

# BERTを用いた テレビドラマに関する関心动向・感想の ウェブマイニング

川口輝太<sup>1</sup> 久保遼馬<sup>1</sup> 藤田拓也<sup>1</sup> 前田竜治<sup>1</sup> 宇津呂武仁<sup>1</sup> 小林彰夫<sup>2</sup> 西崎博光<sup>3</sup>  
河田容英<sup>4</sup>

概要：本論文では、テレビドラマ視聴者がドラマ視聴後にウェブ上で行うドラマ関連関心动向・感想・レビュー類の情報探索過程を支援することを目的として、ブログ・ドラマ関連サイト等のウェブページからの情報収集・集約を行うウェブマイニング技術を提案する。具体的には、本論文では、BERT および Wikipedia を用いて、文単位での当該ドラマ関連判定および主観情報判定を行うとともに、文単位での判定結果に基づいて、ウェブページ単位での当該ドラマ関連判定および主観情報判定を行う手法、および、その評価結果について述べる。

## Web Mining of Concerns and Reviews of TV Drama based on BERT

KOTA KAWAGUCHI<sup>1</sup> RYOMA KUBO<sup>1</sup> TAKUYA FUJITA<sup>1</sup> TATSUYA MAEDA<sup>1</sup> TAKEHITO UTSURO<sup>1</sup>  
AKIO KOBAYASHI<sup>2</sup> HIROMITSU NISHIZAKI<sup>3</sup> YASUHIDE KAWADA<sup>4</sup>

### 1. はじめに

本論文では、テレビドラマ視聴者がドラマ視聴後にウェブ上で行うドラマ関連関心动向・感想・レビュー類の情報探索過程を支援することを目的として、ブログ・ドラマ関連サイト等のウェブページからの情報収集・集約を行うウェブマイニング技術を提案する。本論文でマイニング対象とするウェブページは、ブログ等、ドラマ視聴者がドラマ視聴後に当該ドラマについての関心事項や感想を掲載するページ(図1)、および、ブログ以外にも、ドラマのDVD・関連グッズ等を販売するECサイトのページ、ドラマに関する

記事を掲載するエンターテインメント系のネットニュース記事、ドラマの広告等の動画を掲載する動画サイト、ドラマに出演する俳優等によるインスタグラムのページ等、ドラマ視聴者が継続的にアクセスする多種多様なウェブページ(図2)が含まれる。本論文においてそれらのウェブページ群をマイニングして、情報収集・集約を行った結果を提供する利用者像としては、各ドラマの熱狂的視聴者が挙げられるが、そのような熱狂的視聴者にとどまらず、偶然当該ドラマに関心を持ち、ウェブ上のブログや関連サイトを探して情報収集をしようと試みる一時的な視聴者も重要な利用者として位置付けられる。

以上の背景のもとで、本論文では、まず、2018年10~12月のクールを放送時期とする4つのドラマ「リーガルV」、「下町ロケット」、「僕らは奇跡でできている」、「あなたには渡さない」を対象として、各ドラマに関連するウェブページを実際に収集し、そのウェブページの種類、当該ドラマに関する感想の有無等の人手調査を行った。具体的には、2節で述べる手順により、上述の4つのドラマに関連する話題が掲載されている可能性の高いウェブページを網羅的

<sup>1</sup> 筑波大学 大学院システム情報工学研究科  
University of Tsukuba, Tsukuba, 305-8573, Japan

<sup>2</sup> 筑波技術大学産業技術学部  
Tsukuba University of Technology, Tsukuba, 305-8520, Japan

<sup>3</sup> 山梨大学大学院総合研究部  
Graduate School of Interdisciplinary Research, University of Yamanashi, Kofu, 400-8511, Japan

<sup>4</sup> (株) ログワークス  
Logworks Co., Ltd., Tokyo, 151-0053, Japan

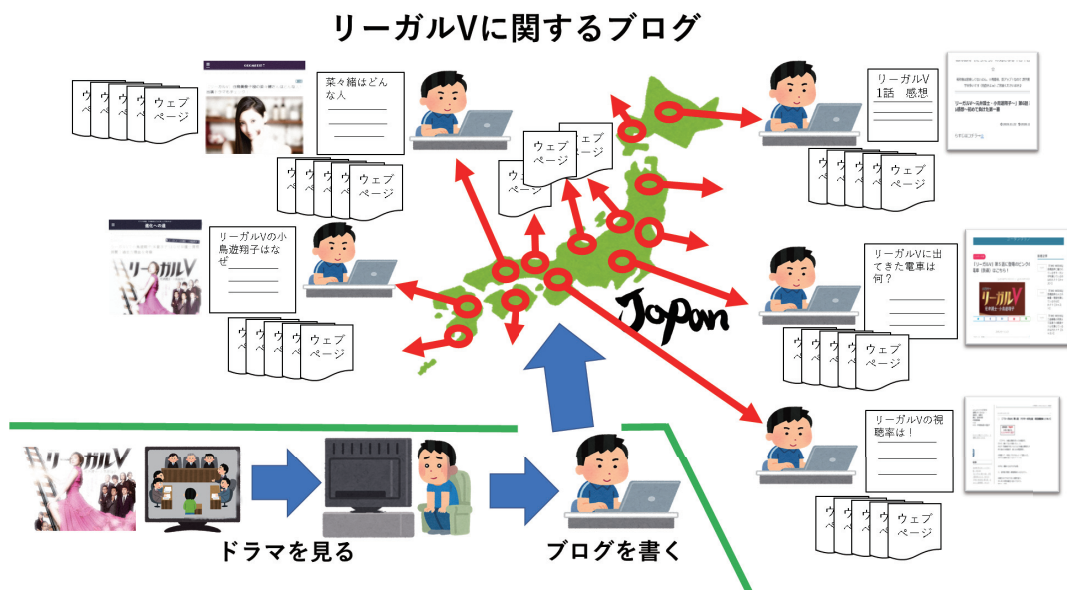


図 1 テレビドラマに関する関心事項・感想についてのブログ記事集合

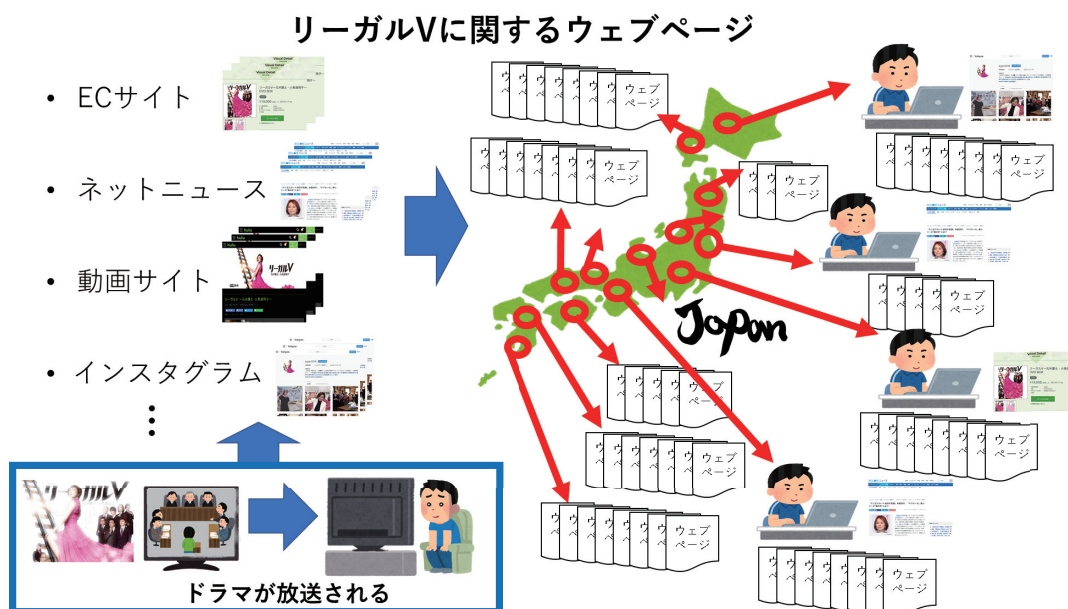


図 2 テレビドラマに関連する各種ウェブページ集合 (ブログ以外)

に収集し、表 1 に示すように、各ドラマに関連する話題の掲載の有無、ブログか否かの区別、ブログの場合の感想の掲載の有無、等について人手分析を行った<sup>\*1</sup>。その結果、表 1 に示すように、各ドラマと密接に関連する話題が掲載されているウェブサイトが、各ドラマごとに約 200~300 程度存在し、そのうち、約 20~50 サイトは、当該ドラマに関する感想を掲載したページが存在するブログであった。

さらに、それらの当該ドラマに関する感想を掲載するブログとは別に、当該ドラマについて、感想以外の関心动向を掲載するブログが約 20~50 サイト存在することが確認された。

この分析結果をふまえて、本論文において、テレビドラマ視聴者がドラマ視聴後にウェブ上で行うドラマ関連関心动向・感想・レビュー類の情報探索過程を支援することを目的として、ブログ・ドラマ関連サイト等のウェブページからの情報収集・集約を行うウェブマイニング技術の枠組みを図 3 に示す。この枠組みにおいては、まず、各ドラマに関して、ブログ記事、EC サイトのページ、ネットニュースの記事、動画サイトのページ、インスタグラムのページ

<sup>\*1</sup> ブログ以外のウェブページにおいても何らかの感想を含む文が掲載されているが、大抵の場合、それらの文において感想が述べられている対象は、当該ドラマの内容そのものではなく、DVD や関連グッズの販売に関してであることが多い。その他、エンターテインメント系のネットニュースにおいて、ドラマのキャストである俳優へのインタビュー記事が掲載されており、その俳優の発言中に主観表現が含まれる場合なども含まれる。

表 1 各ドラマに対して収集されたウェブサイト・ウェブページの数

ドラマタイトル	サジェ スト数	収集された 全サイト対象		Q 中のウェブページ数が 10 以上のサイト対象						
		ウェブ ペー ジ数	ウェブ サイ ト数	ウェブ ペー ジ数	ウェブサイト数				クエリ・ フォーカス のドラマに 関連しない	
					合計	感想を 含む ブログ	感想を 含まない ブログ	ブログ 以外		合計
「リーガルV」	672	2,621	791	2,057	339	47	41	218	306	33
「下町ロケット」	782	4,688	1,393	3,286	340	51	35	229	315	25
「僕らは奇跡で できている」	630	1,912	639	1,457	280	46	34	176	256	24
「あなたには 渡さない」	358	1,019	351	830	197	26	24	136	186	11

等, 多種多様なウェブページが網羅的に収集される。その後, 収集されたウェブページ中のコンテンツが, 当該ドラマについての感想, および, 当該ドラマについての感想以外の, キャスト, 俳優, 当該ドラマ特有のキーワード\*2, 視聴率・ロケ地等のドラマジャンル全体で共通の関心事項等, 主要な分類へと類別され, 各分類において冗長な情報の集約が行われる。そして, それらの集約された情報がドラマ視聴者へ提示される。この枠組みの中の要素技術として, 本論文では特に, ウェブページ単位での当該ドラマ関連判定および感想を含むか否かの判定を行う手法を提案する。具体的には, 本論文では, BERT [2] および Wikipedia を用いて, 文単位での当該ドラマ関連判定および主観情報判定を行うとともに, 文単位での判定結果に基づいて, ウェブページ単位での当該ドラマ関連判定および主観情報判定を行う手法, および, その評価結果について述べる。

## 2. 検索エンジン・サジェストを用いたドラマ関連ウェブページの収集

本論文では, 検索者が詳細な情報を検索したい対象を「クエリ・フォーカス」と呼ぶ [4]。そして, 本論文では, ドラマのタイトルをクエリ・フォーカスとして, 一つのクエリ・フォーカスに対して, 最大約 1,000 語のサジェストを収集する。そして, クエリ・フォーカスに加えて一つの検索エンジン・サジェストを指定した AND 検索によってウェブページを収集する。最大約 1,000 個の検索エンジン・サジェストに対してこの方法を用いることにより, あるクエリ・フォーカスに関する大規模なウェブページ集合を収集することが出来る。

具体的には, 本論文では, 2018 年 10~12 月のクールを放送時期とする 4 つのドラマ「リーガルV」, 「下町ロケッ

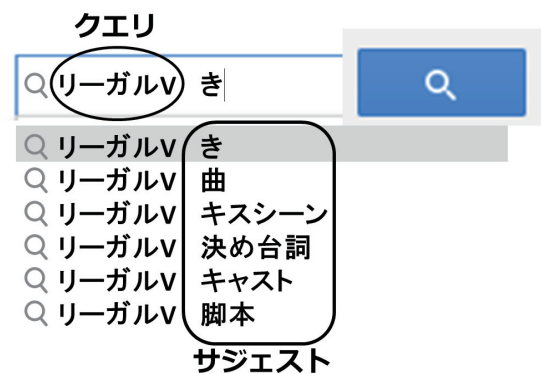


図 4 検索エンジン・サジェストの例

ト」, 「僕らは奇跡でできている」, 「あなたには渡さない」を対象として, 各ドラマに関連するウェブページを収集する。本論文の手順において, 実際に各ドラマについてのサジェストおよびウェブページを収集する手順の詳細を以下の各節で述べる。これらの各ドラマに対して, 以下の各節の手順により収集されたサジェスト数, ウェブページ数, および, ウェブサイト数を表 1 に示す。

### 2.1 検索エンジン・サジェストの収集

本論文では, Google 社の検索エンジン・サジェストを用いて, クエリ・フォーカスに対する情報要求観点を収集する。一つのクエリ・フォーカスあたり 100 通りの文字列を指定する。一文字列あたり最大 10 個のサジェストを収集できるため, 理論上, 最大 1,000 語のサジェストを収集することができる。100 通りの文字列としては, 五十音, 濁音, 半濁音および開拗音を含む。図 4 に示すように, 例えば検索窓に「リーガルV き」と入力すると, 「曲」や「決め台詞」等がサジェストとして提示されるので, それらの収集を行う。この手順により収集したサジェストの数を表 1 に示す。

\*2 例えば, 法廷を題材とするドラマである「リーガルV」の場合には, 法廷用語である「法廷」, 「検事」, 「弁護士」等が当該ドラマ特有のキーワードの一例となる。



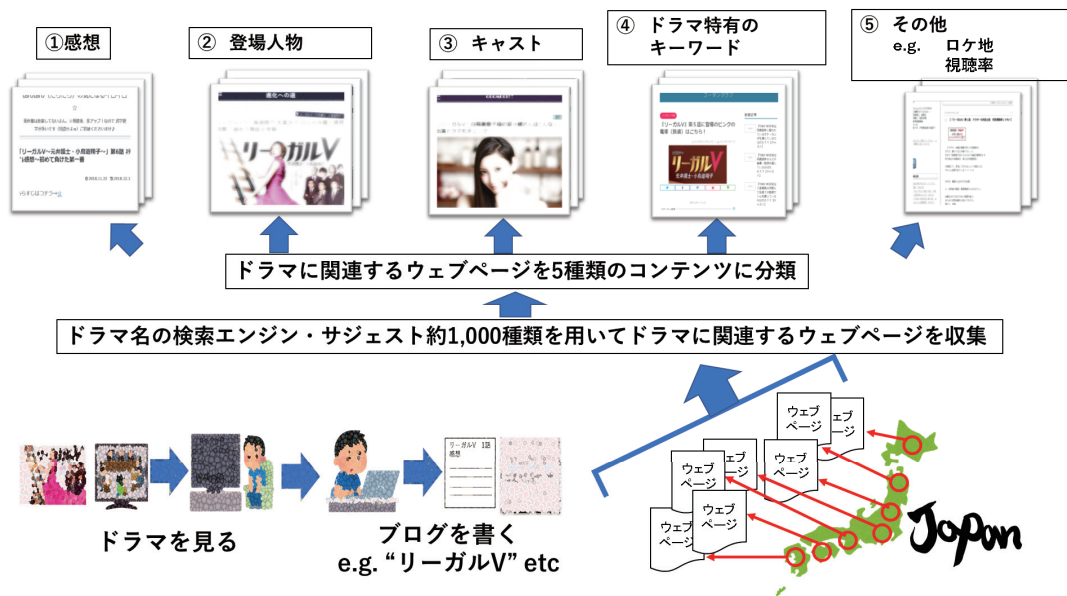


図 3 本論文の枠組み: ドラマ関連ウェブページの収集およびドラマに関する関心事項・感想の分類

## 2.2 ウェブページの収集

ウェブページの収集においては, Google Custom Search API<sup>\*3</sup> を用いて, 検索語として, 「クエリ・フォーカス AND サジェスト」の AND 条件での検索語を指定して, ウェブページを収集する. ここで, クエリ・フォーカスに対して収集されたサジェストの集合を  $S$  として, あるサジェスト  $s \in S$  に対して, クエリ・フォーカスとの AND 検索の検索結果上位  $N$  件以内に検索されるウェブページの集合を  $S(q)$  とする. 本論文では,  $N = 20$  とした. このとき, 各クエリ・フォーカスに対して収集されるウェブページの集合  $S(q)$  を次式で定義する.

$$S(q) = \left\{ s \in S \mid q \in Q(s, N) \right\}$$

また, 収集されたサジェストとクエリフォーカスとの AND 検索により収集されたウェブページを  $Q$  とする.

$$Q = \bigcup_{s \in S} Q(s, N)$$

各ドラマに対して実際に収集されたウェブページ数, および, それらのウェブページの収集元となったウェブサイトを表 1 に示す. また, 本論文における評価実験においては, 各サイトのうち, 本節の手順により収集されたウェブページ数が 10 以上となるサイトのみを対象とする. このことをふまえて, 表 1 においては, 評価対象とするサイトの数, および, それらのサイト上のページのうち本節の手順による収集されたウェブページの数もあわせて示す. さらに, 表 1 においては, 本論文で評価対象とするサイトに対して, 人手により, 「クエリ・フォーカスのドラマに関連する」, 「クエリ・フォーカスのドラマに関連しない」を

分類し, 「クエリ・フォーカスのドラマに関連する」サイトに対して, 「感想を含むブログ」, 「感想を含まないブログ」, 「ブログ以外」に分類した結果もあわせて示す.

## 3. 評価用データセット

評価用データセットの作成にあたっては, 表 1 中の「 $Q$  中のウェブページ数が 10 以上のサイト対象」欄に該当するウェブページ集合から, 各ドラマに対して, 表 2(a) に示す数のページを無作為に選定し評価実験において用いる<sup>\*4</sup>. また, 表 1 中の「 $Q$  中のウェブページ数が 10 以上のサイト対象」欄に該当するウェブページ集合から, 各ドラマに対して, 表 2(b) に示す数の文を無作為に選定し評価実験において用いる. 評価用文の選定においては, まず, 各ドラマごとに 16 サイトを選定した後<sup>\*5</sup>, 16 サイトの各々から文選定用のページを各 1 ページ選定し, 文選定の情報源とした<sup>\*6</sup>.

<sup>\*4</sup> 表 2(a) に示す数のページを選定する際には, 各ドラマごとに無作為に 200 ページ選定した後, 表 2(b) 中の訓練文を選定する際に用いたサイト (当該ドラマ以外の 3 ドラマのタイトルをクエリ・フォーカスとして収集したウェブページのサイトのうちの 48 サイト) のページを除外し, 結果的に, 表 2(a) に示す数のページが選定された.

<sup>\*5</sup> 全 16 サイトの内訳として, 4 サイトは, ドラマ視聴者によるブログサイトのうちクエリ・フォーカスのドラマに関する感想を含むもの, 4 サイトは, クエリ・フォーカスのドラマに関する内容を含むブログサイトであるが, クエリ・フォーカスのドラマに関する感想を含まないもの, 4 サイトは, ブログサイト以外であるが, クエリ・フォーカスのドラマに関する内容を含むサイト, 残り 4 サイトは, 本論文の手順により収集されたページを含むが, クエリ・フォーカスのドラマに関する内容を含まないサイトとした.

<sup>\*6</sup> 文選定の際には, 4 節で述べる手順で各ページ中の文を形態素単位に分割した後, 5 形態素以上から構成される文をすべて選定した.

<sup>\*3</sup> <https://cse.google.com/cse/>

表 2 評価用データセット (評価対象ドラマ: 「リーガルV」, 「下町ロケット」, 「僕らは奇跡でできている」, 「あなたには渡さない」)

(a) ウェブページ数

	クエリ・フォーカスの ドラマに関連する	クエリ・フォーカスの ドラマに関連しない	合計
感想を 含む	19 / 64 / 59 / 15 (合計) 157	40 / 62	677
感想を 含まない	97 / 57 / 77 / 120 (合計) 351	28 / 39 (合計) 169	
total	508	169	677

(b) 文数

	クエリ・フォーカスの ドラマに関連する	クエリ・フォーカスの ドラマに関連しない	合計
感想を 含む	120 / 143 / 65 / 162 (合計) 490	45 / 22 / 84 / 77 (合計) 228	718
感想を 含まない	389 / 722 / 384 / 374 (合計) 1,869	787 / 685 / 761 / 639 (合計) 2,872	4,741
合計	2,359	3,100	5,459

## 4. 日本語版 BERT

BERT [2] の日本語版を用いるにあたって, BERT の実装としては, PyTorch 版<sup>\*7</sup>を用い, 日本語版 BERT の事前訓練済みモデル<sup>\*8</sup>を用いた. 日本語版 BERT の事前訓練済みモデルを利用するには, まず, 収集したウェブページの日本語テキストに対して, mecab-ipadic-NEologd 辞書<sup>\*9</sup>を用いた日本語形態素解析ツール MeCab を適用して形態素単位への分割を行った後, BERT の fine-tuning モジュール内の BPE(byte pair encoding) モジュール<sup>\*10</sup>を適用することにより, 語彙数 32,000 語のサブワード単位へと分割する. そして, BERT の fine-tuning モジュールを用いて fine-tuning された BERT モデルを適用する.

## 5. BERT を用いたドラマ関連判定

本論文の枠組みにおいて, クエリ・フォーカスとして指定されたドラマ名に対して, 検索エンジン・サジェストを用いて収集されたウェブページの話者が当該ドラマに関連するか否かを判定するドラマ関連判定モデルを図 5, および, 図 6 に示す.

### 5.1 文単位判定

本論文の枠組みにおいて, クエリ・フォーカスとして指

定されたドラマ名に対して, 検索エンジン・サジェストを用いて収集されたウェブページ中の文の話者が当該ドラマに関連するか否かを判定する文単位ドラマ関連判定モデルの枠組みを図 5 に示す. この枠組みにおいては, 文の話者が当該ドラマに関連するか否かを判定するタスクにおいて BERT を用いる. そこでは, 話題判定対象の文に加えて, 当該ドラマの Wikipedia エントリ中のあらすじ部分のテキストの二つを BERT に入力して BERT の fine-tuning を行い, 文の話者の判定を行う<sup>\*11</sup>.

本節の文単位ドラマ関連判定モデルの fine-tuning および評価実験において用いる文の数は, 表 2(b) に示すように, 話題が当該ドラマに関連する文が 2,359 文, 話題が当該ドラマに関連しない文が 3,100 文である. 評価実験においては, 評価対象の 4 ドラマを用いて, 3 ドラマをクエリ・フォーカスとして収集したウェブページから抽出した文を fine-tuning の訓練事例とし, 残りの 1 ドラマをクエリ・フォーカスとして収集したウェブページから抽出した文を評価事例とする交差検定により, 文単位ドラマ関連判定モデルの評価を行った.

BERT の fine-tuning モデルにおいて softmax 関数の出力確率の下限値を変化させて再現率・適合率曲線を描いた結果を図 7(a) に示す. この結果から分かるように, 特に, softmax 関数の出力確率の下限値が高い範囲では, 十分に高い適合率となっている.

<sup>\*7</sup> <https://github.com/huggingface/pytorch-pretrained-BERT>

<sup>\*8</sup> 日本語 Wikipedia の 1800 万文を用いて事前訓練された Japanese\_L-12\_H-768\_A-12\_E-30\_BPE.zip モデル (<http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT> 日本語 Pre-trained モデル)

<sup>\*9</sup> <https://github.com/neologd/mecab-ipadic-neologd>

<sup>\*10</sup> <https://github.com/rsennrich/subword-nmt>

<sup>\*11</sup> SQuAD1.1 において BERT の評価を行う際のパラメータ設定 (<https://github.com/google-research/bert>) においては, 質問部分およびコンテキスト (回答の探索範囲となるテキスト) 部分を連結したテキストにおけるトークン数の上限が 384 である. 本論文では, この設定をふまえて, 当該ドラマの Wikipedia エントリから抽出するあらすじのトークン数を 300 形態素とし, 話題判定対象の文の形態素数を 80 以下とする.

## 5.2 ページ単位判定

本論文の枠組みにおいて、クエリ・フォーカスとして指定されたドラマ名に対して、検索エンジン・サジェストを用いて収集されたウェブページの話題が当該ドラマに関連するか否かを判定するページ単位ドラマ関連判定モデルの枠組みを図 6 に示す。本論文の枠組みにおいては、ウェブページ  $q$  中の各文  $x (\in X(q))$ 、ウェブページ  $q$  中の文の集合) に対して文単位ドラマ関連判定モデルを適用した結果に基づき、ウェブページ  $q$  の話題が当該ドラマに関連するか否かを判定する。各文  $x$  に対して文単位ドラマ関連判定モデルを適用した結果は、softmax 関数の出力確率  $p(x)$  (ただし、 $0 \leq p(x) \leq 1$ ) によって表される。ここで、本論文のページ単位ドラマ関連判定モデルにおいては、ウェブページ  $q$  中の各文  $x$  に対する文単位ドラマ関連判定結果の確率値  $p(x)$  ( $x \in X(q)$ ) の分布に対して、ページ単位でのドラマ関連判定を行うタスクに対して SVM<sup>\*12</sup> を適用する。

SVM の素性においては、文単位ドラマ関連判定結果の確率  $p(x)$  の下限値を規定するための以下の 99 個の関数  $\text{lbd}_i(p)$

$$\text{lbd}_i(p) = 1 \quad (p \geq 0.01 \times i \text{ の場合}) \quad (i = 1, \dots, 99)$$

を用いる。次に、ウェブページ  $q$  中の文のうち、文単位ドラマ関連判定結果の確率  $p(x)$  の下限値  $\text{lbd}_i$  を満たす文数  $N_i(q)$  を次式で定義する。

$$N_i(q) = \left| \left\{ x \in X(q) \mid \text{lbd}_i(p(x)) = 1 \right\} \right|$$

そして、ウェブページ  $q$  中の文のうち、文単位ドラマ関連判定結果の確率  $p(x)$  の下限値  $\text{lbd}_i$  を満たす文数を規定する 495 (=99 × 5) 個の素性  $F_{ij}(q)$  ( $i = 1, \dots, 99, j = 1, \dots, 5$ ) を次式で定義し、本論文のページ単位ドラマ関連判定の SVM モデルの素性として用いる。

$$F_{ij}(q) = 1 \quad (3j - 2 \leq N_i(q) \leq 3j \text{ の場合})$$

本節のページ単位ドラマ関連判定モデルの評価実験において用いるウェブページの数は、表 2(a) に示すように、話題が当該ドラマに関連するウェブページ数が 508、話題が当該ドラマに関連しないウェブページ数が 169 である。評価実験においては、評価対象の 4 ドラマを用いて、3 ドラマをクエリ・フォーカスとして収集したウェブページを訓練事例とし、残りの 1 ドラマをクエリ・フォーカスとして収集したウェブページを評価事例とする交差検定により、ページ単位ドラマ関連判定モデルの評価を行った。ただし、訓練においては、無作為抽出により、話題が当該ドラマに関連するウェブページ数を、話題が当該ドラマに関連しないウェブページ数に揃えた。一方、評価時には、二値クラス間の事例数は揃えず、全事例を用いて評価を行った。

SVM の信頼度の下限値を変化させて再現率・適合率曲

<sup>\*12</sup> scikit-learn [6] を用いた。

線を描いた結果を図 7(b) に示す。本論文のページ単位ドラマ関連判定タスクにおいては、SVM の信頼度に下限値を設けない場合のベースライン適合率は相対的に高い値 (0.75) となっているが、特に SVM の信頼度の下限値が高い範囲では、ベースライン適合率と比較しても十分に高い適合率となっている。

## 6. BERT を用いた感想判定

次に、本論文の枠組みにおいて、検索エンジン・サジェストを用いて収集されたウェブページが感想を含むかを判定する感想判定モデルを図 8、および、図 9 に示す。

### 6.1 文単位判定

本論文の枠組みにおいて、検索エンジン・サジェストを用いて収集されたウェブページ中の文が感想を含むかを判定する文単位感想判定モデルの枠組みを図 8 に示す。この枠組みにおいては、文が感想を含むかを判定するタスクにおいて BERT を用いる。そこでは、事前訓練済みの BERT [2] の分散表現を用いた双方向 LSTM による分類器<sup>\*13</sup> を用いた<sup>\*14</sup>。

本節の文単位感想判定モデルの評価実験において用いる文の数は、表 2(b) に示すように、感想を含む文が 718 文、感想を含まない文が 4,741 文である。評価実験においては、評価対象の 4 ドラマを用いて、3 ドラマをクエリ・フォーカスとして収集したウェブページから抽出した文を訓練事例とし、残りの 1 ドラマをクエリ・フォーカスとして収集したウェブページから抽出した文を評価事例とする交差検定により、文単位感想判定モデルの評価を行った。ただし、訓練および評価においては、無作為抽出により、感想を含まない文の数を感想を含む文の数に揃えた。

双方向 LSTM モデルにおいて全結合層の sigmoid 関数の出力確率の下限値を変化させて再現率・適合率曲線を描いた結果を図 10(a) に示す。この結果から分かるように、二値クラスの事例数が均衡するという条件のもとでは、特に、sigmoid 関数の出力確率の下限値が高い範囲では、十分に高い適合率となっている。

### 6.2 ページ単位判定

本論文の枠組みにおいて、検索エンジン・サジェストを用いて収集されたウェブページが感想を含むかを判定するページ感想判定モデルの枠組みを図 6 に示す。本論文の枠組みにおいては、ウェブページ  $q$  中の各文  $x (\in X(q))$ 、ウェブページ  $q$  中の文の集合) に対して文単位感想判定モデルを適用した結果に基づき、ウェブページ  $q$  が感想を含むかを判定する。このページ単位感想判定モデルにお

<sup>\*13</sup> Keras(<https://keras.io/>) 上の実装を用いた。

<sup>\*14</sup> 5.1 節と同様の、BERT の fine-tuning を適用する枠組みの評価も行ったが、本節で述べる枠組みの性能を下回る結果となった。



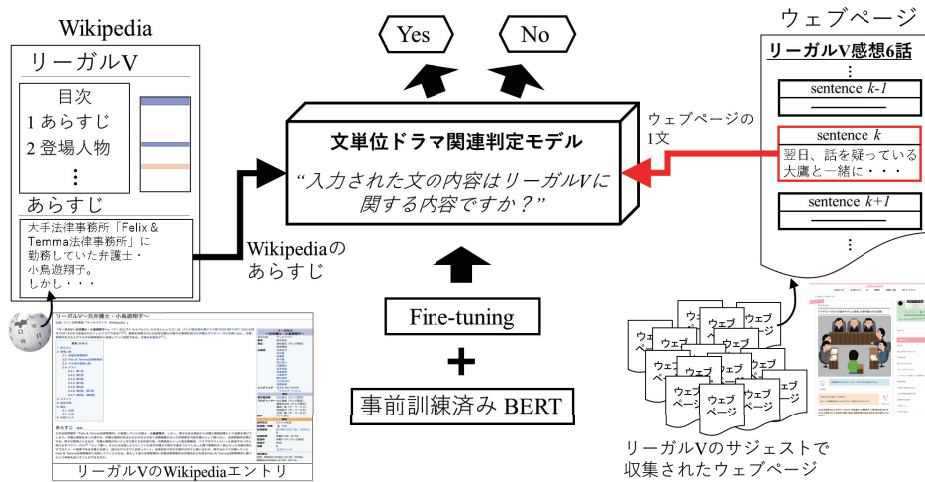


図 5 文単位ドラマ関連判定モデル

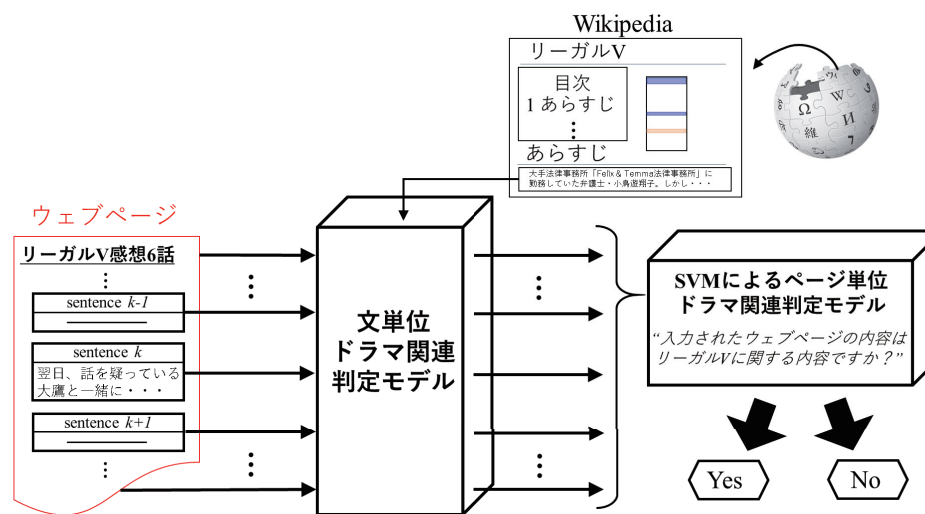


図 6 ページ単位ドラマ関連判定モデル

いても、5.2 節のページ単位ドラマ関連判定モデルと同様に、ウェブページ  $q$  中の各文  $x$  に対する文単位感想判定結果の確率値  $p(x)$  ( $x \in X(q)$ ) の分布に対して、ページ単位での感想判定を行うタスクに対して SVM を適用する。SVM の素性としては、5.2 節で述べた素性を用いる。

本節のページ単位感想判定モデルの評価実験において用いているウェブページの数は、表 2(a) に示すように、感想を含むウェブページ数が 157、感想を含まないウェブページ数が 351 である<sup>\*15</sup>。評価実験においては、評価対象の 4 ドラマを用いて、3 ドラマをクエリ・フォーカスとして収集したウェブページを訓練事例とし、残りの 1 ドラマをクエリ・フォーカスとして収集したウェブページを評価事例とする交差検定により、ページ単位ドラマ関連判定モデルの評価を行った。ただし、訓練においては、無作為抽出により、感想を含まないウェブページ数を、感想を含むウェブ

\*15 前節で述べた文単位感想判定モデルの評価とは異なり、本節のページ単位感想判定モデルの評価においては、ウェブページ収集の際にクエリ・フォーカスとして用いたドラマ名に関連する話題のウェブページを対象として評価を行う。

ページ数に揃えた。一方、評価時には、二値クラス間の事例数は揃えず、全事例を用いて評価を行った。

SVM の信頼度の下限値を変化させて再現率・適合率曲線を描いた結果を図 10(b) に示す。本論文のページ単位ドラマ関連判定タスクにおいては、SVM の信頼度の下限値が高い範囲においても、提案手法の適合率は低めの値 (0.7 程度) となっている。しかし、提案手法の適合率は、SVM の信頼度に下限値を設けない場合のベースライン適合率 (0.31) と比較すると、相対的には高い値となっており、ベースライン適合率を改善していると言える。

## 7. 関連研究

BERT [2] を利用した関連研究の一例として、レビュー理解・主観分析において利用するもの [1], [3], イベント検出において利用するもの [8], SNS マイニングにおいて利用するもの [5], 意見・主張分析において利用するもの [7] 等が挙げられる。一方、本論文においては、テレビドラマ視聴者による情報探索過程支援を目的としてウェブページ

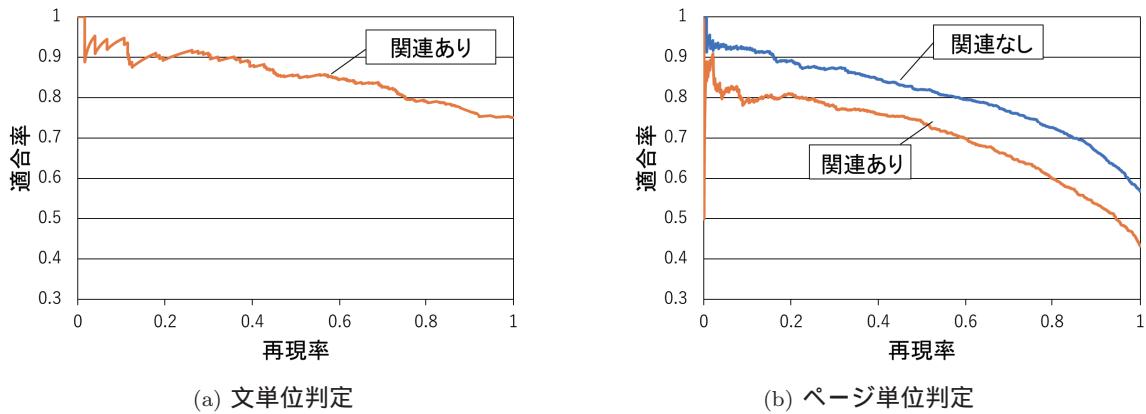


図 7 ドラマ関連判定の評価結果

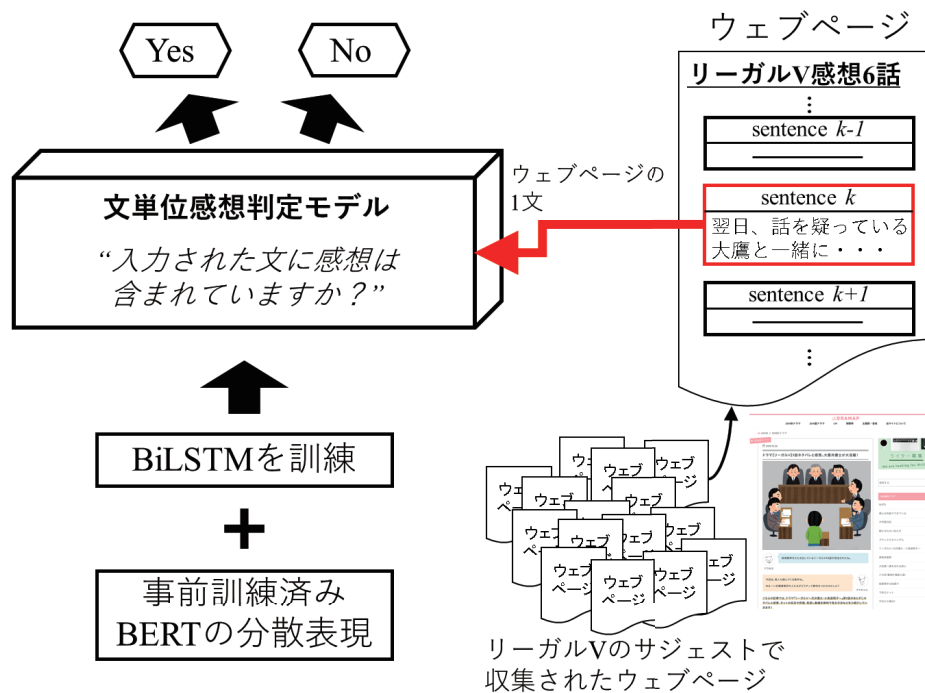


図 8 文単位感想判定モデル

単位での当該ドラマ関連判定および主観情報判定を行う手法を提案しており、先行研究とは問題設定が異なる。

## 8. おわりに

本論文では、テレビドラマ視聴者がドラマ視聴後にウェブ上で行うドラマ関連関心动向・感想・レビュー類の情報探索過程を支援することを目的として、ブログ・ドラマ関連サイト等のウェブページからの情報収集・集約を行うウェブマイニング技術を提案した。具体的には、本論文では、BERT および Wikipedia を用いて、文単位での当該ドラマ関連判定および主観情報判定を行うとともに、文単位での判定結果に基づいて、ウェブページ単位での当該ドラマ関連判定および主観情報判定を行う手法、および、その評価結果について述べた。今後の課題として、ドラマ関連関心动向の詳細分類として、キャスト、俳優、当該ドラマ

特有のキーワード、視聴率・ロケ地等のドラマジャンル全体で共通の関心事項等への識別を行うことが挙げられる。また、ツイッター等のドラマ視聴者によるリアルタイムの関心动向・感想の追跡を行い、関心动向・感想の時系列変化をマイニングし、本論文の枠組みに統合することが挙げられる。

## 参考文献

- [1] Chi, S., Luyao, H. and Xipeng, Q.: Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, *Proc. NAACL-HLT*, pp. 380–385 (2019).
- [2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. NAACL-HLT*, pp. 4171–4186 (2019).
- [3] Hu, X., Bing, L., Lei, S. and Philip, Y.: BERT Post-Training for Review Reading Comprehension and Aspect-



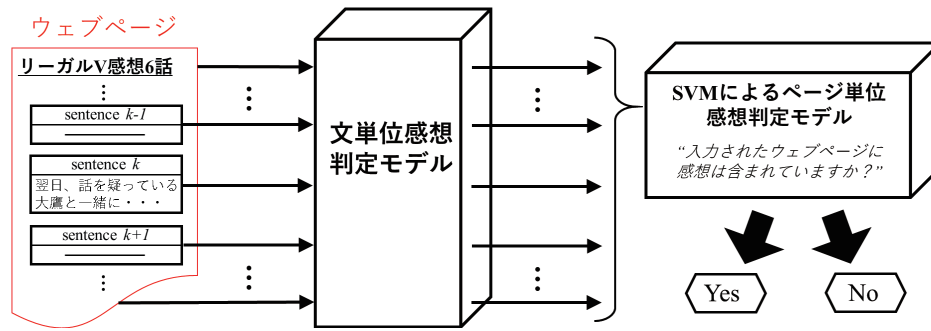


図 9 ページ単位感想判定モデル

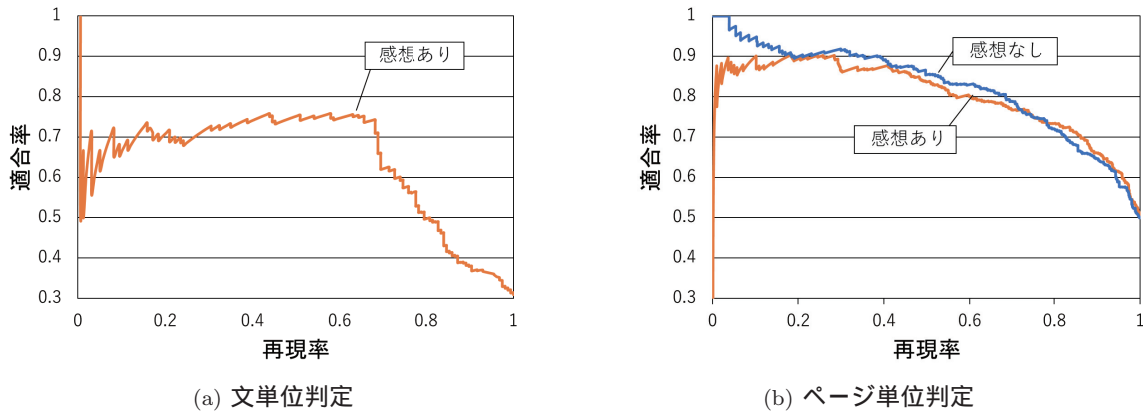


図 10 感想判定の評価結果

based Sentiment Analysis, *Proc. NAACL-HLT*, pp. 2324–2335 (2019).

[4] 井上祐輔, 今田貴和, 陳 磊, 徐 凌寒, 宇津呂武仁, 河田容英: 検索エンジン・サジェストおよびトピックモデルを用いたウェブ検索結果の集約, 第 8 回 DEIM フォーラム論文集 (2016).

[5] Jack, H. and Lillian, L.: Something’s Brewing! Early Prediction of Controversy-causing Posts from Discussion Features, *Proc. NAACL-HLT*, pp. 1648–1659 (2019).

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

[7] Sihao, C., Daniel, K., Wenpeng, Y., Chris, C. and Dan, R.: Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims, *Proc. NAACL-HLT*, pp. 542–557 (2019).

[8] Xiaozhi, W., Xu, H., Zhiyuan, L., Maosong, S. and Peng, L.: Adversarial Training for Weakly Supervised Event Detection, *Proc. NAACL-HLT*, pp. 998–1008 (2019).