単語の U_w は大きくなる．

§2 出現品詞の特徴

表 4 は，文書タイプごとの品詞の出現割合を示したものであるが，特定の品詞が問題のある文書で出現しやすいといった特徴は無い．つまり，問題のある広告文書の表現は現代日本語の文法的な制約を逸脱するものではない．

しかしながら，式 (1) で示した U_w の分布を品詞単位でプロットすると，分布に大きな差が生じていることが分かる．図 2，図 3 は，化粧品および健康食品広告文書

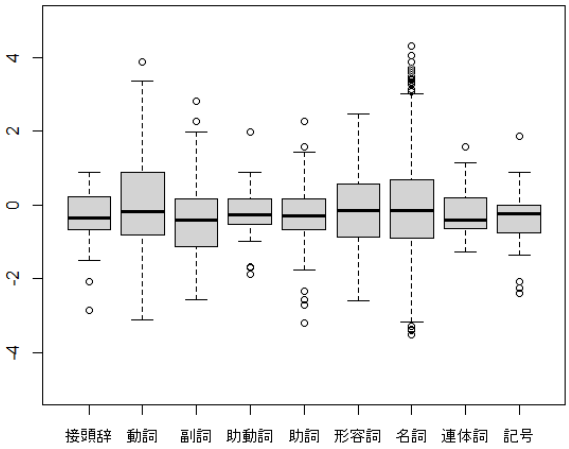


図 2: 各品詞における，出現単語の U_w の分布（化粧品）

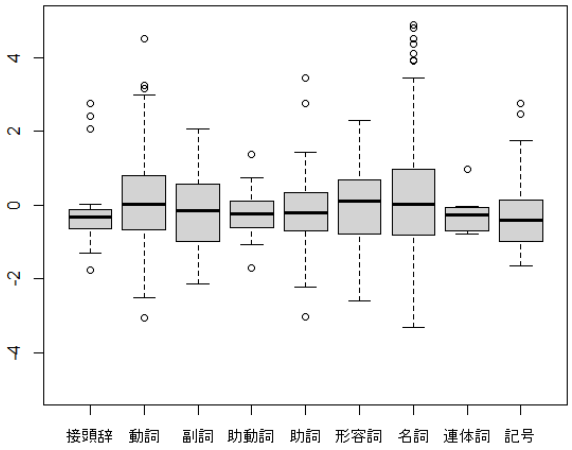


図 3: 各品詞における，出現単語の U_w の分布（健康食品）

における，品詞単位での U_w の分布をプロットしたものである．図 2，図 3 から分かる通り，名詞および動詞は U_w のばらつきが大きく，特に名詞は外れ値も多くなっている．また，接頭辞や助詞および助動詞・記号については，いくつかの外れ値が存在するものの，U_w のばらつきは小さく，広告文書の適法性の影響は小さい．

なお，本研究では，形態素解析には MeCab(ver 0.996) を用いている．また，U_w のばらつきは品詞依存性が強く，U_w の大きな単語は名詞・動詞に集中する．よって，文書中の名詞・動詞を適切に取得するため，辞書データは学校文法に近い品詞体系である IPA 辞書を用いた．

§3 単語の出現位置に関する特徴

広告文書の特徴のひとつとして，特定単語の出現位置の偏りが挙げられる．例えば，図 4(a) は横軸を $k/n\big(k$: 単語の出現位置 $(k=0,1,2,...,n-1)$，n : 文書のトークン数 $\big)$ とし，縦軸は問題のある化粧品広告文書における $U_w > 2.261$（上側 2.5%）となる単語の出現頻度としたヒストグラムであるが，その分布は一様分布とは明らかに異なる．また，図 4(b) は問題の無い化粧品広告における同様のヒストグラムであるが，図 4(a) の分布とは大きく異

表 4: 広告文書の品詞の出現割合

	化粧品 問題あり	化粧品 問題なし	健康食品 問題あり	健康食品 問題なし
助詞	23.272%	23.215%	22.615%	21.322%
助動詞	3.863%	4.391%	4.090%	4.972%
形容詞	1.297%	1.605%	0.907%	1.207%
記号	12.240%	12.980%	12.099%	13.053%
感動詞	0.097%	0.120%	0.023%	0.060%
フィラー	0.015%	0.038%	0.011%	0.025%
接続詞	0.124%	0.145%	0.103%	0.126%
接頭辞	1.126%	1.345%	0.995%	1.068%
動詞	9.503%	10.323%	10.866%	11.732%
副詞	1.712%	2.607%	1.897%	3.147%
連体詞	0.332%	0.545%	0.177%	0.325%
名詞	46.420%	42.685%	46.217%	42.965%
その他	0.000%	0.002%	0.000%	0.000%

なっている．問題のある文書（図 4(a)）では，$U_w > 2.261$ なる単語の出現位置は文書の後半に集中する．対照的に，通常の文書（図 4(b)）では $U_w > 2.261$ となる単語の出現位置の偏りは，やや弱い．なお，図 4(a), 図 4(b) の曲線は，カーネル密度推定（Gaussian カーネル）により得られた確率密度の推定値である．また，曲線のモード数（峰の数）は Silverman の検定 [Silverman 81] により，有意水準を 5%として推定している．Silverman の検定では，帰無仮説 H_0 および対立仮説 H_1 を下記の通りに設定した検定を行う．l を 1 から徐々に増やしながら検定を繰り返し，H_0 を棄却できなくなる値が推定モード数となる．

H_0 : 峰の数は l 個以内である（l : 自然数）

H_1 : 峰の数は l 個より多い

注意すべき点として，Silverman の検定では第二種の過誤が過大になることや，分布の裾で偽のモードが出現しやすいことが指摘されており [楠橋 15]，推定されたモード数は必ずしも正確であるとはいえない．そのため，本研究では，分布が多峰性 ($l \geq 2$) を持つかどうかを検定し，離散フーリエ変換が有効なのかどうかを議論するために Silverman の検定を用いるにとどめる．

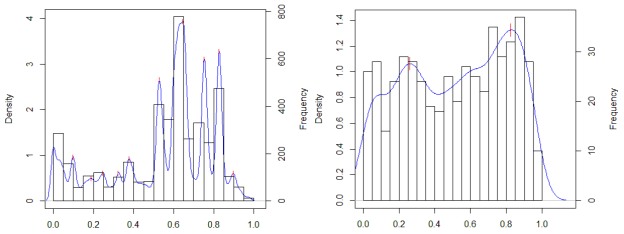
表 5 は化粧品広告文書における，単語の出現位置の多峰性を Silverman の検定で評価したものである．峰の数の正確性を前提とした議論はできないが，問題のある文書において，U_w の大きな単語は出現位置の多峰性が強く，周期的に同一単語が出現している可能性がある．また，表 6 は健康食品広告文書を対象とした同様の表である．問題のある文書では，U_w の大きな単語の出現位置の多峰性が強いが，U_w の小さい単語は単峰性である．問題の無い文書では，U_w の小さい単語に強い多峰性がある．

4·4 離散フーリエ変換後の文書特徴

4·2·1 節にて，式 (2) による離散フーリエ変換を用いた文書特徴量を示した．本節では，広告文書セット Γ の離散フーリエ変換の重心 $\mathbf{G}_\Gamma(\theta)$ の性質を示す．ある文書セット Γ の総文書数を N_Γ とし，文書 $D(\in \Gamma)$ の離散フーリエ変換を $\mathbf{F}_D(\theta)$ としよう．文書セット Γ における $\mathbf{F}_D(\theta)$ の平均値 $\mathbf{G}_\Gamma(\theta)$ を

$$\mathbf{G}_\Gamma(\theta) = \frac{1}{N_\Gamma} \sum_{D \in \Gamma} \mathbf{F}_D(\theta) \tag{9}$$

と定義する．



(a) 問題のある文書　　**(b)** 問題の無い文書

図 4: 化粧品広告文書の単語の出現位置の頻度（$U_w > 2.261$）

表 5: 化粧品広告文書における単語の出現位置の偏り

U_wの範囲	上側%	文書タイプ	峰の数	p 値
$2.261 < U_w \leq 4.305$	$0 \sim 2.5\%$	illegal	11	0.495
$1.791 < U_w \leq 2.261$	$2.5 \sim 5\%$	illegal	5	0.109
$-1.897 < U_w \leq 1.791$	$5 \sim 95\%$	illegal	4	0.149
$-2.260 < U_w \leq -1.897$	$95 \sim 97.5\%$	illegal	1	0.470
$-3.511 \leq U_w \leq -2.260$	$97.5 \sim 100\%$	illegal	2	0.309
$2.261 < U_w \leq 4.305$	$0 \sim 2.5\%$	legal	2	0.556
$1.791 < U_w \leq 2.261$	$2.5 \sim 5\%$	legal	2	0.567
$-1.897 < U_w \leq 1.791$	$5 \sim 95\%$	legal	2	0.337
$-2.260 < U_w \leq -1.897$	$95 \sim 97.5\%$	legal	2	0.349
$-3.511 \leq U_w \leq -2.260$	$97.5 \sim 100\%$	legal	2	0.153

ここで，問題のある文書セット Γ_1 と通常の文書セット Γ_2 における $|\mathbf{G}_{\Gamma_1}(\theta) - \mathbf{G}_{\Gamma_2}(\theta)|$ に注目する．図 5 は，横軸を θ とし，縦軸を化粧品広告および健康食品広告の $|\mathbf{G}_{\Gamma_1}(\theta) - \mathbf{G}_{\Gamma_2}(\theta)|$ としたグラフである．なお，単語ベクトルを式 (1) で重み付けしたグラフを Cosmetics(U_w), Foods(U_w) と表記し，重み付けしていないものを Cosmetics, Foods と表記している．

単語ベクトルに重み付けがなされているとき，健康食品広告は $\theta \leq 35$ 程度の領域で $\mathbf{G}_{\Gamma_1}(\theta)$ と $\mathbf{G}_{\Gamma_2}(\theta)$ の差が比較的大きい．また，化粧品広告は $\theta \leq 3$ 程度の低周波領域および $\theta = 40$ 近傍の領域で $|\mathbf{G}_{\Gamma_1}(\theta) - \mathbf{G}_{\Gamma_2}(\theta)|$ が大きくなっている．

$\mathbf{G}_{\Gamma_1}(\theta)$ と $\mathbf{G}_{\Gamma_2}(\theta)$ との距離が大きいと，問題のある文書セットと通常の文書セットの空間的な分離が明確になり，

表 6: 健康食品広告文書における単語の出現位置の偏り

U_wの範囲	上側%	文書タイプ	峰の数	p 値
$2.770 < U_w \leq 4.897$	$0 \sim 2.5\%$	illegal	5	0.109
$2.300 < U_w \leq 2.770$	$2.5 \sim 5\%$	illegal	5	0.652
$-1.784 < U_w \leq 2.300$	$5 \sim 95\%$	illegal	4	0.624
$-2.066 < U_w \leq -1.784$	$95 \sim 97.5\%$	illegal	1	0.638
$-3.308 \leq U_w \leq -2.066$	$97.5 \sim 100\%$	illegal	1	0.236
$2.770 < U_w \leq 4.897$	$0 \sim 2.5\%$	legal	2	0.121
$2.300 < U_w \leq 2.770$	$2.5 \sim 5\%$	legal	2	0.179
$-1.784 < U_w \leq 2.300$	$5 \sim 95\%$	legal	2	0.325
$-2.066 < U_w \leq -1.784$	$95 \sim 97.5\%$	legal	3	0.325
$-3.308 \leq U_w \leq -2.066$	$97.5 \sim 100\%$	legal	6	0.101

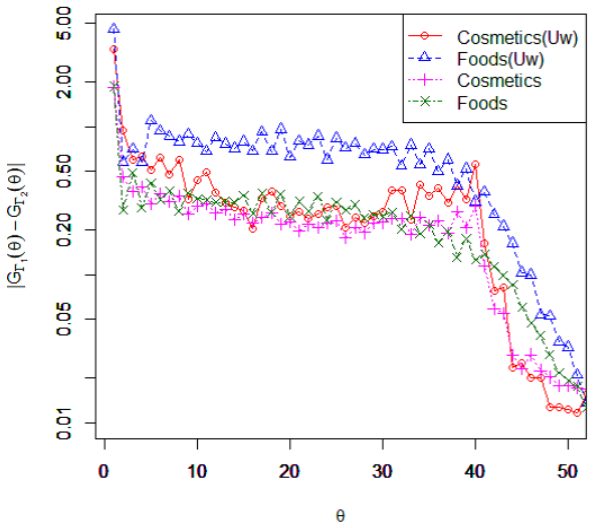


図 5: 化粧品広告および健康食品広告の $|\mathbf{G}_{\Gamma_1}(\theta) - \mathbf{G}_{\Gamma_2}(\theta)|$

CV-SVM による判別が容易になる可能性がある．このことを踏まえて式 (3) の Θ の探索範囲を決定する．図 5 に示される通り，化粧品広告，健康食品広告のいずれも $\theta = 40$ 程度を境に $|\mathbf{G}_{\Gamma_1}(\theta) - \mathbf{G}_{\Gamma_2}(\theta)|$ の低下傾向が強くなる．そこで本研究では，Θ の探索範囲の上限を $\Theta = 40$ とし，探索する Θ の範囲を $\Theta \in \{1,2,3,4,5,6,10,20,30,40\}$ と設定する．

5. 広告文書の判別シミュレーション

5·1 判別モデルの性能指標

広告文書の適法性を効果的に判別するモデルを作るためには，性能を評価するための具体的な指標を定義する必要がある．判別性能の高いモデルの特徴として，問題のある文書を正確に検出しつつ（高い Precision），Recall も高い水準を維持することが望ましい．よって，Precision と Recall の調和平均である F 値を，判別モデルの性能を評価する指標とする．

5·2 ホールドアウト法による判別シミュレーション

シミュレーション対象となるデータには，表 1 における化粧品広告および健康食品広告の文書データを用いた．また，本研究では実装および検証の容易性と計算負荷の低さからホールドアウト法を採用し，シミュレーションを行った．具体的には，表 1 のデータを 2:1:1 の割合に分割し，それぞれ表 7 に示されるような訓練用データ，検証用データ，テストデータとした．すなわち，訓練用データを用いてモデルの学習を行い，検証用データを用いて過学習を防ぎつつ F 値が高くなるパラメータを探索し，実際のモデルの性能はテストデータを用いて評価した．本研究で用いるデータは問題のある文書（正例）よりも通常の文書（負例）の方が多い，偏ったデータである．そのため，訓練データの負例をランダムな非復元抽出によりアンダーサンプリングし，正例数と負例数が同一になるように調整し，学習モデルに与えた．例えば，化粧品広告の判別モデルの学習時には，訓練データの正例は 1 504 件のデータを利用し，負例は 4 052 件のうち 1 504 件をランダムに抽出して利用している．同様に，健康食品広告の訓練用データは正例および負例を 744 件ずつ用いて学習処理を行っている．判別シミュレーションを実施する際，式 (1) による単語ベクトルの重みは訓練データのみを用いて計算するが，訓練データに出現しない単語が検証データおよびテストデータに出現する可能性がある．その場合，未知語 w の重み係数は $U_w = 0$ とする．また，SVM, CV-SVM の C パラメータは $C = 256$ で固定しており，RBF カーネルのパラメータは，$\sigma^2 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ から，グリッドサーチにより決定している．

表 7: 分割後データ数

文書タイプ	訓練データ	検証データ	テストデータ
化粧品 (問題のある文書)	1 504	752	752
化粧品 (通常文書)	4 052	2 026	2 025
健康食品 (問題のある文書)	744	372	371
健康食品 (通常文書)	6 500	3 250	3 249

5·3 比較対象モデルおよび文書ベクトル

4·3 節で議論したように，U_w の大きな単語は，問題のある文書において相対的に出現しやすい．そのため，単語ベクトルを U_w で重み付けすることで判別に有利な特徴量が作成されると見込まれる．そこで本シミュレーションでは，単語ベクトルを U_w で重み付けしたモデルと，重み付けしていないモデルを比較した．

文書ベクトルには SWEM-Aver [Shen 18] および，式 (3) で定義された離散フーリエ変換情報を埋め込んだ文書ベクトルを用いた．高周波成分を用いるかを決定するパラメータ Θ は $\Theta \in \{1,2,3,4,5,6,10,20,30,40\}$ の 10 パターンでシミュレーションを行い，比較した．

判別モデルは，Tang ら [Tang 14] が用いた実数 SVM，Huang ら [Huang 17] が用いた CNN，勾配ブースティング決定木を用いたモデルである XGBoost，LightGBM，および BERT，ALBERT，DistilBERT の学習済み日本語モデルをファインチューニングしたモデル，および本研究で用いる CV-SVM をそれぞれ比較している．BERT の日本語学習済みモデルは東北大学乾研究室が公開しているモデル [東北 19] を用い，ALBERT の日本語学習済みモデルはストックマーク株式会社の公開モデル [スト 20] を用い，DistilBERT の日本語学習済みモデルはバンダイナムコ研究所の公開モデル [バン 20] を用いた．

また，Huang らは単語ベクトルだけではなく，係り受け構造も入力ベクトルに加えることで CNN の判別性能が向上することを示していた．本シミュレーションでは，word2vec で作成した単語ベクトルデータと，日本語 Wikipedia をコーパスとした係り受けに基づく学習済み単語ベクトルデータ（以降 dependency-based と表記）

表 8: シミュレーションパターン（該当項目無しは - 表記）

判別モデル	単語ベクトル	文書ベクトル
SVM	word2vec	SWEM-Aver
CNN	word2vec	-
CNN	dependency-based	-
XGBoost	word2vec	SWEM-Aver
LightGBM	word2vec	SWEM-Aver
CV-SVM	word2vec	SWEM-Aver
CV-SVM	word2vec	DFT（$\Theta \in \{1,2,3,4,5,6,10,20,30,40\}$）
SVM	word2vec(U_w)	SWEM-Aver
CNN	word2vec(U_w)	-
CNN	dependency-based(U_w)	-
XGBoost	word2vec(U_w)	SWEM-Aver
LightGBM	word2vec(U_w)	SWEM-Aver
CV-SVM	word2vec(U_w)	SWEM-Aver
CV-SVM	word2vec(U_w)	DFT（$\Theta \in \{1,2,3,4,5,6,10,20,30,40\}$）
BERT	-	-
ALBERT	-	-
DistilBERT	-	-

表 9: CNN の構成

ユニット	詳細
入力層	200(単語ベクトルの次元数)× 文書の単語数
畳み込み層 (ReLU)	200×3 : 100 チャンネル
	200×4 : 100 チャンネル
	200×5 : 100 チャンネル
プーリング層	Max Pooling
全結合層	活性化関数 : ReLU
出力層	活性化関数 : y=x

[Matsuno 19] の両パターンのシミュレーションを行った．

以上をまとめると，シミュレーションのパターンは表 8 の通りとなる．また，比較対象とした CNN の構成は表 9 の通りである．畳み込み層は 1 層で構成されており，$(200 \times 3), (200 \times 4), (200 \times 5)$ の畳み込みフィルタがそれぞれ 100 チャネルずつ存在する．ストライドは 200 である．畳み込み層を通った後は，各チャネルで Max Pooling 処理が行われ，全結合層へと連結される．ドロップアウト率は 0% としている．学習時に与える正解ラベル y は，問題のある文書には $y = 1$ とし，通常の広告文書には $y = -1$ としている．

5·4 シミュレーション結果

§1 各モデルの判別性能の比較

表 8 のシミュレーションパターンにて，化粧品広告文書判別シミュレーションを行った結果を表 10，健康食品広告文書の判別シミュレーションの結果を表 11 に示した．

化粧品広告文書の判別では，ALBERT および BERT の判別能力が高く，F 値はそれぞれ 0.9038，0.8896 であった．また DistilBERT の Accuracy, Precision は 0.8 を超えるが，Recall が小さく，F 値は 0.6053 程度となっている．CV-SVM の F 値は 0.7989 であり，DistilBERT の F 値を上回るが，ALBERT および BERT モデルの F 値を下回る．また，LightGBM の F 値は 0.7973，XGBoost の F 値は 0.7629 となり，CV-SVM を僅かに下回った．

健康食品文書の判別についても ALBERT や BERT の判別能力は高い．ALBERT の F 値は 0.8954 であり，BERT の F 値は 0.8886 であった．CV-SVM の F 値は 0.8862 と，ALBERT および BERT の判別能力よりも少し劣るが，DistilBERT，LigthGBM，XGBoost の F 値よりも高い．健康食品広告文書の判別において DistilBERT は高い Recall を持つものの，Precision が低めの水準となっている．

§2 各モデルの推論時間の比較

各モデルにおける，1 広告文書あたりの推論時間を表 12 に示した．なお推論時間は，文書を Tokenizer で分割する処理時間と判別モデルにて判別を行う処理時間の合算としている．また，各モデルの推論は Intel(R) Core(TM) i7-9700 CPU 3.00GHz を用いて，CPU モードにて計測した．

BERT および ALBERT は 1 文書の推論に，おおよそ 30ms 程度の時間を要した．DistilBERT の推論時間は，BERT および ALBERT よりも大幅に短縮されている．CV-SVM は実数の SVM よりも長い推論時間を要するが，DistilBERT よりも短い推論時間となっている．LightGBM および XGBoost の推論時間は 2ms 前後であった．

広告配信システムによっては広告表示までの諸処理が 100ms 以内で完了するよう制約される場合もある．よって推論時間が 8ms 未満で完了する点において，CV-SVM には一定の強みがある．しかしながら，LightGBM および XGBoost と比較して，提案モデルの推論時間は長くなっている．

§3 単語の重み付けの効果

全体的な傾向として，式 (1) を用いた単語ベクトルの重み付けにより Accuracy，Precision，Recall，F 値のいずれの指標も凡そ向上する．この傾向は判別モデルが SVM，CNN，CV-SVM，XGBoost，LightGBM において共通であり，簡単な手法ではあるが「問題のある文書で相対的に出現しやすい単語の重みを大きくする」方法が一定の有効性を持つことを示している．

また，CNN に与える入力として係り受けに基づく単語ベクトルを用いたとき，word2vec を用いたときよりも，化粧品広告文書の判別結果の F 値は向上した．しかしながら健康食品広告文書の判別については，Recall が上昇したものの，全体的な判別性能は低下している．CNN モデル全体の傾向として他モデルよりも判別性能は低く，限られた文書数で高い汎化性能を得ることは，CNN では難しいものとなっている．

§4 乱数行列の影響

式 (3) を用いた文書ベクトルを作る際，Θ に設定する値は $\Theta \in \{1,2,3,4,5,6,10,20,30,40\}$ の範囲の任意の数値を選択することができるが，$\Theta = 1$ のときは，単語の出現位置 k が変わっても位相に変化は無い．つまり，文書ベクトル $\mathbf{x}_D$ は「出現単語ベクトルの平均値に乱数行列を掛けたもの」になる．これは BOREP [Wieting 19] のプーリング処理を Average Pooling としたときの特徴量と同一である．SWEM-Aver [Shen 18] との違いは，乱

数行列の有無の差である．

CV-SVM を判別モデルとして用いたとき，SWEM-Aver を特徴量として文書判別を行った結果と，DFT（$\Theta = 1$）を特徴量として文書判別を行った結果を比較しよう．化粧品広告の判別では，単語ベクトルの重み付けをしている条件下で，SWEM-Aver を文書ベクトルとした際の F 値は 0.7669 であり，DFT（$\Theta = 1$）を文書ベクトルとした際の F 値は 0.7583 である．微小な差ではあるが，F 値は低下している．また，健康食品広告を対象とした判別結果も，SWEM-Aver を文書ベクトルとしたときの F 値（0.8717）よりも DFT（$\Theta = 1$）を文書ベクトルとしたときの F 値（0.8415）が下回っている．

Wieting[Wieting 19] は，SentEval[Conneau 18] を用いた特徴量の性能評価で BOE（Bag of Embeddings）よりも，BOREP が有効な特徴量であることを示しているが，本シミュレーションでは，乱数行列を掛けた結果，微小量ながら F 値が低下している．この結果は，単純に乱数行列を掛けることだけが特徴量の性能を向上させることを意味しない．言い換えると，乱数行列を効果的に用いるための前提条件の存在を意味する．

§5 Θ の変化による判別性能への影響

$\Theta = 2$ のときに得られる文書ベクトルは，BOREP に語順情報が追加された特徴量となる．さらに $\Theta = 3$ では，文書内の出現周期が $n/2$（n:文書のトークン数）近傍の単語が強調され，文書ベクトルに埋め込まれる．同様に $\Theta = 4 \sim 6$ では，おおよそ等間隔で 3 ～ 5 回出現する単語を強調した情報が文書ベクトルに埋め込まれる．このように，Θ が増加するにつれ，単語の統計情報・語順情報・周期情報が文書ベクトルに順次埋め込まれていく．

化粧品広告の判別シミュレーションでは，単語ベクトルが U_w で重み付けされている条件下で，$\Theta = 3$ のときに Accuracy および F 値が最も高くなった．Precision と Recall の乖離も小さい．$\Theta \geq 4$ の範囲では，Θ が大きくなるほど判別性能が徐々に低下する傾向が見られた．

健康食品広告の判別シミュレーションでは，単語ベクトルに重み付けを施している条件下で，$\Theta = 10$ であるときに Accuracy および F 値が最も高くなった．このとき，Precision および Recall も高い水準で両立されている．また，表 6 に示される通り，問題のある文書では，$U_w \leq -1.784$ となる単語の出現位置のヒストグラムは単峰性の分布となっている．そのため，$\Theta = 2$ のとき，文書ベクトルへの語順情報の埋め込みにより，False Nagative が少なくなり，問題のある文書の検出漏れが少なくなっている．しかしながら，BERT および ALBERT モデルよりも Recall が比較的小さくなっている．

6. シミュレーション結果の考察

本章では，4·3 節 で触れた広告文書の特徴の側面から，提案手法が広告文書の判別シミュレーションにおいて効果的であった理由を記述する．

まず，Tang の指標による単語ベクトルの重み付けにより，判別結果が改善した理由について記述する．表 5，表 6 に示されている通り，名詞および動詞は U_w のばらつきが大きく，外れ値も多い．次いで形容詞や副詞において，ばらつきが大きい．名詞・動詞・形容詞は内容語であり，文書の意味を構成するために必須となる品詞である．単語の重み付けによって内容語が強調され，不適切な内容の文書を判別しやすい特徴量が得られる．また，副詞による文章の誇張表現も単語ベクトルの重み付けにより強調され，判別性能の向上に寄与した可能性がある．

次に，離散フーリエ変換を用いた文書特徴量を用いることで，高い判別性能が得られた理由について記述する．4·3·3 節 にて示したように，特定の単語は文書内の出現位置に偏りがある．また，U_w の大きさによって多峰性を示したり，単峰性であったりするなどの特性が Silverman の検定にて示されている．ある単語 w が文書中で複数回出現したとしても，出現位置に偏りが無い（文書内で一様に出現する）場合，フーリエ変換をしても有効な特徴量は得られない．しかし出現位置のヒストグラムが多峰性を示すとき，その単語の出現間隔は周期的な特性を持ち，フーリエ変換によって強調される．この強調により，CV-SVM にて効果的に判別可能な文書特徴量が得られる．但し，シミュレーション結果から示される通り，判別対象となる文書次第で，適切な Θ の値は異なる．適切な Θ の値は自明ではないものの，探索によって求めることが可能である．

また，4·4 節にて，問題のある文書セット Γ_1 と通常文書セット Γ_2 の離散フーリエ変換の重心 $\mathbf{G}_{\Gamma_1}(\theta), \mathbf{G}_{\Gamma_2}(\theta)$ に関する議論を行った．健康食品広告については，式 (1) による単語の重み付けと離散フーリエ変換の組み合わせにより，$\mathbf{G}_{\Gamma_1}(\theta)$ と $\mathbf{G}_{\Gamma_2}(\theta)$ との間の距離が十分に確保される．但し，化粧品広告については，$|\mathbf{G}_{\Gamma_1}(\theta) - \mathbf{G}_{\Gamma_2}(\theta)|$ が比較的小さい．そのため，化粧品広告の判別では CV-SVM による識別境界が明確化せず，提案モデルは DistilBERT，XGBoost，LightGBM よりも判別能力が高いものの，BERT や ALBERT よりも判別能力が劣る結果となった可能性がある．しかしながら，$\mathbf{G}_{\Gamma_1}(\theta)$ と $\mathbf{G}_{\Gamma_2}(\theta)$ との距離は，あくまで重心間の距離であるため，化粧品広告において提案モデルの判別性能が頭打ちになった理由を明確化していくことが課題となる．また，4·3 節にて，広告文書の特性に関する分析を行ったが，単語の出現周期特性や出現位置の特性以外に，明確になっていない文書特徴が存在する可能性がある．その特徴を定量的に明らかにしていくことにより，提案モデルの判別性能が向上することが見込まれる．

また，本研究では広告文書データをもとに，ウィンドウサイズを 10 と設定して単語ベクトルを作成したが，適切なコーパスおよびウィンドウサイズを探索することにより，判別性能が向上する可能性もある．

表 10: シミュレーション結果（化粧品広告の分類）

	σ^2	WordVector	DocumentVector	TP	TN	FP	FN	Accuracy	Precision	Recall	F-value
SVM	1.0×10	word2vec	SWEM-Aver	303	1 974	51	449	0.8199	**0.8559**	0.4029	0.5479
CNN	-	word2vec	-	616	859	1 166	136	0.5311	0.3457	0.8191	0.4862
CNN	-	dependency-based	-	484	1 374	651	268	0.6691	0.4264	0.6436	0.5130
XGBoost	-	word2vec	SWEM-Aver	615	1 695	330	137	0.8318	0.6508	0.8178	0.7248
LightGBM	-	word2vec	SWEM-Aver	624	1 729	296	128	0.8473	0.6783	0.8298	0.7464
CV-SVM	1.0×10	word2vec	SWEM-Aver	638	1 753	272	114	0.8610	0.7011	0.8484	0.7677
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 1$)	646	1 608	417	106	0.8117	0.6077	0.8590	0.7118
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 2$)	469	1 931	94	283	0.8642	0.8330	0.6237	0.7133
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 3$)	529	1 846	179	223	0.8552	0.7472	0.7035	0.7247
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 4$)	473	1 899	126	279	0.8542	0.7896	0.6290	0.7002
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 5$)	448	1 854	171	304	0.8290	0.7237	0.5957	0.6535
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 6$)	364	1 942	83	388	0.8304	0.8143	0.4840	0.6072
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 10$)	544	1 811	214	208	0.8480	0.7177	0.7234	0.7205
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 20$)	592	1 622	403	160	0.7973	0.5950	0.7872	0.6777
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 30$)	579	1 597	428	173	0.7836	0.5750	0.7699	0.6583
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 40$)	551	1 593	432	201	0.7721	0.5605	0.7327	0.6352
SVM	1.0	word2vec(U_w)	SWEM-Aver	567	1 807	218	185	0.8549	0.7223	0.7540	0.7378
CNN	-	word2vec(U_w)	-	684	435	1 590	68	0.4030	0.3008	**0.9096**	0.4521
CNN	-	dependency-based(U_w)	-	500	1 359	666	252	0.6694	0.4288	0.6449	0.5214
XGBoost	-	word2vec(U_w)	SWEM-Aver	621	1 770	255	131	0.8610	0.7089	0.8258	0.7629
LightGBM	-	word2vec(U_w)	SWEM-Aver	647	1 801	224	105	0.8815	0.7428	0.8604	0.7973
CV-SVM	1.0×10	word2vec(U_w)	SWEM-Aver	579	1 846	179	173	0.8732	0.7639	0.7699	0.7669
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 1$)	596	1 801	224	156	0.8632	0.7268	0.7926	0.7583
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 2$)	650	1 770	255	102	0.8714	0.7182	0.8644	0.7846
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 3$)	590	1 890	135	162	**0.8931**	0.8138	0.7846	**0.7989**
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 4$)	568	1 874	151	184	0.8794	0.7900	0.7553	0.7723
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 5$)	554	1 753	272	198	0.8308	0.6707	0.7367	0.7022
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 6$)	410	1 925	100	342	0.8408	0.8039	0.5452	0.6498
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 10$)	553	1 856	169	199	0.8675	0.7659	0.7354	0.7503
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 20$)	527	1 817	208	225	0.8441	0.7170	0.7008	0.7088
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 30$)	536	1 740	285	216	0.8196	0.6529	0.7128	0.6815
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 40$)	538	1 686	339	214	0.8009	0.6135	0.7154	0.6605
BERT	-	-	-	713	1 887	138	39	**0.9363**	**0.8378**	**0.9481**	**0.8896**
ALBERT	-	-	-	719	1 904	121	32	**0.9449**	**0.8560**	**0.9574**	**0.9038**
DistilBERT	-	-	-	365	1 936	89	387	0.8286	0.8040	0.4854	0.6053

表 11: シミュレーション結果（健康食品広告の分類）

	σ^2	WordVector	DocumentVector	TP	TN	FP	FN	Accuracy	Precision	Recall	F-value
SVM	1.0×1	word2vec	SWEM-Aver	262	3 078	171	109	0.9227	0.6051	0.7062	0.6517
CNN	-	word2vec	-	221	3 037	212	150	0.9000	0.5104	0.5957	0.5498
CNN	-	dependency-based	-	250	2 652	597	121	0.8017	0.2952	0.6739	0.4105
XGBoost	-	word2vec	SWEM-Aver	320	2 867	382	51	0.8804	0.4558	0.8625	0.5965
LightGBM	-	word2vec	SWEM-Aver	331	2 931	318	40	0.9011	0.5100	0.8922	0.6490
CV-SVM	1.0×10	word2vec	SWEM-Aver	287	3 013	236	84	0.9116	0.5488	0.7736	0.6421
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 1$)	241	3 085	164	130	0.9188	0.5951	0.6496	0.6211
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 2$)	158	3 227	22	213	0.9351	0.8778	0.4259	0.5735
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 3$)	228	3 139	110	143	0.9301	0.6746	0.6146	0.6432
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 4$)	255	3 116	133	116	0.9312	0.6572	0.6873	0.6719
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 5$)	219	3 170	79	152	0.9362	0.7349	0.5903	0.6547
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 6$)	169	3 230	19	202	0.9390	0.8989	0.4555	0.6047
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 10$)	266	3 009	240	105	0.9047	0.5257	0.7170	0.6066
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 20$)	213	3 161	88	158	0.9320	0.7076	0.5741	0.6339
CV-SVM	1.0×10^3	word2vec	DFT($\Theta = 30$)	180	3 184	65	191	0.9293	0.7347	0.4852	0.5844
CV-SVM	1.0×10^4	word2vec	DFT($\Theta = 40$)	219	3 078	171	152	0.9108	0.5615	0.5903	0.5756
SVM	1.0×1	word2vec(U_w)	SWEM-Aver	321	3 159	90	50	0.9613	0.7810	0.8652	0.8210
CNN	-	word2vec(U_w)	-	203	3 108	141	168	0.9146	0.5901	0.5472	0.5678
CNN	-	dependency-based(U_w)	-	241	2 709	540	130	0.8149	0.3086	0.6496	0.4184
XGBoost	-	word2vec(U_w)	SWEM-Aver	333	3 165	84	38	0.9663	0.7986	0.8976	0.8452
LightGBM	-	word2vec(U_w)	SWEM-Aver	343	3 137	112	28	0.9613	0.7538	0.9245	0.8305
CV-SVM	1.0×10	word2vec(U_w)	SWEM-Aver	326	3 198	51	45	0.9735	0.8647	0.8787	0.8717
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 1$)	292	3 218	31	79	0.9696	**0.9040**	0.7871	0.8415
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 2$)	329	3 167	82	42	0.9657	0.8005	0.8868	0.8414
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 3$)	298	3 210	39	73	0.9691	0.8843	0.8032	0.8418
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 4$)	313	3 124	125	58	0.9494	0.7146	0.8437	0.7738
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 5$)	325	3 204	45	46	0.9749	0.8784	0.8760	0.8772
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 6$)	297	3 214	35	74	0.9699	**0.8946**	0.8005	0.8450
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 10$)	327	3 209	40	44	**0.9768**	0.8910	0.8814	**0.8862**
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 20$)	301	3 220	29	70	0.9727	**0.9121**	0.8113	0.8588
CV-SVM	1.0×10^3	word2vec(U_w)	DFT($\Theta = 30$)	284	3 203	46	87	0.9633	0.8606	0.7655	0.8103
CV-SVM	1.0×10^4	word2vec(U_w)	DFT($\Theta = 40$)	300	3 129	120	71	0.9472	0.7143	0.8086	0.7585
BERT	-	-	-	367	3 161	88	4	**0.9746**	0.8066	**0.9892**	**0.8886**
ALBERT	-	-	-	364	3 171	78	7	**0.9765**	0.8235	**0.9811**	**0.8954**
DistilBERT	-	-	-	363	3 051	198	8	0.9431	0.6471	**0.9784**	0.7790

表 12: 各モデルの 1 広告文書あたりの推論時間

判別モデル	文書タイプ	推論時間 (ms)	パラメータ
SVM	化粧品広告	3.010	$\sigma = 1.0$
CNN	化粧品広告	41.630	-
XGBoost	化粧品広告	1.964	-
LightGBM	化粧品広告	2.029	-
CV-SVM	化粧品広告	7.658	$\sigma = 1.0 \times 10^3, \Theta = 3$
BERT	化粧品広告	29.291	-
ALBERT	化粧品広告	31.042	-
DistilBERT	化粧品広告	15.712	-
SVM	健康食品広告	1.630	$\sigma = 1.0$
CNN	健康食品広告	41.148	-
XGBoost	健康食品広告	2.076	-
LightGBM	健康食品広告	2.078	-
CV-SVM	健康食品広告	5.377	$\sigma = 1.0 \times 10^3, \Theta = 10$
BERT	健康食品広告	30.519	-
ALBERT	健康食品広告	29.402	-
DistilBERT	健康食品広告	17.797	-

7. ま と め

本研究では，化粧品および健康食品の日本語広告文書の適法性を判別するための文書ベクトルおよび判別モデルを提案し，その性能評価と他モデルとの比較を行った．

具体的には，単語ベクトルに対して Tang[Tang 14] の指標による重み付けを施し，かつ離散フーリエ変換情報を文書ベクトルに埋め込むことで，文書判別に有効な特徴量を得た．また，判別モデルとして CV-SVM を用いることで，高速な推論と高い判別能力が両立されることを示した．CV-SVM の判別能力は BERT，ALBERT を下回るものの，処理速度では BERT 系モデルを大きく上回った．しかしながら処理速度については XGBoost，LightGBM を下回り，全体的な傾向として，モデルの判別能力と処理速度との間にトレードオフの関係が成り立った．

提案モデルは，健康食品広告の判別については，BERT モデルとほぼ同等の判別能力を維持したまま大幅な高速化が成されている．しかし，化粧品広告の判別能力は BERT モデルを下回る．また，構造化された特徴を文書ベクトルに埋め込むことにより，SWEM-Aver を文書ベクトルとしたときよりも判別能力は向上したが，判別対象の文書タイプに応じた適切な構造は明らかになっておらず，適切な構造発見が今後の課題となる．

自然言語処理に関して，複素数を活用した研究は多くない．Mahajan [Mahajan 15] は，ウェーブレット係数を用いた特徴選択の方法を提案したが，その目的は次元削減であり，判別モデルに与える特徴量は実数を前提としている．FNet [Lee-Thorp 21] は，BERT の Self-Attention 層をフーリエ変換層に置き換えたモデルであり，学習の安定性も示されている．しかし，フーリエ変換後の特徴量の虚数成分は除去されており，虚数成分の有効性に関する議論は行われていない．しかしながら本研究にて示した通り，文書特徴量を複素数領域に拡張することによって，シンプルな方法で語順情報や単語の周期的な出現特徴を文書ベクトルに埋め込むことが可能になる．本研究では，単語ベクトルの重み付けと離散フーリエ変換を組み合わせた特徴量が，広告文書の判別において一定の有効性を持つことを示した．また少ない計算負荷で，柔軟な文書特徴量を作成することができることが複素モデルの大きな特徴であるが，効果的に判別可能な文書タイプあるいは判別が困難な文書タイプの詳細な特性は現状明らかではなく，今後明確にしていくことが課題である．また，BERT モデルなど他の言語モデルと比較して，どのようなタスクで複素特徴量が有効に働くかを明確化していくことも課題である．

◇ 参 考 文 献 ◇

[Achlioptas 03] Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins, *Journal of Computer and Systems Science*, Vol. 66, No. 4, pp. 671–687 (2003)

[Bouboulis 14] Bouboulis, P., Theodoridis, S., Mavroforakis, C., and Evaggelatou-Dalla, L.: Complex support vector machines for regression and quaternary classification, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 6, pp. 1260–1274 (2014)

[Chen 16] Chen, T. and Guestrin, C.: XGBoost: a scalable tree boosting system, in *Proceedings of KDD 2016 the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM: 2016*, pp. 785–794 (2016)

[Conneau 18] Conneau, A. and Kiela, D.: Senteval: An evaluation toolkit for universal sentence representations, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018)

[Devlin 18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding, in *arXiv preprint, arXiv:1810.04805* (2018)

[Huang 17] Huang, H., Wen, Y., and Chen, H.: 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)

[Kaur 20] Kaur, S., Kumar, P., and Kumaraguru, P.: Automating fake news detection system using multilevel voting model, *Soft Computing*, Vol. 24, No. 12, pp. 9049–9069 (2020)

[Kawamoto] Kawamoto, S., Akimitsu, T., and Asai, K.: Legality identification of Japanese online advertisements using complex-valued support vector machines with DFT-coded document vectors, in *New Frontiers in Artificial Intelligence, Lecture Notes in Artificial Intelligence (LNAI)*: (accepted)

[Kawamoto 21] Kawamoto, S., Akimitsu, T., and Asai, K.: Identifying legality of Japanese online advertisements using complex-valued support vector machine with DFT-based document features, in *Proceedings of International Workshop: Artificial Intelligence of and for Business (AI-Biz2021) associated with JSAI International Symposia on AI 2021 (IsAI-2021)* (2021)

[Ke 17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.: LightGBM: A highly efficient gradient boosting decision tree, in *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017*, pp. 3147–3155 (2017)

[Kim 14] Kim, Y.: Convolutional neural networks for sentence classification, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751 (2014)

[Lan 19] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations, in *arXiv preprint, arXiv:1909.11942* (2019)

[Lee-Thorp 21] Lee-Thorp, J., Ainslie, J., Eckstein, I., and Ontanon, S.: FNet: mixing tokens with fourier transforms, in *arXiv preprint, arXiv:2105.03824* (2021)

[Ma 15] Ma, M., Huang, L., Xiang, B., and Zhou, B.: 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*, pp. 174–179 (2015)
[Mahajan 15] Mahajan, A., Jat, S., and Roy, S.: Feature selection for short text classification using wavelet packet transform, in *Proceedings of the 19th Conference on Computational Language Learning*, pp. 321–326 (2015)
[Matsuno 19] Matsuno, T.: Dependency-based Japanese Word Embeddings, https://github.com/lapras-inc/dependency-based-japanese-word-embeddings (2019)
[Melamud 16] Melamud, O., McClosky, D., Patwardhan, S., and Bansal, M.: The role of context types and dimensionality in learning word embeddings, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016)
[Sanh 19] Sanh, V., Debut, L., Chaumond, J., and Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in *arXiv preprint, arXiv:1910.01108* (2019)
[Shen 18] Shen, D., Wang, G., Wang, W., Min, M., Su, Q., Zhang, Y., Li, C., Henao, R., and Carin, L.: Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 440–450 (2018)
[Silverman 81] Silverman, B.: Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society. Series B(Methodological)*, Vol. 43, No. 1, pp. 97–99 (1981)
[Tang 14] Tang, Y. and Chen, H.: 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)
[UCI] UCI, : UCI machine learning repository, http://archive.ics.uci.edu/ml
[Wieting 19] Wieting, J. and Kiela, D.: No training required: Exploring random encoders for sentence classification, in *arXiv preprint, arXiv:1901.10444* (2019)
[Zhang 20] Zhang, J., Dong, B., and Philip, S.: Fakedetector: Effective fake news detection with deep diffusive neural network, in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1826–1829 (2020)
[スト 20] ストックマーク：大規模日本語ビジネスニュースコーパスを学習した ALBERT モデルを公開！, https://stockmark.co.jp/news/2020-02-18-3426 (2020)
[バン 20] バンダイナムコ研究所：Japanese DistilBERT Pretrained Model, https://github.com/BandaiNamcoResearchInc/DistilBERT-base-jp (2020)
[厚生 17] 厚生労働省：医薬品等適正広告基準, https://www.mhlw.go.jp/file/06-Seisakujouhou-11120000-Iyakushokuhinkyoku/0000179263.pdf (2017)
[篠田 11] 篠田 北斗, 服部 元信, 小林 正樹：複素サポートベクターマシン, 情報処理学会第 73 回全国大会, pp. 315–316 (2011)
[電通 21] 電通：2020 年日本の広告費, https://www.dentsu.co.jp/news/release/2021/0225-010340.html (2021)
[東北 19] 東北大学乾研究室：Pretrained Japanese BERT models released / 日本語 BERT モデル公開, https://github.com/cl-tohoku/bert-japanese (2019)
[楠橋 15] 楠橋 直, 岡本 隆：ノンパラメトリックな多峰性検定-Silverman の検定-とその古生物学への導入, 化石, Vol. 97, pp. 23–37 (2015)

〔担当委員：吉田 光男〕

2022 年 5 月 12 日 受理

著 者 紹 介

河本 哲（正会員）

株式会社アイモバイル技術本部所属．広告配信機能の開発および広告効果の最適化に従事．2005 年東京大学工学部電気工学科卒業．2007 年東京大学大学院新領域創成科学研究科先端エネルギー工学専攻修士課程修了．2018 年放送大学大学院文化科学研究科修士課程情報学プログラム修了．2020 年より放送大学大学院文化科学研究科博士後期課程情報学プログラムに在学中．

秋光 淳生

放送大学教養学部准教授．博士 (工学)．2000 年東京大学大学院工学研究科博士課程退学．同年東京大学先端科学技術研究センター助手．2002 年東京大学大学院新領域創成科学研究科助手．2004 年東京大学大学院工学系研究科電気工学専攻助手．2007 年放送大学准教授．生涯学習における数理情報技術の活用に関する研究に従事．神経回路学会, 情報処理学会, 教育工学会各会員

浅井 紀久夫（正会員）

放送大学教養学部教授．博士（工学）．1996 年名古屋大学大学院工学研究科博士後期課程退学．同年大学共同利用機関放送教育開発センター助手，2000 年メディア教育開発センター助教授，2009 年放送大学准教授．この間，イリノイ大学シカゴ校，アルバータ大学，カンタベリー大学客員研究員．知識情報処理，ヒューマンコンピュータインタラクションに関する研究に従事．電子情報通信学会，電気学会，ACM 各会員．