Vol.2022-IFAT-147 No.2
2022/7/29

ゲームソフトの評価レビューに対するマルチラベル分類におけるSVMとBERTの比較

岸田 和明[1,a)]　伊藤 夕希也[2]　門脇 夏紀[3]

概要：本研究では，ゲームソフトの評価レビューを観点別に自動分類することを試みる。観点としては，「戦闘システム」「物語」「サウンド・グラフィックス」を取り上げ，評価レビューの集合の中からそれぞれ「肯定」「否定」しているものを自動検出することが最終的な目的である。この際，1 つの評価レビューが複数の観点に言及している可能性があり，したがって，これはマルチラベル分類の問題である。今回はそのための方法として，SVM に基づく Binary relevance 法と Binary cross entropy による BERT に焦点を当てる。実際に，Amazon.com から収集した 600 件の評価レビューを使って，それぞれの方法の性能を確認する。SVM に関しては，自作辞書による特徴抽出の効果の検証も行う。

1. はじめに

新型コロナウイルス COVID-19 の感染拡大により，多くの業界が苦境に陥った一方で，ゲームソフト業界はその売り上げを伸ばした。その一因は，外出しづらい状況での「巣ごもり需要」にあるのかもしれない。いずれにせよ，専用ゲームマシンあるいはパソコンに加えて，タブレットやスマートフォンがそのための機器として加わり，数多くのゲームソフトが有料または無料で利用可能となっている。

そのような状況において新たに始めるゲームソフトを探す際には，評価レビューが欠かせない。専門家や一般利用者によるレビューが重要な探索手段として機能することは，小説や映画あるいはその他の商品でも同様である。例えばレビュー中での「良い／悪い」「面白い／面白くない」といった記述は，選択のための貴重な判断材料となる。実際，ゲームソフトの場合，ウェブ上で閲覧できるレビューのサイトとして Amazon.com，価格.com，4Gamer.net などがあり，広く活用されていると推測される。

評価レビューの数もまた膨大なので，その閲覧の際には，何らかのタグを使って関連レビューを絞り込めれば便利である。実際，Amazon.com ではゲームによっては「操作方法」や「オンライン対戦」のようなタグを使って特定のトピックを限定できる。さらに進んで，意見分析（または極性分析）を応用して，「戦闘（バトル）の仕組みが良い」「物語が素晴らしい」といった指摘が含まれているレビューを自動的に検出するシステムがあれば便利かもしれない。

つまり，1 件のレビューを単に「高評価」「低評価」に仕分けるのではなく，「戦闘システム」「物語（ストーリー）」「サウンド・グラフィックス」といった観点（viewpoint）別での肯定と否定を割り当てる分類器を用意すれば，より一層高度な絞り込み機能が提供可能となる。例えば物語を重視する利用者は，この分類器によってそれを「肯定」と評価しているレビューを特定し，その中からゲームソフトの候補を選ぶことができる。また，特定ソフトの物語についての評価が知りたい時には，まずはソフト名でレビューを検索してから，その結果集合を分類器で「肯定」「否定」「物語とは無関係」に三分割すれば，その閲覧が容易となる。

1 件のレビューが複数の観点に対する肯定あるいは否定に言及する可能性があるので，ここでの問題はマルチラベル分類として定式化される。例えば，「戦闘システム」と「物語」について「(1) 戦闘：肯定」「(2) 戦闘：否定」「(3) 物語：肯定」「(4) 物語：否定」の 4 つのカテゴリを設定すれば，「戦闘は低評価，物語は高評価」のレビューのラベルは $[0, 1, 1, 0]$ となる。

これは，標準的な「テキストデータに対するマルチラベル分類」である。本研究の目的は，この問題に対して既に考案されているサポートベクタマシン（SVM）を使った仕組みを，若干の工夫を加えた上で適用してみることにある。さらには，BERT でもマルチラベル分類が可能であり [25]，その結果を SVM によるものと比較する。以下，2 節では関連研究を概観し，3 節にて本研究でのマルチラベル分類

[1] 慶應義塾大学文学部図書館・情報学専攻
Faculty of Letters, Keio University, Minato-ku, Tokyo 108–8345, Japan
[2] 2022 年 3 月まで慶應義塾大学文学部
[3] 慶應義塾大学大学院文学研究科後期博士課程
Graduate School of Letters, Keio University
[a)] kz_kishida@keio.jp

の仕組みと実験データを説明する。それに対する実験結果は 4 節で述べる。

2. 関連研究

本節では，まずゲームソフトの推薦に関連する研究を概観し，次にマルチラベル分類の技法について確認する。

2.1 ゲームソフトの推薦システム

COVID-19 の感染拡大以前からゲームの市場は年々拡大しており，様々なゲームソフトが販売・提供されるとともに，それに対する愛好者のニーズもまた多様化している。多種多様なゲームソフトの中から自分の好みに合ったものを選ぶには，他の商品やサービスと同様に，情報推薦が有力な手段である。実際に，インターネットで商品購入を可能とするサイトには推薦システムが組み込まれ，ゲームソフトについても協調フィルタリングなどのアルゴリズムによる推薦がなされる。同時に，ゲームソフトに関するその種の推薦を向上させるための研究も試みられている。

中谷ら (2008)[23] はゲームをプレイした経験から得られる経験的価値を感覚的知覚，創造的思考，身体性，社会性，衝動的感情の 5 つに分類した。その上で，レビューに含まれる経験的価値と利用者の経験に対する嗜好が記録されているユーザープロファイルとの類似度を計算することによって推薦するゲームを決めている。

大山ら (2017)[26] は利用者の検索での要望と合致したゲームの情報を推薦するシステムの設計と実装を行った。この研究では，利用者のゲーム体験で得られる経験的価値を多く含んでいるレビュー文を訓練用コーパスとし，「言葉の足し引き」による意味処理とゲームのベクトル化が試みられている。この「言葉の足し引き」では，例えば「謎解き要素を含んだアクションゲームをプレイしたい」という要望が「謎解き+アクション」と表現される。具体的にはこのために Word2vec が応用されている。

2.2 テキストデータに対するマルチラベル分類

レビューやブログ記事などのテキストデータに対する自動分類の研究・試みには枚挙のいとまがない。ここでは，特にマルチラベル分類に焦点を当て，その方法を整理する。なお，マルチラベル分類の技術は Zhang & Zhou(2014)[45] により全般的にレビューされているが，このレビューの後，ニューラルネットワークの応用が数多くなされているのが現状である。

2.2.1 マルチラベル分類手法の類型

本稿ではラベルの集合を $\mathcal{Y} = \{y_1, \ldots, y_L\}$ と表記する。テキストに対するマルチラベル分類は，ある文書のテキストデータ x に対して $\mathcal{Y}$ の部分集合 Y（$\subseteq \mathcal{Y}$）を対応付けることに相当する。この問題に対する初期的な試みとしては，確率分布の混合モデルの適用 [21] や Boosting の応用 [30] が挙げられる。既存の分類器をマルチラベル分類用に修正する場合にはアルゴリズム適応法（Algorithm adaptation method）と呼ばれ，データを変形してシングルラベル分類用の手法をそのまま利用する問題変形法（Problem transformation method）と区別される [34]。この「問題変形法」「アルゴリズム適応法」の区分はよく使われており（例えば文献 [45]），場合によってはそこにアンサンブル法（Ensemble method）を加えることがある [19]。

2.2.2 BR 法とペアワイズ法

問題変形法の中で最も単純な方法が BR 法（Binary relevance method）であり，数多くの研究が試みられている（Zhang et al., 2018[42] 参照）ほか，実験にてベースラインとして使われることも多い。この場合には，L 個のラベルそれぞれに対して 2 値分類器を設定し，「1 対残り（one-to-the-rest）」で学習する。新規文書に対しては，それを L 個の分類器に投入し，「+1」が出力された分類器に対応したラベルを付与することになる。実際，k-NN，C4.5，ナイーブベイズ，SVM による BR 法が試されている [34]。一方，「1 対残り」ではなく「1 対 1（one-to-one）」で学習する場合はペアワイズ法（Pairwise method）である（例えば文献 [15] など）。

2.2.3 BR 法の改良：ラベル相関の考慮

BR 法を改良する試みは数多く，Classifier chain（CC）[37] はその代表例である。CC では，y_k についての分類器への入力に，当該文書のテキストデータ x に加えてそれ以前のラベル $y_1, \ldots, y_{k-1}$ の分類結果を含める。これにより，ラベル間の相関を考慮したマルチラベル分類が可能になる。単純な CC はラベルの順序に依存するため，さらに 1 つの CC を弱学習器としたアンサンブル学習（ECC: Ensembles of classifier chains）も試みられている [37]。また CC に対しては，条件付き確率やベイジアンネットワークによる修正もなされている（文献 [31] など）。

ラベル相関を組み込む方法には，この種の「chaining」の他にスタッキング（stacking）がある。これは，標準的に BR 法を適用して L 個の分類器を学習した後に，各文書に対するそれらの分類器の予測結果を入力に付加した新たな分類器を構成する方法である [11]。一方，Dependent binary relevance（DBR）モデル [22] では，スタッキングにおける第 1 段階での分類器を構成せず，正解ラベル自体を特徴として組み込む（新規文書には別の分類器を用意）。また，ラベル相関を考慮するには，共起頻度などから計算されるファイ係数などの指標を使って各ラベルと相関の高いラベルのみを BR 法に組み込むことが考えられ，実際にいくつかの試みがある（文献 [41] など）。

2.2.4 ラベルのべき集合の要素を採用する方法

ラベルの集合 $\mathcal{Y}$ のべき集合 $2^{\mathcal{Y}}$ の要素それぞれを 1 つのラベルと見なせば，多クラスでのシングルラベル分類に帰着する [3]。ここでの $2^{\mathcal{Y}}$ は LP（Label powerset）と呼ばれ

る。ただしその際，LP 中の要素数が膨大になるかもしれず，そしてそのすべてが正解ラベルとして訓練データに含まれるとは限らないという問題がある。

RAkEL (Random k-labelsets）[35] はこの問題を解決するための一種のアンサンブル学習で，$\mathcal{Y}$ を分割し，それぞれに対して LP を個別に構成する方法である。RAkEL の改良版として RAkEL++[29] や CP-RAkEL[39] も考案されている。また，HOMER（Hierarchy of Multilabel classifers）[36] も LP の要素数が多い場合の効率化の仕組みで，クラスタリングの手法で LP の要素自体を階層的に構造化することにより分類器の数を減少させる工夫である。

2.2.5 アルゴリズム適応法の例

アルゴリズム適応法の場合には，問題変形法とは異なり分類器自体にマルチラベル分類のための工夫が組み込まれる。例えば，Rank-SVM[7] は，本来は 2 値分類である SVM をマルチラベル分類用に拡張したものである。なお，Rank-SVM では新規文書に対して各ラベルの得点が出力されるのみなので，別の方法で閾値を設定して最終的なラベルを決めなければならない。SVM と同様に判別分析もまたマルチラベル分類に応用されており，カーネル判別分析を使った試み [33] などがある。

Boosting もまた早くからマルチラベル分類に応用され，AdaBoost.MH と AdaBoost.MR がその代表である [30]。この場合には 1 つの語に対して 1 個の弱学習器が設定されるが，大規模な文書集合では計算量が多くなるため，効率化のための工夫がいくつか考案されている [8][1]。

k-NN については，マルチラベル分類用の ML-kNN[44] がよく知られている。k-NN は訓練データ中の実例（インスタンス）に基づく方法であるが，ロジスティック回帰によりこれをモデルによる学習とを組み合わせた IBLR-ML (Instance-based logistic regression for multi-label classification)[4] や BR 法として k-NN を実行する際に計算上の工夫を加えた BRkNN[32] もある。

決定木をマルチラベル分類に適応させた例として ML-C4.5[5] がある。また ML-C4.5 を弱学習器としたランダムフォレストが RFML-C4.5 である [19]。一方，文書集合に対する分割型クラスタリングにより得られた木構造である PCT（Predictive clustering tree）[2] も利用されており，PCT でのアンサンブル学習は RF-PCT[14] と呼ばれる。なお，決定木では各ノードでの分岐におけるルールが推計される点に特徴があるが，マルチラベル分類におけるこの種のルールを自動構築する試みもなされている [20]。

2.2.6 ニューラルネットワークの応用

ニューラルネットワークをマルチラベル分類に応用した初期的な例として，FNN に基づく Backpropagation for multilabel learning（BP-MLL）[43] がある。BP-MLL に対しては，ラベル付与のための閾値推計の組み込み [12]，ReLU や Cross entropy（損失関数）の利用 [24] などの改良が試みられている。

その後，Word2vec などの語の埋め込み，CNN や RNN 等の技術がマルチラベル分類に応用された。例えば，XML-CNN[18] は，CNN に対してマルチラベル分類用の調整を加えたものである。なお，ラベルの数が膨大な場合のマルチラベル分類を Extreme multi-label text classification (XMTC) と称することがあり，ここでの「XML」は「Extreme multi-label」の略である。XML-CNN では損失関数としては BCE（Binary cross-entropy）が採用され, 語の埋め込みには GloVe が使われている。また，双方向 RNN[6] の応用や，複数のニューラルネットワークによるアンサンブル学習 [16][13] も試みられている。

Transformer 系の BERT をマルチラベル分類に応用したモデルとして BERTMeSH[40] がある。MeSH の自動付与についての研究の歴史は長く，さらに BioASQ チャレンジ（http://bioasq.org/）の中で MeSH の自動付与が取り上げられた時期があり，そこでも様々な技術が試された。比較的最近のものとしては，MeSHLabeler，DeepMeSH，MeSH Now，AttentionMeSH，MeSHProbeNet，FullMeSH がある（各出典は文献 [40] を参照）。この中で AttentionMeSH と MeSHProbeNet が RNN，FullMeSH が CNN に基づいている。なお，CNN を MeSH の自動付与に試みた例は他にもある（文献 [10] など）。BERTMeSH は，これらの技術の一部を活用しつつ，医学論文に対して BERT を適用する仕組みであり，損失関数としては BCE が使われ，最終的な MeSH の決定は閾値に基づいてなされている。

なお，日本語テキストに対するニューラルネットワークによるマルチラベル分類の試みとしては，藤井ら (2020)[9] などがある。

2.2.7 最適な閾値を推定する方法

Rank-SVM などでは，新規文書に対して各ラベルの得点（あるいは重み）が算出されるのみなので，当該文書の複数ラベルを確定するには，閾値の設定が必要となる。そのためのアルゴリズムとしては，分類結果の評価指標である F 値を目的関数として反復的に最適解を探索するものが知られている（詳細は文献 [28] を参照）。すなわち，全体的な F 値が最大になるように，ラベルごとに閾値を変えるわけであり，例えば BERTMeSH では，Pillai et al.(2013)[27] によるアルゴリズムが利用されている。ここでは，F 値のマイクロ平均を利用した Pillai et al.(2013)[27] によるアルゴリズムのみ，以下簡単に説明する。

このアルゴリズムではまず，ラベルごとに，訓練データ中の文書の得点を昇順に並べる。そして，それらの間隔のそれぞれの中間の値を閾値の候補としてまずは設定する。その中間値の中からラベルごとに 1 つの閾値を選ぶことになるが，具体的にはそれらの組み合わせのうち全体での F 値のマイクロ平均を最大化するものを探して，最終的な閾値とする。その探索を「虱潰し」に実行するのではなく，

効率的に求める手順が文献 [27] では示されている。

3. データの作成とマルチラベル分類の方法

3.1 評価レビューについてのデータ集合の作成

今回の目的は，ゲームについての新規レビューが与えられた際に，例えば「戦闘システムは良く，物語は悪いと評価している」のように観点別に自動判定するシステムを構築することにある。観点はとりあえず「戦闘システム（以下「戦闘」）」「物語」「サウンド・グラフィックス（以下「サウンド」）の 3 つとし，発表者の 1 人がそれらに関する日本語での評価レビューを Amazon.com から探して手作業で抽出した。作業の過程において，1 つの観点のみに言及している評価レビューに絞り，その「肯定（良い）」「否定（悪い）」を割り付けるのが，正解ラベル付与の効率と正確性の点で望ましいことがわかった。そこで結果的に次の 3 つのデータセットを作成した。

A: 「戦闘」肯定 100 件，否定 100 件
B: 「物語」肯定 100 件，否定 100 件
C: 「サウンド」肯定 100 件，否定 100 件

これらの合計 600 件のレビューにはそれぞれ 1 つのみ正解ラベルが付与されているわけである（シングルラベル）。標本としてはかなり小さいが，今回はこれで実験を試みた。

以下の手順により，この 600 件のレコードを「訓練用」「評価用」に分けた上で分類器の学習と評価を行う。

(1) データ集合 A，B，C の「肯定」「否定」からそれぞれ 90 件を無作為抽出して訓練データとする。
(2) マルチラベルの評価レコード 60 件を，残りのレコードから「人工的に」生成する。
(3) 上記 (1) の 270 件で分類器を学習し，(2) の 60 件のマルチラベルを予測して評価指標を算出する。

そしてこの (1)〜(3) の手順を 10 回反復したのち，最終的に評価指標のそれぞれの値の平均を求めることとした。

上記の (2) におけるマルチラベルデータの「人工的」作成方法は以下のとおりである。

a) 評価用の 1 件のレコードに対してそれぞれ「観点」を無作為抽出する。
b-1) もし自分自身の観点が抽出された場合，そのままシングルラベルのレコードとする。
b-2) もし自分自身の観点とは異なる観点が抽出された場合，その観点に該当する評価レコード 20 件から 1 レコードを無作為抽出する。そして，そのレビューテキストおよび正解ラベルを単純に併合してマルチラベルのレコードとする。

例えば，評価レコード「“戦闘が煩雑である。”：戦闘-否定」に対して別の評価レコード「“物語が泣ける。”：物語-肯定」が選ばれた場合，「“戦闘が煩雑である。物語が泣ける。”：戦闘-否定，物語-肯定」というマルチラベルの人工的レコードが生成される。

3.2 マルチラベルでの分類器

上で見たように，マルチラベル分類のための技法は数多いが，今回は SVM による BR 法と BERT とを取り上げる。

3.2.1 複数の SVM によるマルチラベル分類

この実験でのラベル集合 $\mathcal{Y}$ は，「戦闘-肯定」「戦闘-否定」「物語-肯定」「物語-否定」「サウンド-肯定」「サウンド-否定」の 6 つの要素から構成される。これらに対してそのまま 6 個（1 対他）あるいは 15 個（1 対 1）の分類器を作成するのは効率的でない。そこで，データ集合 A，B，C それぞれに 1 つの SVM を割り当て，「肯定」「否定」「無関係」の多クラスで学習することとした（つまり分類器は観点別に 3 個）。この際，「無関係」に対応する訓練データが必要であり，このため当該観点以外の 2 つの観点の訓練データから 90 件のレコードを無作為抽出した。したがって，各分類器の訓練データ中のレコード件数はそれぞれ 270 となる。もちろんこの作業は，上記 (1)〜(3) の手順を 10 回反復する過程で，それぞれ別個に実行した。

評価データ中のレビューに対しては 3 つの分類器の結果を統合してそのまま単純にマルチラベルとした。例えば「戦闘」が肯定，「物語」が無関係，「サウンド」が肯定ならば，ラベルは $[1, 0, 0, 0, 1, 0]$ となる。つまり，「無関係」の場合には，「肯定」「否定」の両者を「0」とした。

SVM に投入する語をレビューから抽出する方法としては，(i) 形態素解析での語分割の結果から「名詞」「形容詞」を選択，(ii) 辞書を自作しその登録語を選択，の 2 つを試した。後者についてはさらに表記のゆれを統一した。例えば語として「カクカク」「かくかく」「ガクガク」「カクつく」「かくつく」を登録し，いずれも「カクカク」に統一して SVM に投入した。登録語としては，今回の 3 つの観点に関連すると予想された 630 語を設定した。

3.2.2 BERT によるマルチラベル分類

BERT によるマルチラベル分類については，まずは BERT の教科書 [25] で説明されている方法をそのまま使った。すなわち，BERT の出力層での [PAD] 以外の値の平均に対する線型結合で得点を計算し，損失は BCE（torch.nn.BCEWithLogitsLoss）で求めた。マルチラベルの確定は「予測確率が 0.5 を超える場合に 1」と判定することになる（そうでなければ 0）。それに加えて今回は，F 値のマイクロ平均でラベルごとに最適な閾値を求める方法（Pillai et al., 2013[27]）も実行してみた。すなわち，教科書 [25] によるファインチューニング後，ラベルごとに事後的に閾値を調整した。

なお今回の実験では，訓練データにはマルチラベルは含まれず，シングルラベルのみでの学習である。そしてそのチューニング結果をマルチラベル分類の状況で用いたことになる。また，実験用データの規模が小さいことから検証データは設定せず，エポック数は 5 に固定した。以上の点はやや変則的であり，実験結果の解釈には注意を要する。

4. 実験とその結果

4.1 実験環境

SVM については sklearn.svm.SVC で線型カーネルを設定した。その際のトークナイザとしては Janome を用いた。BERT については，Hugging Face のライブラリを使用し，東北大学による訓練済日本語 BERT モデル[*1] で実装を行った（そのトークナイザを各レビューにそのまま適用）。その他の処理（自作辞書での SVM 用の特徴抽出や閾値の最適化 [27] など）については Python でソースコードを自作した。

4.2 実験結果と考察

上で述べたように，SVM の学習では，観点ごとの 180 件のレコードに，それぞれ「無関係」なレコードを単純無作為抽出で 90 件追加することを 10 回繰り返した。そのため，それぞれの 270 件のデータは異なっており，それらに対して Janome および自作辞書により抽出される語の集合も当然，変化する。その平均文書長についての 10 回の反復での平均値を表 1 に示す。

表 1 実験用データの平均文書長

語分割	「戦闘」	「物語」	「サウンド」	全体
Janome	28.48	20.91	18.39	22.60
自作辞書	13.34	9.65	9.24	10.74

念のため，これらの 270 件の集合それぞれに対して，SVM の 10 交差検証（多クラス・シングルラベル）を実行してみたところ，その正解率の平均（10 回の反復での平均）は，3 観点での平均としては Janome で 0.712，自作辞書で 0.743 となった。次にマルチラベル分類の結果を表 2 に示す。ここでの精度や再現率は，マルチラベル分類の実験にて標準的に用いられる定義 [45] に従って算出している。精度については自作辞書による SVM，再現率については閾値を最適化した BERT が最も高かった。F 値については自作辞書による SVM が上回るため，今回の実験では，自作辞書により特徴抽出した SVM による BR 法の性能が最も高いという結果が得られたことになる。

表 2 マルチラベル分類の評価結果

指標	SVM:Janome	SVM:辞書	BERT	BERT+閾値
精度	0.593	<u>0.628</u>	0.274	0.333
再現率	0.552	0.581	0.222	<u>0.868</u>
F 値	0.571	<u>0.603</u>	0.235	0.481

今回の実験では検証データを省略するなど，BERT の実行が粗削りである。また，訓練データのサイズは大きくなく，BERT はこの点で十分なチューニングができなかったのかもしれない。加えて，自作辞書は実験用データを参照しつつ作成したため，こちらはこの点でも有利である。実際，Janome での語の抽出と比較して，自作辞書での方法がより高い性能を示した。

[*1] https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/

5. おわりに

本研究では，「戦闘システム」「物語」「サウンド・グラフィックス」の 3 つの観点に関してゲームの評価レビューの「肯定」「否定」を判定するためのマルチラベル分類の方法を検討した。小さな標本を使った実験では，自作辞書に基づく SVM での BR 法が最も高い性能を示した。マルチラベル分類での BERT の実行手順（閾値の最適化を含む）は今回の実験では不十分で，これについてはさらに探究する必要がある。

参考文献

[1] Al-Salemi, B., Noah, S. A. M., and Aziz, M. J. A.: RF-Boost: An improved multi-label boosting algorithm and its application to text categorisation. *Knowledge-Based Systems*, vol.103, pp.104–117 (2016).
[2] Blockeel, H., De Raedt, L., and Ramon, J.: Top-down induction of clustering trees. *Proceedings of the 15th International Conference on Machine Learning*, vol.cs.LG/0011032, pp.55–63 (1998).
[3] Boutell, M. R., Luo, B., Shen, X., and Brown, C. M.: Learning multi-label scene classification. *Pattern Recognition*, vol.37, pp.1757–1771 (2004).
[4] Cheng, W. and Hüllermeier, E.: Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, vol.76, p.211–225 (2009).
[5] Clare, A. and King, R. D.: Knowledge discovery in multi-label phenotype data. *Principles of Data Mining and Knowledge Discovery. PKDD 2001* (2001).
[6] Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., and Lu, Z.: ML-Net: Multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, vol.26, no.11, pp.1279–1285 (2019).
[7] Elisseeff, A. and Weston, J.: A kernel method for multi-labelled classification. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01)*, pp.681–687 (2001).
[8] Esuli, A., Fagni, T., and Sebastiani, F.: MP-Boost: A multiple-pivot boosting algorithm and its application to text categorization. *String Processing and Information Retrieval, SPIRE 2006*, Springer, pp.1-12 (2006).
[9] 藤井美娜, 阿部智彦, 高橋宏治, 岩城安浩, 加藤恒昭: 契約書のリスク判定のための条文マルチラベル分類. 第 34 回人工知能学会全国大会論文集, 4P3-OS-8-02 (2020).
[10] Gargiulo, F., Silvestri, S., Ciampi, M.: Deep convolution neural network for extreme multi-label text classification. *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF 2018)*, pp.641–650 (2018).
[11] Godbole, S. and Sarawagi, S.: Discriminative methods for multi-labeled classification. *Advances in Knowledge Discovery and Data Mining, PAKDD 2004*, pp.22–

30(2004).
[12] Grodzicki, R., Mańdziuk, J., and Wang, L.: Improved multilabel classification with neural networks. *Parallel Problem Solving from Nature – PPSN X. PPSN 2008*, Springer, pp.409–416 (2008).
[13] Haralabopoulos, G., Anagnostopoulo, I., and McAuley, D.: Ensemble deep learning for multilabel binary classification of user-generated content. *Algorithms*, vol.13, no.4, 83 (2020).
[14] Kocev, D., Vens, C., Struyf, J., and Džeroski, S.: Ensembles of multi-objective decision trees. *Machine Learning: ECML 2007*, pp.624–631 (2007).
[15] Lauser, B. and Hotho, A.: Automatic multi-label subject indexing in a multilingual environment. *Research and Advanced Technology for Digital Libraries (ECDL 2003)*, pp.140–151(2003).
[16] Lenc, L. and Král, P.: Ensemble of neural networks for multi-label document classification. *ITAT 2017 Proceedings*, pp.186–192 (2017).
[17] Li, Y.-K. and Zhang, M.-L.: Enhancing binary relevance for multi-label learning with controlled label correlations exploitation. *PRICAI 2014: Trends in Artificial Intelligence*, pp.91–103 (2014).
[18] Liu, J., Chang, W.-C., Wu, Y., and Yang, Y.: Deep learning for extreme multi-label text classification. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.115–124 (2017).
[19] Madjarov, G., Kocev, D., Gjorgjevikj, D. and Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, vol.45, no.9, pp.3084–3104 (2012).
[20] Mencía, E. L. and Janssen, F.: Learning rules for multi-label classification: A stacking and a separate-and-conquer approach. *Machine Learning*, vol.105, pp.77-126 (2016).
[21] McCallum, A. K.: Multi-label text classification with a mixture model trained by EM. *AAAI 99 Workshop on Text Learning* (1999).
[22] Montañes, E., Senge, R., Barranquero, J., Quevedo, J. R., del Coz, J. J., and Hüllermeier, E.: Dependent binary relevance models for multi-label classification. *Pattern Recognition*, vol.47, pp.1494–1508 (2014).
[23] 中谷知博, 星野准一: 経験的価値の分類に基づくゲーム推薦システム, 情報処理学会研究報告. EC, エンタテインメントコンピューティング, No.11, p.49–56(2008).
[24] Nam, J., Kim, J., Mencía, E. L., Gurevych, I., and Fürnkranz, J.: Large-scale multi-label text classification: Revisiting neural networks. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014*, Springer, pp.437–452 (2014).
[25] 近江崇宏, 金田健太郎, 森長誠, 江間見亜利: BERT による自然言語処理入門：Transformers を使った実践プログラミング, オーム社 (2021).
[26] 大山浩暉, 竹川佳成, 平田 圭二: レビュー文を考慮したゲーム推薦システムの実現に向けた単語の類似度調整の取り組み, エンタテインメントコンピューティングシンポジウム 2017 論文集. p. 223–227(2017).
[27] Pillai, I., Fumera, G., and Roli, F.: Threshold optimisation for multi-label classifiers. *Pattern Recognition*, vol.46, no.7, pp.2055–2065 (2013).
[28] Pillai, I., Fumera, G., and Roli, F.: Designing multi-label classifiers that maximize F measures: State of the art. *Pattern Recognition*, vol.61, pp.394–404 (2017).
[29] Rokach, L., Schclar, A., and Itach, A.: Ensemble methods for multi-label classification. *Expert Systems with Applications*, vol.41, pp.7507–7523 (2014).
[30] Schapire, R. E. and Singer, Y.: BoosTexter: A boosting-based system for text categorization. *Machine Learning*, vol.39, pp.135–168 (2000).
[31] Sucar, L. E., Bielza, C., Morales, E. F., Hernandez-Leal, P., Zaragoza, J. H., and Larrañaga, P.: Multi-label classification with Bayesian network-based chain classifiers. *Pattern Recognition Letters*, vol.41, pp.14-22 (2014).
[32] Spyromitros, E., Tsoumakas, G., and Vlahavas, I.: Empirical study of lazy multilabel classification algorithms. *SETN 2008: Artificial Intelligence: Theories, Models and Applications*, Springer, pp 401–406 (2008).
[33] Tahir, M. A., Kittler, J., and Bouridane, Multi-label classification using stacked spectral kernel discriminant analysis. *Neurocomputing*, vol.171, pp.271–137 (2016).
[34] Tsoumakas G. and Katakis, I.: Multi-label classification: An overview. *International Journal of Data Warehouse and Mining*, vol.3, no.3, pp.1–13 (2007).
[35] Tsoumakas G. and Vlahavas, I.: Random k-labelsets: An ensemble method for multilabel classification. *Machine Learning: ECML 2007*, pp.406–417 (2007).
[36] Tsoumakas, G., Katakis, I., and Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. *Proceedings of the ECML/PKDD Workshop on Mining Multidimensional Data*, pp.30–44 (2008).
[37] Read, J., Pfahringer, B., Holmes, G., and Frank, E.: Classifier chains for multi-label classification. *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2009)* (2009).
[38] Rios, A. and Kavuluru, R.: Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles. *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB) 2015*, pp.258–267 (2015).
[39] Yang, F., Wang, H., Feng, L., and Lai, Y.: CP-RAkEL: Improving random k-labelsets with conformal prediction for multi-label classification. *Proceedings of Machine Learning Research*, vol.60, pp.1–14 (2017).
[40] You, R., Liu, Y., Mamitsuka, H., and Zhu, S.: BERTMeSH: Deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics*, vol.37, no.5, pp.684–692 (2021).
[41] Zhang, Y., Li, Y., and Cai, Z.: Correlation-based pruning of dependent binary relevance models for multi-label classification, *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI*CC 2015*, pp.399–404 (2015).
[42] Zhang, M.-L., Li, Y., Liu, X.-Y., and Geng, X.: Binary relevance for multi-label learning: An overview. *Frontiers of Computer Science*, vol.12, no.2, pp.191–202 (2018).
[43] Zhang, M.-L. and Zhou, Z.-H.: Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, vol.18, no.10, pp.1338–1351 (2006).
[44] Zhang, M.-L. and Zhou, Z.-H.: ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, vol.40, no.7, pp.2038–2048 (2007).
[45] Zhang, M.-L. and Zhou, Z.-H.: A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, vol.26, no.8, pp.1819–1837 (2014).