International Journal on Digital Libraries (2021) 22:63–83
https://doi.org/10.1007/s00799-020-00293-5

Feature selection for classifying multi-labeled past events

Yasunobu Sumikawa[1] · Ryohei Ikejiri[2]

Received: 23 September 2018 / Revised: 24 January 2020 / Accepted: 13 July 2020 / Published online: 8 September 2020

Abstract

The study and analysis of past events can provide numerous benefits. While event categorization has been previously studied, it usually assigned only one event category to an event. In this study, we focus on multi-label classification for past events, which is a more general and challenging problem than those approached in previous studies. We categorize events into thirteen different types using a range of diverse features and classifiers trained on a dataset that has at least 50 labeled news articles for each category. We have confirmed that using all the features to train classifiers has statistical significance and improves all micro- and macro-average F_1, multi-label accuracy, average precision@5, area under the receiver operating characteristic curve and example-based loss functions.

Keywords Multi-label classification · Document classification · History · Event

1 Introduction

Study and analysis of past events can provide numerous benefits, including an enhanced perception of the legacies of the past in the present and enabling learners to make valuable connections through time [19,45,58]. Indeed, one of the goals of imparting recent history education at high schools is to enable students to study how people or organizations in history tried to solve problems in described events. Students can then apply this knowledge to consider creative solutions to social problems in present events [4,38]. In addition, there are many applications of this knowledge if we correctly understand event documents. For example, by being able to tell the categories of mentioned events one could better understand thanks to studying which past event types are mentioned in news articles. Equipped with knowledge on the categories of past event mentions one could also foster collective memory studies [1] as well as support search methods for finding historical events. Finally, the classification technique could be used for constructing thematic timelines or event lists (e.g., list of disasters/accidents in Asia, timeline of armed conflicts in the USA).

We focus in this work on the problem of *multi-label classification (MLC) for past events* that assigns more than one category to each event. For example, if we read the Wikipedia article[1] to know what the 2014 West Africa Ebola outbreak caused in our life, we can see that it killed many both human and animal (environment event), some researchers developed a vaccine (technology event), they then reported the details and their statistics (study event). Table 1 shows other examples of multi-labeled events.

The main challenge in MLC for past events lies in the scarcity of data, the ambiguity of expressions and variety of diverse means by which events can be referred to. Furthermore, frequently, in realistic scenarios, events are not called by their explicit names, or, they may have no known names.[2] Consequently, their automatic detection using named entity recognition (NER) tools is problematic. We make an assumption that the context of such documents (e.g., surrounding sentences in the original text) is not available to cover also the case of standalone documents like the lists of significant events in each month of the Wikipedia's Current Portal.[3] Hence, we rely only on the event document itself.

To provide sufficient data, we use a range of features based on lexical analysis as well as ones based on distributional

✉ Yasunobu Sumikawa
sumikawa-yasunobu@tmu.ac.jp

Ryohei Ikejiri
ikejiri@iii.u-tokyo.ac.jp

[1] Tokyo Metropolitan University, Tokyo, Japan

[2] The University of Tokyo, Tokyo, Japan

[1] https://en.wikipedia.org/wiki/West_African_Ebola_virus_epidemic.

[2] Usually, only very popular or important events have own names.

[3] https://en.wikipedia.org/wiki/Portal:Current_events.

Table 1 Example events. Our classifier takes documents of events; however, we include only short documents or names of events here for simplicity. The abbreviated category names are used: `Reign` (`Rg`), `Diplomacy` (`Dp`), `War` (`Wr`), `Production` (`Pr`), `Commerce` (`Cr`), `Study` (`St`), `Religion` (`Rl`), `Literature and Thought` (`LT`), `Technology` (`Tc`), `Popular Movement` (`PM`), `Community` (`Cn`), `Disparity` (`Ds`) and `Environment` (`En`)

Event	Categories
Agnes Chan named UNICEF Regional Ambassador for East Asia and Pacific Region	`Dp`, `Cn` and `LT`
The World Strikes a Deal on Climate Change	`En`
Paris attacks	`Dp`, `Rg` and `PM`
ISIS Terrorists Strike on Three Continents	`Dp`, `Rl`, `Wr` and `PM`
Same-Sex Marriage Debate	`LT` and `Cn`
Ebola outbreak	`En`, `St` and `Tc`
The Scottish independence referendum	`Rg`, `PM` and `Cn`

word representation using neural networks. Though there are several labeled event datasets on the Web such as Wikipedia's Current Portal, many of them assign only one category to each event. To perform MLC for events, we have created a new database and opened it on a public repository (Sect. 3). We use news articles that have one or more than two event categories in the created dataset to train our classifiers from the features.

We conducted experimental evaluations to confirm how well using all feature types improve classification accuracies on the created new database. We confirmed that our method achieved approximately 60% in the micro-average F_1 score that is the best among all compared methods. We also evaluated that this score is a statistically significant improvement from using each feature type to train classifiers. In addition, we performed other measurements (macro-average F_1, multi-label accuracy, average precision@5, hamming loss, log loss, ranking loss scores, and area under the receiver operating characteristic curve (AUC)), which are widely used in MLC studies, and confirmed that our method achieved the best scores of all compared methods.

Problem statement In our classification, each document describing an event may have more than one label. The formal definitions are given as follows: Let $\mathcal{L}$ be a finite and non-empty set of labels $\{l_1, l_2, \ldots, l_m\}$. Let $\mathcal{X}$ and $\mathcal{Y}$ be the input and the output spaces, respectively. Given a dataset $\mathcal{D}^l = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$, $y_{i_j} \in \{0, 1\}^m$, MLC predicts labels $\hat{y} = \{y_k \mid 1 \leq k \leq m\}$ for a document. The assigned labels are usually referred to as the relevant labels for the input document.

Definition of the past In this paper, we define *past events* as events that occurred before training classifiers.

Paper organization The remainder of this paper is organized as follows: Sect. 2 provides summaries of several related works. Section 3 describes the dataset this paper used. Next, the proposed method in this paper is described in Sect. 4. We perform experimental evaluations to confirm the effectiveness of the proposed method in Sect. 5 and then conclude remarks in Sect. 6.

2 Related works

As this study performs classification for past events, we summarize differences of this study from classification and history-related studies. In particular, this study explores how well it is able to create feature vectors to train classifiers; thus, we separately describe differences of past feature selection studies and training classifiers in Sects. 2.1 and 2.2, respectively.

We then focus on identifying differences of this study from history-related past studies by dividing this sub-section in events focused on classification, history education, collective memory, information retrieval and data mining in Sects. 2.3, 2.4, 2.5, 2.6 and 2.7, respectively.

2.1 Feature selections and extractions

To train classifiers, it is necessary to create feature vectors from documents. One of the simplest methods is to count the number of word occurrences in documents to set them as indexes of vectors corresponding to the documents. However, this simple method presents problems such as high computational complexity and overfitting as there are numerous kinds of words in documents leading to the high feature space dimensionality.

Semantic analysis such as latent semantic analysis (LSA), latent Dirichlet allocation (LDA) [2] and Doc2Vec [35] has become a popular way to reduce the dimensionality of feature space. Using semantic analysis creates feature vectors from topic distributions for all documents and trains classifiers from the feature vectors. Another popular method of dimensional reduction is to use statistical approaches [5,63] or mutual information [12,36,37].

Feature extraction is another widely applied method, for example principal component analysis (PCA) [72]. Gopal and Yang [17] defined meta-level features by transforming conventional representations of data and categories into a relatively small set of link-based features.

Similar to the case of past studies, we use word- and semantic-based feature vectors and reduce the dimensionality of feature space after combining all feature vectors. In addition to the popular feature types, we take temporal nouns that occur only for specific durations to replace them with their contexts (meaning the top-similar words for the nouns) as several events tend to include names of persons, groups and other entities. From our experimental results, we show that this replacement of temporal nouns with their contexts plays a key role in this objective.

2.2 Training classification methods

MLC algorithms typically consist of two steps: learning to rank categories for data and learning to put a threshold on whether each category is assigned or not. For the first step, the simplest method is to employ a binary-classifier, such as support vector machine (SVM), naive bayes probabilistic classifiers or random forests [67]. This binary-classifier-based method learns a scoring function for each category independently from the scoring functions for other categories, and then scores test data for each category independently.

Several studies utilize global optimization techniques. Elisseeff and Weston proposed Rank-SVM to maximize the sum of the margins for all categories at the same time whereas binary SVM maximizes the margin for each category independently [13]. As another method, the k-Nearest Neighbor is widely used for multi-label classification [10,70,73].

McCallum [41] trains a classifier by EM algorithm to take mixture models into account for the training because each data point of a category can be considered as it is generated from a distribution of the category. However, the EM algorithm is typically used to train classifiers as the semi-supervised learning (SSL) style [8,16,48,76]. SSL is a well-known procedure to train classifiers with decreasing manual category assignment costs in the case where few labeled data and many unlabeled data are acquired. For the other SSL-based classification type, graph-based methods are proposed [76]. In graph-based methods, data are represented as nodes and the similarity between two data points is represented as the weight of an edge.

In the task of categorizing short documents, data scarcity becomes a more severe problem than in long documents. To overcome this problem, some studies use contextual information. Sriram et al.'s [57] approach classifies tweets by using author information, URL and hashtags. Nie et al. [47] use a Naive Bayes classifier equipped with texts, image and video contents for Q&A classification. Lee et al. [39] classify queries using user-click behavior to identify user goals in web searches. On the other hand, using external information such as Wikipedia is also a popular approach. Zelikovitz and Marquez [71] train a classifier with LSA [11] based on Wikipedia data, and Phan et al. [53] propose a generalized classification framework with the topic model. This framework first trains the topic model on texts of an external resource. It then trains classifiers after building a moderate size labeled training dataset. Explicit semantic analysis (ESA) is applied in [61] to map short texts to Wikipedia articles.

Training classifier assumes that feature vectors are given; in other words, the training designs an algorithm to project feature vectors into categories. Since this paper proposes a novel feature selection method, this paper takes any algorithms of training models. Indeed, this paper applies SVM, Naive Bayes and graph-based methods that are described above.

2.3 Event classification

Kosmerlj et al. proposed event categories that were originally defined by Wikipedia editors and then investigated automatic classification using TF-IDF created from news articles [33]. Several events can be mentioned with a few sentences, such as news articles containing references to related events, historical accounts or biographies. Sumikawa and Jatowt propose a feature selection method to classify short documents of past events [59]. These studies propose classifying event document frameworks; however, they are designed as multi-class classification that assigns only one category to an event.

2.4 Analyzing history for education

Studying history is beneficial to understand how the present shapes, and it can be used to predict the future to some extent. Indeed, there are classes to learn history starting from elementary school in many countries. Recently, some historians, education researchers and national guidelines consider that it is important to support learners to connect past and present to examine what knowledge we can use about the past to consider solutions to present social issues. This ability is called *historical analogy*, and several researchers have studied effective and efficient methods for enhancing the historical analogy.

Drie and Boxtel [66] find that there are basic components to enhance historical reasoning. Mansilla examined what the triggers for successfully using historical analogy are [3]. Lee proposes a framework that makes connections between events in the past and potentially between events in the past and present [38]. This framework is an overview of long-term change patterns, and an open structure capable of being modified, tested, improved and even abandoned. Ikejiri designed a competitive card game [22] where players construct causal relations for an event. From the construction, players can identify causal relationships within modern societal issues and compare how two past and present issues are similar from a viewpoint of causal relationship structures to stimulate historical analogy. In addition, Ikejiri et al. propose another competitive card game in which players learn economic policies that are actually enforced in the past to create new policies that would revitalize Japan's economy [23].

2.5 Analyzing collective memories

The concept of *collective memory* (*social memory*) popularized by Halbwachs [18,20] describes the shared reflection of the past within social groups. Collective memory can be contrasted with the concept of *collective amnesia* defined by Jacoby [26] as forceful or unconscious suppressions of memories, especially those related to disgraceful or inconvenient events.

Traditionally, research on collective memory has been based on small-scale investigations of personal accounts and the activities of political and cultural institutions. Recently, some researchers have used computational approaches for the quantification of the characteristics of social memory over large text datasets. Cook et al. [9] investigated the decay of fame over time based on a collection of news articles spanning the twentieth century. Au Yeung and Jatowt [1] have studied memory decay and the way in which past years are remembered based on the dataset of English news articles spanning 90 years. When it comes to other document genres, Ferron and Massa [14] and Kanhabua et al. [32] proposed using Wikipedia as a global memory space. There are several literature reports using Wikipedia to study collective memory [15,28]. Social media has been commonly utilized to study public attitudes toward real-time events such as the US American elections [65]. Microblogs are popular tools for sharing and finding information related to the past as well as media. There is an ongoing project that focuses on the First World War [7] and compares commemorative cultures across countries. Sumikawa et al. [60] attempt to fill this gap by focusing on Twitter as a common social media platform frequently used in computational social science. This study is exploratory and aims to provide an initial broad investigation of history-related content sharing in social networks.

2.6 History-related information retrieval

The current Web has numerous digital archives including historical images and documents, as results of digitization. For effective searching for what users want, searching past objects is becoming popular research topic to aid historians [50,56].

Singh et al. [56] proposed an IR framework to support historians. According to the literature, if historians investigate an entity, they first try to see a big picture of the entity. Then, they further search the entity for specific aspects. Thus, it is important for historians' information seeking behavior to show not only important time windows but also different aspects. Zhang et al. [74] propose a framework that detects entities counterparts over time. The framework bridges two different vector spaces that are created for different time-ranges such as [1900–1950] and [1960–2010] by transformation matrix, which maps an entity from one vector space into another; thus, this framework takes a word. They extend this framework to make use of hierarchical cluster structures [75]. Searching images related to history is also proposed [6].

2.7 Mining history-related knowledge

Growing the number of digital archives gives us studies that extract beneficial history-related knowledge, such as evaluating the significance of historical entities[62], the importance of historical persons [29]and semantic change of words [27]. These studies try to find beneficial information from a large amount of data. Considering mining the past, there is literature to add useful information, such as timestamps to entities [30], analysis trends [21] and for trying to predict future from past events [31,54,55].

3 Data collection

3.1 Event categories

This paper uses thirteen categories defined in [24,25] in order to connect past and present events. The event categories are: `Commerce` (`Cr`), `Diplomacy` (`Dp`), `Production` (`Pr`), `Reign` (`Rg`), `Environment` (`En`), `Religion` (`Rl`), `Disparity` (`Ds`), `Study` (`St`), `Community` (`Cn`), `Literature and Thought` (`LT`), `Popular Movement` (`PM`), `Technology` (`Tc`) and `War` (`Wr`). These categories are defined by Encyclopedia of Historiography [51]. Table 1[4] shows example events for the 13 categories.

3.2 Datasets

In this paper, we use news articles describing events, which were published by Japanese companies including NHK news and Mainichi news.[5] These articles typically have enough words for classification; however, most news articles are assigned categories defined by their companies. Thus, they are usually different from the above 13 event categories. To train our classifiers, we manually assigned more than one event category from the list to several news articles. The assignment processes were done by two Japanese researchers working on history education research and HistoInformatics. They all have Ph. D. degrees; therefore, the dataset is created

[4] We use Japanese news articles to evaluate classifications in this paper as described in Sect. 5. Even though we did not use the listed example events in the evaluation, we show them to aid understanding what kinds of events can be assigned to from the 13 categories.

[5] Some articles are stored in CD-Mainichi Newspapers 2012 data, Nichigai Associates, Inc., 2012 (Japanese). The others are collected by Web crawling.

Table 2 Statistics of the whole dataset

Num. of categories	13
Num. of labeled articles	435
Ave. length	1641.8
Ave. num. of categories per document	2.6
Ave. num. of document per category	87.2

Table 3 Statistics for each category. This table shows the numbers of labeled data for a category c ($\mathcal{D}_c^l$)

Cat.	Num. of $\mathcal{D}_c^l$	Cat.	Num. of $\mathcal{D}_c^l$
Cr	179	St	59
Dp	187	Cn	60
Pr	108	LT	55
Rg	93	PM	52
En	69	Tc	77
Rl	50	Wr	59
Ds	86		

by experts. This new ground truth dataset has been opened on a public repository.[6]

3.3 Statistics of dataset

Table 2 shows the dataset statistics. We have prepared 435 labeled articles from Web crawling and the Mainichi news dataset.

We show the number of articles per category in Table 3. For each event category, there are at least 50 labeled articles.

4 MLC for past events

For classifying past events, our algorithm first creates effective feature vectors to train multi-label classifiers (Sect. 4.1). It then trains the classifiers including probabilistic, non-probabilistic and graph-based ones (Sect. 4.2).

Algorithm 1 shows an overview of classifier training for past events. First, this algorithm applies *Preprocess* to create tokens after removing stop words and stemming. As several natural languages such as Japanese and Korean do not distinguish words by spaces, it is necessary to apply morphological analysis to divide words during this preprocess. Since this paper uses Japanese documents in experiments, we apply MeCab [34] as a morphological analysis.

Algorithm 1 Algorithm of multi-label classification for past events.

```
Input: A set of labeled documents $docs_l$, a set of unlabeled documents $docs_u$ and a set of labels for the labeled documents $l$
Output: A classifier $FVecs$
Function $MLCPastEvents(docs_l, docs_u, l, k)$
    $tokens_l, tokens_u = Preprocess(docs_l), Preprocess(docs_u)$
    $models = MakeModels(tokens_l, tokens_u)$
    $FVecs_l = FeatureSelection(tokens_l, models, k)$
    $FVecs_u = FeatureSelection(tokens_u, models, k)$
    $clf = TrainClassifier(FVecs_l, FVecs_u, l)$
    return $clf$
Function $Preprocess(docs)$
    $tokens = \emptyset$
    for $d \in docs$
        $tlist = SplitWords(d)$ // If it is necessary, morphological analysis is applied here
        $validtlist = RemoveStopWords(tlist)$
        $tokens.add(validtlist)$
    return $tokens$
Function $MakeModels(tokens_l, tokens_u)$
    $lda, lsa, d2v, w2v = TrainSemModels(tokens_l, tokens_u)$
    $dr = TrainDimReductModel(tokens_l)$
    return $lda, lsa, d2v, w2v, dr$
```

Algorithm 2 Feature Vector Creation

```
Input: A set of documents $docs$ and a set of models $models$ used in semantic-based feature vectors
Output: A set of feature vectors $FVecs$
Function $FeatureSelection(tokens\_list, models, k)$
    $FVecs = \emptyset$
    for $tokens \in tokens\_list$
        // Word-based features
        $v_1 = WordFVecs(tokens, tokens\_list)$
        // Semantic-based features
        $v_2, v_3, v_4 = SemanticFVecs(tokens, models)$
        // Noun-context-based features
        $v_5 = NounContextFVecs(tokens, tokens\_list, k, models)$
        $fvec = DimReduction(v_1, v_2, v_3, v_4, v_5)$
        $FVecs.add(fvec)$
    return $FVecs$
```

The algorithm then performs *MakeModels* to train the LDA, LSA, Doc2Vec and Word2Vec models on the whole dataset to create semantic-based feature vectors described in Sect. 4.1.2. In *MakeModels*, dimensional reduction models are also trained on the labeled documents. Then, it applies the function *FeatureSelection* to create feature vectors by applying the dimensional reduction models. As this paper performs SSL training for graph-based classifiers using unlabeled documents ($docs_u$), this algorithm describes how to use $docs_u$. If the training method is supervised learning instead of SSL, the $docs_u$ should be $\emptyset$ and the training classifier function (*TrainClassifier*) does not use it to train classifiers.

4.1 Feature selection

In this sub-section, we describe how our approach creates feature vectors to train classifiers. At the beginning, Algorithm

[6] https://doi.org/10.5281/zenodo.3258150. This opened dataset excludes all texts of the articles to respect copyright law. However, it is possible to obtain the texts because the opened dataset includes event IDs defined in Mainichi Newspapers 2012 data or URLs used to Web crawling. Thus, after buying Mainichi Newspapers 2012 data or recrawling the URLs with Wayback Machine (the accessed day is 18 June, 2019), their corresponding texts can be retrieved.

2 shows the overview of the feature selection. First, this algorithm creates three types of feature vectors. It then combines these feature vectors to be a single feature vector. Simply combining the vectors increases the number of dimensions leading to *the curse of dimension*, it finally performs dimensional reduction to the combined feature vectors. Once these processes are applied to tokens for all documents, this function returns the results of applying dimensional reduction methods. In the remainder of this sub-section, we describe how to produce each type of feature. To help understanding how and why we create these features, we use actual texts of Web news that are used in our experimental evaluation as examples to see how to create feature vectors. The texts are:

1. D_1: イギリスでＥＵ＝ヨーロッパ連合からの離脱協定案を批判する閣僚の辞任が相次ぐなど政局が混乱する中、メイ首相は新しい離脱担当相を任命するなど政権の立て直しを急いでいます。ただ離脱派の閣僚が協定案の修正を迫る構えだとも伝えられ、メイ政権がＥＵとの正式合意にこぎつけられるか、見通しは立っていません。イギリスでは、ＥＵとの間で取りまとめた離脱協定案を批判する離脱担当相はじめ閣僚など４人が相次いで辞任するなど、政局の混乱が続いています。メイ首相は16日、ラジオ番組に出演して、国民からの質問に直接答える形で離脱協定案の支持を訴えました。その後、新しい離脱担当相にスティーブン・バークレー氏を任命し、政権の立て直しを急いでいます。こうした中、離脱派を代表するフォックス国際貿易相は「協定案に反対している議員は合理的な判断をすべきだ。なんらかの合意があるほうが何もないよりはましで、いまは経済界に安定を与えることが国益にかなうことだ」と述べ、首相を支持する姿勢を示しました。また、辞任の可能性が取り沙汰されていたゴーブ環境相も「首相を支持する」と述べ、閣僚の辞任の連鎖は、ひとまず止まった形です。ただ、ゴーブ氏をはじめ離脱派の閣僚は週末にも会合を開き、離脱協定案を修正するよう首相を説得する方策を探るとも伝えられ、メイ政権がＥＵとの正式合意にこぎつけられるかどうか見通しは立っていません。[7]
2. D_2: 安倍総理大臣は政府の会議で、ことしのノーベル医学・生理学賞に選ばれた京都大学特別教授の本庶佑さんから、若手研究者などへの支援の重要性を指摘されたことを踏まえ、関係閣僚に対し、支援策の具体化に取り組むよう指示しました。総理大臣官邸で開かれた政府の「総合科学技術・イノベーション会議」で、ことしのノーベル医学・生理学賞に選ばれた京都大学特別教授の本庶佑さんが講演し、日本は基礎研究費と若手研究者の数が減少しており、政府による支援が必要だと指摘しました。これに対し、安倍総理大臣は「本庶先生から、基礎研究の重要性について大変率直なお話をうかがい、改めて国として若手研究者に挑戦の機会を作ることの重要性を強く認識した」と述べました。そのうえで、安倍総理大臣は「政府の科学研究費補助金を、若手研究者へ重点的に配分することなどを着実に実行してほしい」と述べ、関係閣僚に対し、来年度予算案の編成作業などを通じて、基礎研究や若手研究者への支援策の具体化に取り組むよう指示しました。会議のあと、本庶さんは記者団の取材に対し「科学的な力がない国は将来性がなくるので、次の世代の研究者を育てることが重要だ」と話していました。[8]

The following English texts are translations of the above Japanese texts.

1. D_1: With the government in a state of confusion, such as the resignation of cabinet ministers criticizing the European Union (EU)'s draft withdrawal agreement, Prime Minister May hastened to rebuild her administration, including appointing a new Secretary of State for exiting the European Union. It is also reported that the ministers who are in a position to criticize the draft agreement on withdrawal are willing to make amendments to the draft agreement, and there is no prospect of the May administration being able to form a formal agreement with the EU. In the UK, there has been confusion over the political situation. It includes the resignation problem of the ministers who are in a position to criticize the draft agreement on withdrawal with the EU and four other ministers one after another. Prime Minister May appeared on a radio program on the 16th and appealed for support for the withdrawal agreement by answering questions from the public directly. Later, she appointed Stephen Berkeley as the new Secretary of State for exiting the European Union and is in a hurry to rebuild her regime. Under these circumstances, Fox, the International Trade Minister, who

[7] https://www3.nhk.or.jp/news/html/20181122/k10011720261000.html accessed on 22 Nov. 2018.

[8] https://www3.nhk.or.jp/news/html/20181117/k10011714161000.html accessed on 17 Nov. 2018.

represents the withdrawal party, said that *Legislators who oppose the draft agreement should make reasonable judgments. It is better to have some kind of agreement than to have nothing, and now it is in the national interest to provide stability to the business community.* In addition, Gove, the Minister of the Environment, whose possibility of resignation was discussed, also stated that he supports the prime minister, and the chain resignation of ministers has been halted for the time being. However, Gobe and other ministers who are in a position to criticize the draft agreement on withdrawal will also meet on the weekend to find ways to persuade the Prime Minister to amend the draft withdrawal agreement, and it is forecasted that the May administration will be in agreement with the EU.

2. D_2: Prime Minister Abe directed ministers to work on the implementation of supporting young researchers at the government council as Professor Honjo, a special professor at Kyoto University who was selected for the Nobel Prize in Medicine and Physiology, highlighted the importance of supporting young researchers. Professor Honjo gave a presentation at the government's conference called General Science, Technology and Innovation Conference held at the Prime Minister's Office in order to suggest that both amounts of national research funds for fundamental researches and the number of young researches have been decreasing and therefore government support is needed. Prime Minister Abe stated *Professor Honjo gave a very candid talk about the importance of fundamental research and reaffirmed the importance of creating opportunities for young researchers to challenge as national support.* On that basis, Prime Minister Abe stated, *We want the government to carry out steady distribution of scientific research grants to young researchers*, and told related ministers to work on the budget proposal for the next fiscal year according to the presentation given by Professor Honjo. After the meeting, Professor Honjo told reporters that *It is important to bring up the next generation of researchers, as countries without scientific power have no future.*

4.1.1 Word-based features

First, we create TF-IDF vectors (v_1) from all the documents to measure similarity based on their terms. Each element of this vector is a TF-IDF score for a word indicating how important the element is to a document in the dataset. This score is a multiplication of term frequency and inverse document frequency. The term frequency means how frequently each term (word) occurs in each document whereas the inverse document frequency represents how rarely each term occurs in whole documents. The formal definition is given as follows:

$$\text{TFIDF}(w, d, \mathcal{D}^l) = tf_{w,d} * \frac{|\mathcal{D}^l|}{|\{d' \in \mathcal{D}^l \mid w \in d'\}|} \tag{1}$$

where $tf_{w,d}$ is the number of times a word w occurs in a document d, $|\bullet|$ is the size of $\bullet$. The second term of this equation gives the number of all labeled data divided with labeled data including w.

Algorithm 3 Word-based feature vector creation

```
Input: A token list for a document $t$ and a set of token list for all documents $tokens$
Output: A tf-idf vector
Function $WordFVecs(tokens, tokens\_list)$
    $vec = [0.0, \ldots, 0.0]$ // its size equal to the size of a set of tokens.
    for $w \in tokens$
        $vec_w = TFIDF(w, tokens, tokens\_list)$
    return $vec$
```

Algorithm 3 shows how to use Eq. 1 for a given document. As a document is already converted to a token list, this algorithm calculates the TF-IDF score for each word list. In the algorithm, vec_w represents an index for a word w.

Table 4 lists words that are portions of the results of applying MeCab and removing stop words to the two actual example texts (D_1 and D_2). As word-based features use all words to make feature vectors, the dimensionality of this feature's type tends to be high, leading to sparse vectors. In our experiments, there are 24,594 words; thus, the dimensionality of v_1 is 24,594.

4.1.2 Semantic-based features

Next, we create feature vectors by applying Doc2Vec (v_2), LSA (v_3) and LDA (v_4) models to capture latent semantic text structures. These three models capture the latent semantics, but their algorithms differ. Doc2Vec is a neural network-based algorithm. LSA performs a matrix decomposition (SVD) on the term-document matrix whereas LDA is a probabilistic model assuming a Dirichlet prior over the latent topics. As shown in Algorithm 4, these feature vectors are simply created from paragraph vectors by applying *D2vFeature* and topic distributions of LSA and LDA by *LSATopicDist* and *LDATopicDist*, respectively. Note that $models_A$ represents a model of A.

Algorithm 4 Semantic-based vector creation

```
Input: A token list for a document $t$ and a set of token list for all documents $tokens$
Output: Doc2vec, lda and lsa feature vectors
Function $SemanticFVecs(tokens, models)$
    $v_2 = D2vFeature(tokens, models_{d2v})$
    $v_3 = LSATopicDist(tokens, models_{lsa})$
    $v_4 = LDATopicDist(tokens, models_{lda})$
    return $v_2, v_3, v_4$
```

Table 4 Words from example texts. The listed words are sub-sets of words of the example

D_1	イギリス(UK), ヨーロッパ(Europe), 離脱(break from), ..., メイ(May), 首相(the Prime minister), 新しい(new), ..., 迫る(are willing to make), 質問(questions), 直接(directly), 答える(answer), 合意(agreement), 見通し(forecasted).
D_2	安倍(Abe), 総理(the Prime), 大臣(minister), 政府(Government), 会議(council), ノーベル(the novel prize), ..., 教授(professor), 講演(presentation), 研究(research), 支援(support), 踏まえ(according to), 取り組む(work).

Applying Algorithm 4 for tokens that are results of *Preprocess* generates feature vectors whose dimension is 100 as we set the value as a parameters in all the semantic analysis (LDA, LSA, Doc2Vec and Word2Vec). As each element of these vectors indicates a distribution of topics, these vectors tend to be dense compared with the word-based feature.

4.1.3 Noun-context-based features

The objective of this feature type is to smooth the temporal effects of words used for specific temporal durations. In this paper, we call this kind of word *a temporal word*. As this study classifies texts of events, names of persons, events and groups are often used in documents such as, who does what, locations where the events occurred and so on. Since some words occur for specific durations, our algorithm replaces temporal words with semantically equivalent words that are commonly used for all durations. For example, two texts *Abraham Lincoln won the election* and *Donald Trump won the election* can be targets of this study. If it is possible to identify that both *Abraham Lincoln* and *Donald Trump* can be replaced with *the president*, then the two texts become completely the same texts.

In this paper, we apply a simple strategy that replaces all nouns with their semantically similar words. From this, temporal words can be removed by the replacements. This study focuses on nouns as this type of word plays a key role in distinguishing event categories. For example, diplomatic events tend to include names of politicians whereas commercial events frequently mention production items. To perform the replacement, we use two assumptions: (1) words that are frequently used together are semantically similar to each other and (2) meanings of frequent terms (such as president, the prime minister, propose, accident and cause) can be the same at different points in time. The first assumption is widely used in natural language processes such as LSA and Word2Vec. This study uses this assumption to locate words for replacement. The second assumption uses the observation that it is hard for the dominant meaning of frequently used words to change in several languages [40,52]. For example, the president tends to be used in political activities such as nomination, defeat, serve, party, vote and congress. Indeed, these words are commonly used in Wikipedia articles for Abraham Lincoln and Donald Trump. Thus, this study uses the second assumption as a reason for why our algorithm performs the replacements to fill in the gap of temporal words.

The simple strategy may perform replacements for non-temporal words even though the processes are unnecessary for them. However, this strategy has two benefits. First, it does not require any additional analysis to identify whether each word is temporal or not. Second, it is beneficial to reduce the sparsity of created feature vectors as the replacement should increase the number of words describing events.

Figures 1, 2, 3 and 4 show the top-5 words for two temporal and two non-temporal nouns from the two example texts (D_1 and D_2). These two texts are related to government policies for the UK and Japan; therefore, there are two names of the politicians (*May* and *Abe*). Once their terms as the Prime Minister expire, and different persons contribute as the Prime Ministers, the two names may not be used in event description frequently. Thus, once the noun (*May*) is replaced with the 5 words displayed in Fig. 1 (postpose, leaving, defeat, withdrawal and "form a Government"[9]) that represent what the noun tries to accomplish, it is possible to use words widely appearing in political event descriptions instead of the temporal words. Figure 3 also shows similar effects of the replacement for *Abe*; it is able to use the 5 words representing which party he belongs to (*Liberal Democratic Party*), which position he is in the party (*president*), what is the objective of his policy (*deflation* and *break out*) instead of using the temporal word. Figures. 2 and 4 show the top-5 words for two non-temporal words (council and research). We can see that there are no temporal words in the top-5s; therefore, the replacement reduces sparsity without changing the semantics of the word.

One possible concern of this replacement is that it is possible to include temporal words in the top-k such as *Tanigaki*, which is the name of a person, as shown in Fig. 3. However, the figure shows that the replacement that incorporates 4 non-temporal ones that contribute to improvement share common words with other documents. Indeed, the non-temporal words should have strong similarity within non-temporal words as they can be used on all temporal durations. Although there is a concerning situation, our experimental results show that

[9] In Japanese, this term can be represented as a word.

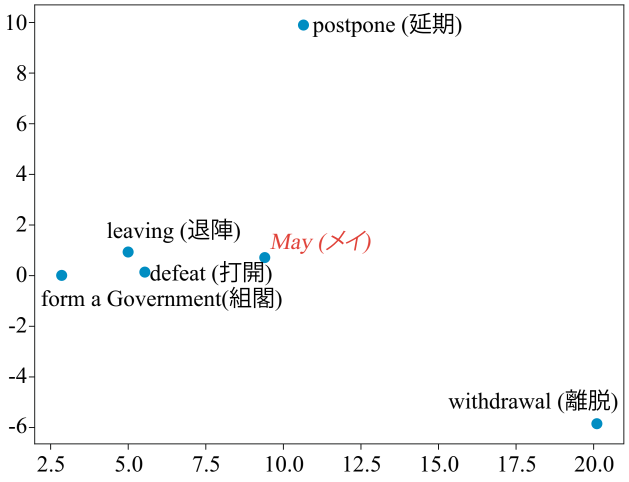

Fig. 1 Example top-5 for May. This figure plots top-5s for two nouns (person and non-named entity)

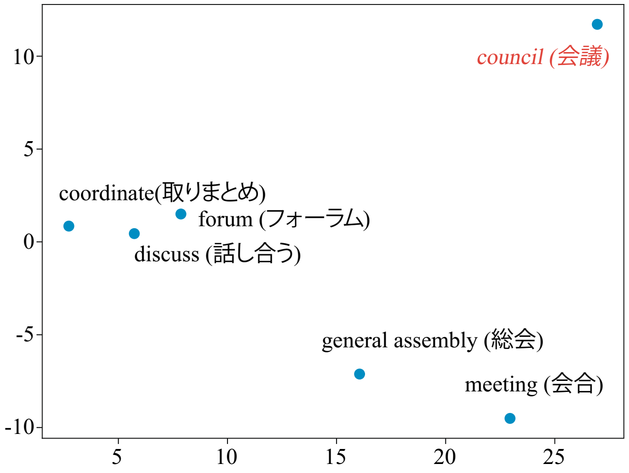

Fig. 2 Example top-5 for council. This figure plots top-5s for two nouns (person and non-named entity)

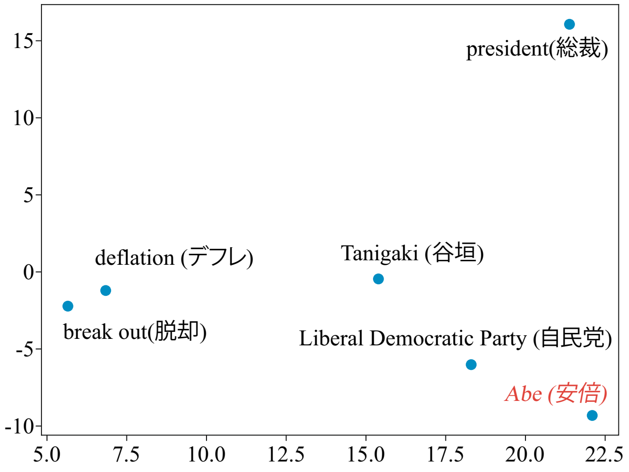

Fig. 3 Example top-5 for Abe. This figure plots top-5s for two nouns (person and non-named entity)

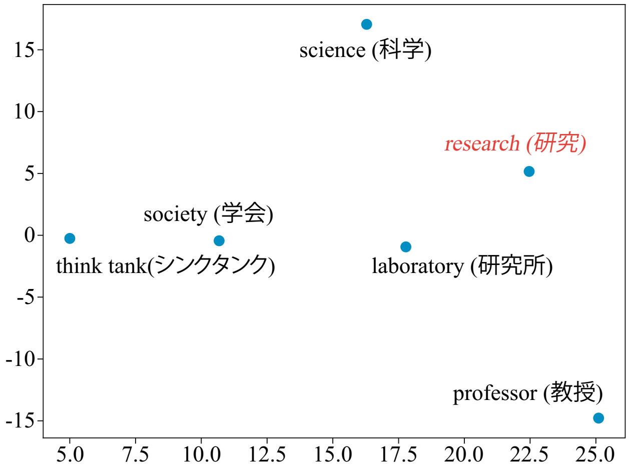

Fig. 4 Example top-5 for research. This figure plots top-5s for two nouns (person and non-named entity)

this type of feature contributes to improving classification accuracy.

To capture the semantic similarity of words, we perform word embedding through the Skip-gram model [42–44]. Since this technique assigns vectors to words so the closer their meaning, the greater similarity they indicate, we replace all nouns in documents with their top-k closest words on the vectors. Algorithm 5 shows how to create this type of feature vector. After training the Skip-gram model, the proposed method locates top-k similar words for each noun. It then replaces the nouns by their top-ks to create TF-IDF vectors (v_5) from the replaced words.

Algorithm 5 Noun-context vector creation

Input: A token list for a document t and a set of token list for all documents $tokens$

Output: A noun-context-based feature vector

```
Function $NounContextFeature(tokens, tokens\_list, k, models)$
  $new\_tokens = []$
  for $n \in Noun(tokens)$
    $new\_tokens \leftarrow TopSimWords(n, k, , models_{w2v})$
  return $TFIDFFeature(new\_tokens, tokens\_list)$
```

4.1.4 Combining feature vectors

Finally, we combine all the feature types and then perform dimensional reduction to avoid sparsity. Let s_i be a size of the ith feature vector. For each document, we create 5 feature vectors ($v_1, v_2, \ldots, v_5$), and then combine them to form a feature vector; therefore, the size of a combined feature vector is $s_1 + s_2 + \cdots + s_5$. For the combined feature vectors, we apply a method of dimensional reduction. In this paper, we train the following three popular methods of dimensional reduction on labeled data:

1. L1 Norm Regularization (L1): This method trains the linear model penalized with the L1 norm, and then selects the non-zero coefficients.
2. Random Forests (RFs): This method calculates the importance for each feature and discards irrelevant features according to the values of importance.
3. PCA: This method decomposes a multivariate dataset in a set of successive orthogonal components that explain the maximum amount of variance.

4.2 Classification

In this paper, training classifiers is performed on the results of combined feature vectors. Since this paper focuses on designing a feature selection method for effective multi-label event classification, this paper implemented three popular classifiers: naive bayes (NB), random forests (RFs), and SVMs with RBF or linear kernels. These classifiers are trained as one-vs-rest classification.

In addition, this paper trains the following graph-based classifiers to estimate how well utilizing correlation between labels works on the proposed feature selection scheme.

1. Label propagation (LP): LP is a graph-based SSL classification algorithm [76]. This algorithm employs cluster assumption meaning that similar nodes tend to have common labels to calculate scores for assigning categories. This calculation is performed by iteratively multiplying label scores with similarities between nodes.
2. Dynamic LP (DLP): DLP is an extension of LP to take label correlation [68].
3. LP through linear neighborhoods (LNP) [69]: LNP is an extension of the LP algorithm to efficiently construct graphs by applying KNN to incorporate similarity of linear neighborhoods into a probability matrix.
4. LP using amendable clamping (LPAC) [46]: LPAC is an extension of the LP-based algorithm. LPAC is originally designed for the label completion task of MLC by emphasizing the cluster assumption; however, this algorithm achieves better than traditional classifiers on a simple MLC task. We use LPAC as a baseline in this study.

As these classifiers perform SSL, we additionally prepared 9000 unlabeled news articles from CD-Mainichi Newspapers 2012 data.

5 Experimental results

5.1 Parameter settings

In this paper, we set 100 as parameters of LDA, LSA, Doc2Vec and Word2Vec and 5 as k for creating v_5. We use implementations of LDA, LSA, Doc2Vec and Word2Vec from gensim.[10]

5.2 Evaluation criteria

There are several methods to measure MLC performances from several different points of view. Usually, these performances are measured by two kinds of methods: label-based measures and example-based loss functions [64]. The label-based measures decompose the evaluation with respect to each label whereas the example-based loss functions compute the average differences of the actual and the predicted sets of labels over all examples.

For the label-based measurement, we use micro- and macro-average precision, recall and F_1 score. Formal equations of micro-average precision, recall and F_1 score are defined as follows:

$$miP = \frac{\sum_i \mathrm{TP}_i}{\sum_i (\mathrm{TP}_i + \mathrm{FP}_i)} \tag{2}$$

$$miR = \frac{\sum_i \mathrm{TP}_i}{\sum_i (\mathrm{TP}_i + \mathrm{FN}_i)} \tag{3}$$

$$miF_1 = \frac{2\ miP\ miR}{miP + miR} \tag{4}$$

where TP, FP and FN are true positive, false positive and false negative, respectively. The precision is defined as the proportion of predicted labels that are truly relevant. The recall is defined as the proportion of truly relevant labels that are included in predictions. The trade-off between precision and recall is formalized by their harmonic mean, called the F_1 score.

These micro-average measurements calculate metrics globally by counting the total true positives, false negatives and false positives. In contrast, the macro-average measurements treat all classes equally; in other words, they compute the metrics independently for each class and then take the average. The formal definitions of macro-average precision, recall and F_1 are as follows:

$$maP_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{5}$$

$$maR_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{6}$$

$$maF_1 = \left(\sum_i \frac{2\ maP_i\ maR_i}{maP_i + maR_i} \right) / \mid \mathcal{L} \mid \tag{7}$$

[10] https://radimrehurek.com/gensim/models/ldamodel.html, https://radimrehurek.com/gensim/models/lsimodel.html, https://radimrehurek.com/gensim/models/doc2vec.html and https://radimrehurek.com/gensim/models/word2vec.html.

Further, we use average precision at K (ap@K), which is one of the most popular metrics in information retrieval. This metric corresponds to average precision among the top K documents. The formal equation of average precision is defined as follows:

$$\text{Precision}(k) = \frac{1}{k}\sum_{i}^{k} r_i \tag{8}$$

$$\text{Average}P = \frac{1}{\mid \mathcal{D}^t \mid}\sum_{k<N} r_k \text{Precision}(k) \tag{9}$$

where r_i represents whether the prediction is correct or not by using 1 (correct) or 0 (wrong), $\mathcal{D}^t$ is a set of test data and N is the last rank where a classifier assigns a correct label to the test data.

In addition, for multi-label accuracy, we use the Jaccard index-based measurement (MA) and area under the receiver operating characteristic curve (AUC). The MA measurement calculates a score for the dissimilarity between two sets by dividing the difference of the sizes of the union and the intersection of the two sets with the size of the union. The formal definition is given as follows:

$$\text{MA} = \frac{1}{N}\sum_{i}^{N}\frac{|y_i \wedge \hat{y}_i|}{|y_i \vee \hat{y}_i|} \tag{10}$$

where $\hat{y}_i$ is a label predicted by a classifier.

AUC is one of the most important metrics for evaluating classifier models. This metric measures the area under the receiver operating characteristics (ROC) curve that represents the relationship between true positive and false positive rate.

In the case where we use these metrics, the higher scores they are, the better the evaluated classifiers are.

As for the example-based loss functions, hamming loss (HL), ranking loss (RL) and log loss (LL) are popular measurements in MLC. HL calculates the fraction of the wrong labels to the total number of labels. RL means a proportion of pairs of labels that are not correctly ordered. Finally, LL calculates scores from probabilistic confidence. This metric can be seen as a cross-entropy between the distribution of the true labels and predictions. Their formal definitions are given as follows:

$$\text{HL} = \frac{1}{NL}\sum_{i}^{N}\sum_{l}^{L}[\![y_{i,l} \neq \hat{y}_{i,l}]\!] \tag{11}$$

$$\text{RL} = \frac{1}{N}\sum_{i}^{N}\sum_{y_j > y_k}\left([\![\hat{y}_i < \hat{y}_j]\!] + \frac{1}{2}[\![\hat{y}_i = \hat{y}_j]\!]\right) \tag{12}$$

Table 5 Feature selection accuracies. Micro-average F_1 scores of SVM-Lin. for different feature selection methods

	PCA (%)	RFs (%)	L1 (%)
NB	30.0	57.1	**60.2**
RFs	28.8	**55.2**	50.3
SVM-Lin.	56.4	57.1	**57.2**
SVM-RBF	**58.0**	56.8	56.8

The bold-faced numbers indicate the best score for a classifier over the three-dimensional reduction methods

$$\text{LL} = -\sum_{i}^{L} y_i \log(p_i) \tag{13}$$

In these measurements, the smaller the scores, the better the model performances.

We calculate all the above scores by averaging of 5-fold cross-validation.

5.3 Discussions of accuracies

In the remainder of this section, we discuss the results of MLC for past events. First, Sect. 5.3.1 compares accuracies for all dimensional reduction methods to fix methodologies used in the following discussions. Then, Sect. 5.3.2 investigates how well the proposed feature vector creation correctly predicts categories. Section 5.3.3 performs error analysis to better understand why each classifier performed mis-predictions. Finally, Sect. 5.3.4 analyzes which feature types contribute to the improvement in the proposed method.

5.3.1 Accuracies of dimensional reductions

> **Q.** Which dimensional reduction method was the best for icro-average F_1?
> **A.** The L1-based dimensional reduction method exhibited the best micro-averaging performance for F_1.

Table 5 shows the micro-average F_1 scores of NB, RFs, SVM-Lin and SVM-RBF, which are trained on feature vectors created by the three different feature selection methods. We can see that L1-based feature selection for NB obtained the best score; therefore, we show results of classifiers using this method in the following section.

5.3.2 Accuracies of feature vectors creations

> **Q.** Which classifier was the best for micro-average F_1 scores?
> **A.** The NB equipped with all features demonstrated the best micro-average F_1 scores.

Table 6 $mi F_1$ of baselines. Scores for NB and LP-based algorithms obtained when using individual feature groups and TF-IDF, respectively

Category	TF-IDF (v_1) (%)	Doc2Vec (v_2) (%)	LSA (v_3) (%)	LDA (v_4) (%)	Noun-context (v_5) (%)	LP (%)	DLP (%)	LPAC (%)	LNP (%)
Cr	23.4	64.7	66.1	56.1	69.7	58.2	59.7	70.7	66.1
Dp	53.4	71.6	60.5	65.6	73.8	59.9	64.1	71.2	67.7
Pr	0.0	53.4	45.0	31.4	38.9	39.2	41.8	45.9	46.0
Rg	0.0	20.1	13.1	4.8	7.6	23.9	26.0	26.4	27.8
En	1.9	46.8	41.3	29.7	5.5	26.2	35.8	59.6	45.5
Rl	0.0	50.1%	40.0	44.5	26.0	20.6	22.4	53.7	43.5
Ds	8.8	**60.4**	60.1	24.9	28.6	32.5	22.4	35.6	47.2
St	6.4	36.6	50.0	19.7	13.5	23.8	22.4	33.3	42.3
Cn	0.0	31.0	31.0	15.8	27.5	35.0	31.1	45.2	44.3
LT	0.0	36.2	**48.1**	22.8	25.3	22.1	16.8	0.0	19.5
PM	0.0	38.5	35.7	12.8	0.0	21.1	29.4	**53.3**	40.1
Tc	2.5	52.5	45.3	36.8	14.9	29.8	31.1	42.1	43.9
Wr	0.0	53.2	35.9	48.2	14.8	23.8	29.4	44.1	40.4
Total	16.5	52.9	50.1	40.8	43.1	33.4	34.5	48.9	48.4

The bold-faced numbers indicate the best for a particular term given the metric

Table 7 miF_1 of proposed methods. Scores when using all features used together for NB, RFs, SVM and LP-based algorithms settings for each class

Category	All+NB (%)	All+RFs (%)	All+SVM-RBF (%)	All+SVM-Lin. (%)	All+LP (%)	All+DLP (%)	All+LPAC (%)	All+LNP (%)
Cr	**75.8**	69.9	73.0	71.4	65.1	58.2	64.9	60.6
Dp	75.8	69.6	**78.0**	76.4	73.8	59.9	74.1	69.2
Pr	**60.6**	49.0	39.3	36.2	52.4	39.2	51.7	50.3
Rg	**31.3**	22.1	4.4	0.0	26.2	23.9	26.0	26.2
En	54.4	25.2	58.4	**59.7**	30.3	26.2	30.3	37.8
Rl	**60.1**	57.1	52.1	35.3	59.5	20.6	59.5	55.8
Ds	57.4	45.1	56.6	52.1	46.9	32.5	46.9	44.9
St	52.8	41.5	**53.4**	52.7	38.9	23.8	38.9	41.4
Cn	**57.3**	33.0	34.6	30.7	35.7	35.0	37.7	38.0
LT	39.5	18.8	31.6	15.9	16.6	22.1	16.6	20.3
PM	42.4	24.3	17.7	19.2	40.2	21.1	40.2	40.5
Tc	**58.6**	34.1	50.0	35.6	40.1	29.8	40.1	38.7
Wr	**54.8**	35.0	35.5	26.7	50.7	23.8	50.4	41.6
Total	**60.2**	50.3	56.8	57.2	52.6	33.4	52.5	49.3

The bold-faced numbers indicate the best for a particular term given the metric

The micro-average F_1 scores for all baselines and our approaches are shown in Tables 6 and 7. Initially, we can see that the combination of all features achieved the best results for almost all categories as well as for the whole dataset. Thus, we can conclude that our method is better than the baselines. In particular, the F_1 scores for 6 categories, `Cr`, `Pr`, `Rg`, `Rl`, `Cn` and `Tc`, were improved more than 5 points compared with the best results of individual feature groups. Weaker results for `Rg`, `LT` and `PM` categories were likely due to the relatively small size of training data compared with the number of training data for the co-occurring categories (`Cr` and `Dp`) as shown in Fig. 7 and Table 2.

To confirm the conclusions, we perform approximate randomization tests [49] for the top-2 baselines (Doc2Vec and LSA) with All+NB on micro-average F_1. The comparison results (Doc2Vec vs. All+NB and LSA vs. All+NB) showed 0.0310 and 0.0301, respectively, in the case where we repeated comparisons 1000 in the test. Thus, we can claim that the result of our classifier is statistically significant.

> **Q.** How well did all classifiers perform in other measurements (macro-average F_1, MA, average precision@5, HL, LL, RL and AUC)?
> **A.** For macro-average F_1, multi-label accuracy, average precision@5 and RL, All+NB performed the best.
> **A.** For HL and LL, SVM-RBF was the best.
> **A.** For AUC, SVM-Lin. was the best.

We evaluated two other kinds of accuracies (macro-average F_1 (MF), multi-label accuracy (MA), average precision@5, three loss functions (HL, LL and RL) and AUC in Table 8. Similar to the results of micro-average F_1, we can see that combining all feature vectors improved the scores and generated the best outcomes. In particular, all the best scores were generated from the combined all feature vectors by All+NB for MF, MA and RL, All+SVM-RBF for average precision@5, HL and LL or All+SVM-Lin. for AUC.

From the above results, we can conclude that combining all the features improves scores for all the categories. However, the results also showed that two classifiers NB and SVM-RBF achieved the best scores in many measurements. To better understand the differences between the two classifiers, we focus on the two classifiers in the remainder of this experimental evaluations.

> **Q.** Which was better, All+NB or All+SVM-RBF?
> **A.** All+NB was better overall because All+NB's micro-average recall and F_1 were on average approximately 30% and 10% higher than the values of All+SVM-RBF. However, comparing micro-average precision scores, All+SVM-RBF was better than All+NB approximately by 10%.

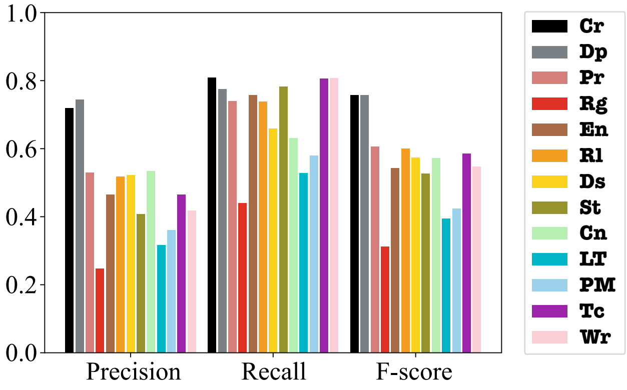

Fig. 5 Micro-average precision, recall and F_1 scores for All+NB

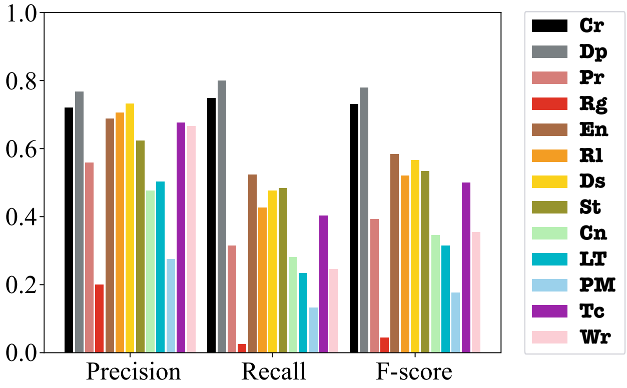

Fig. 6 Micro-average precision, recall and F_1 scores for All+SVM-RBF

Figures 5 and 6 show micro-average precisions and recalls as well as F_1 scores for All+NB and All+SVM-RBF. Looking at All+NB's result, this classifier achieved over 70% scores in recall for 8 categories: `Cr`, `Dp`, `Pr`, `En`, `Rl`, `St`, `Tc` and `Wr`. On the other hand, there are only 2 categories `Cr` and `Dp` in which the classifier achieved over 70% scores in precision and F_1-score. Looking at results of All+SVM-RBF, this classifier obtained high precision scores in 4 categories: `Cr`, `Dp`, `Rl` and `Ds`. However, its recall scores tended to be low excluding `Cr` and `Dp`. If improving precision or reducing the loss scores is important, SVM-RBF may be a good classifier instead of NB.

Comparing the two classifiers, we can say that All+NB is better than All+SVM-RBF because All+NB's micro-average recall and F_1 were on average approximately 30% and 10% higher than the corresponding All+SVM-RBF values.

To confirm these conclusions, we performed approximate randomization tests for the two classifiers on micro-average F_1. The result showed 0.0213 in the case where the comparison test was repeated 1000 times. Thus, we can claim that All+NB is statistically significant from All+SVM-RBF.

Table 8 Label- and example-based results without micro-average F_1. Scores of macro-average F_1 (MF), multi-label accuracy, (MA), average precision@5 (ap@5), hamming loss (HL), log loss (LL), ranking loss (RL) and AUC

	MF (%)	MA (%)	ap@5 (%)	HL	LL	RL	AUC
TF-IDF (v_1)	7.4	9.2	41.7	0.185	31.589	0.326	0.510
Doc2Vec (v_2)	47.3	38.4	40.9	0.274	10.668	0.258	0.512
LSA (v_3)	43.9	34.9	41.0	0.154	8.177	0.145	0.520
LDA (v_4)	31.8	26.7	37.8	0.168	6.821	0.187	0.508
Noun-context (v_5)	26.6	28.9	39.3	0.174	44.741	0.212	0.526
LP	32.0	20.1	28.6	0.799	6.450	0.373	0.518
DLP	33.3	20.9	28.6	0.791	6.705	0.361	0.500
LPAC	44.7	33.4	40.4	0.343	6.354	0.176	0.510
LNP	46.7	35.1	26.4	0.320	17.889	0.314	0.500
All+NB	**55.4**	**45.8**	**42.5**	0.191	9.314	**0.138**	0.519
All+RFs	40.3	35.6	40.1	0.171	10.565	0.242	0.528
All+SVM-RBF	45.0	43.1	42.4	**0.145**	**5.292**	0.143	0.530
All+SVM-Lin.	39.4	38.5	41.1	0.154	5.514	0.162	**0.534**
All+LP	44.5	34.2	38.8	0.189	5.674	0.152	0.518
All+DLP	32.0	20.1	28.0	0.799	6.523	0.386	0.500
All+LPAC	44.4	34.1	33.3	0.190	5.612	0.151	0.510
All+LNP	43.5	31.0	28.8	0.209	23.178	0.540	0.500

The bold-faced numbers indicate the best score for a classifier over the three-dimensional reduction methods

5.3.3 Error analysis

Q. Why was the result of Rg weak for all classifiers?
A. Events of the category tend to co-occur with other categories (Cr and Dp) where the numbers of training data for the co-occurring categories were approximately twice those of the Rg category.

To better understand the reasons why the Rg category was weak result in Table 7, we plot the number of co-occurring category pairs in Fig. 7, in which clearly Cr and Dp are often used with Rg. However, the numbers of training data of Cr and Dp were approximately twice those of Rg as shown in Table 3; therefore, the small size of training data can be considered the reason for weak results for the Rg category.

Q. How and why did ALL+NB incorrectly predict results?
Q. How and why did ALL+NB perform missing correct label assignments?
A. The Pr and St categories were often wrongly assigned to each other.
A. If several categories such as Cn, PM and Ds co-occurred together on the same events, the categories tended to be wrongly assigned to each other.

We next analyzed how and why our classifier performed mis-predictions. In particular, we analyze (1) what categories ALL+NB wrongly assigned data to and (2) what suitable categories the classifier did not assign data to. Figure 8 shows the categories that were incorrectly assigned to events. We can see that several Ds (Disparity) events were assigned to the Rg (Reign), Cn (Community), PM (Popular Movement) and Wr (War) categories. One possible reason for assigning Cn or PM is that Ds category has a high co-occurrences with the two categories as shown in Fig. 7. The three categories of events, Wr, Rg and Ds, often refer to locations, which can be a reason for wrongly assigning Wr or Rg to Ds events. For example, an event text *what economic or political policy issues may trigger for causing economic disparity on a specific location* can be seen as a disparity-related event. As for both Wr- and Rg-related events, a text *which countries began to invade to another country* can belong to the categories.

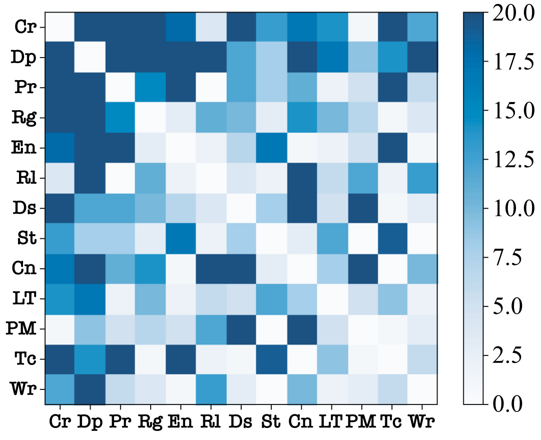

Fig. 7 Co-occurrences of labels

We also observe that several Pr (Production) events were wrongly assigned to the Ds or St (Study) categories. First, it is relevant that there is a strong relationship between Pr

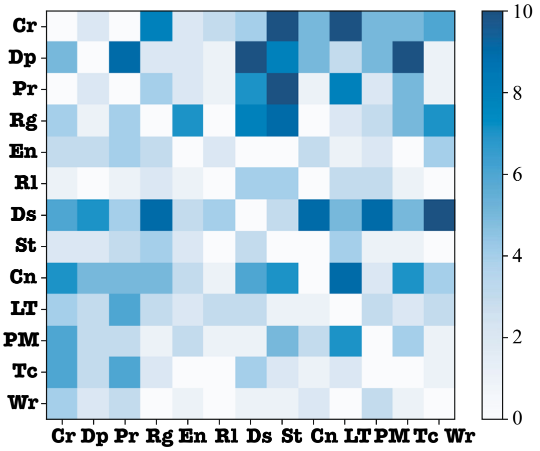

Fig. 8 Wrongly assigned categories by NB. The x axis represents categories that were wrongly assigned to events of y axis's categories

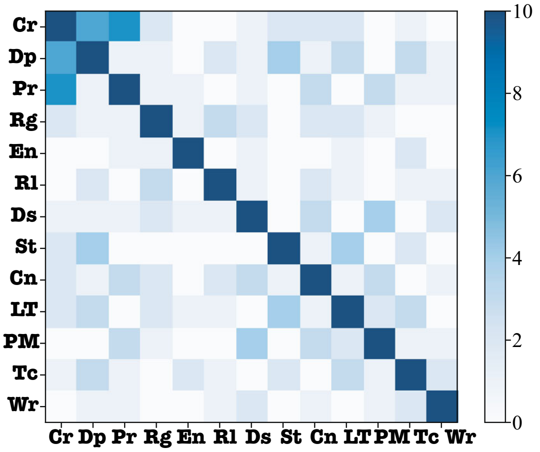

Fig. 9 Missed categories by NB. The x axis represents categories that were correct but were not assigned to events of y axis's categories

and `St` since many production events can be results of study events. Also, these two category events may generate disparity; thus, these wrong assignments can be considerable.

We can see that `Dp` (Diplomacy) was often incorrectly assigned to `Pr`, `Ds`, `St` or `Tc` (Technology) events. `Dp` events include negotiation such as the Trans-Pacific Partnership Agreement which is a trade agreement involving several products; therefore, the classifier wrongly assigned `Dp` to `Pr`. The same reason can be considered for reasons of assigning the category to `St` and `Tc` because it is possible to regard outputs of these two categories as productions. Next, looking at Fig. 7, we can see that some events commonly can have `Dp` and `Pr`; it is a possible reason for the mis-prediction.

Figure 9 shows the number of categories that are attached in the test data for which no classifiers were assigned. We can see that several test data that were attached to two categories such as `Cr` (commerce) and `Dp` were often assigned to only one of them. Similarly, the `Pr` category also missed `Cr` events.

> **Q.** Were there dependencies between different categories?
> **A.** Regarding mutual information measurements, there were strong dependencies between 3 categories, `Tc`, `LT` and `PM`.
> **A.** Regarding Euclidean distance measurements, events of `Cr`, `Dp` and `Pr` or `PM`, `Tc` and `Wr` were similar to each other within the 3 categories.

We then calculated scores of the mutual information (MI) that represent dependencies between categories. The formal definition of MI is as follows:

$$\mathrm{MI}(X'_1, X'_2) = \sum_{x_1 \in X'_1} \sum_{x_2 \in X'_2} p(x_1, x_2) \log\left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right) \tag{14}$$

Figure 10 plots the obtained MI values between the categories. We can see that there are strong dependencies among categories `Tc`, `LT` (`Literature and Thought`) and `PM`. This result is considerable because some technologies can affect our social life. For example, IT items, such as personal computer, can change our work styles in ways like office or remote working. The items also have potential of creating movements such as affecting the work style not only for specific persons/organizations but also for societies.

In addition, we measure similarities between different categories by Euclidean distance. The lower the score of the distance between two feature vectors is, the more similar they are. We calculate the distances for all combinations of feature vectors of 2 categories and plot them in Fig. 11. The scores among 3 categories of `Cr`, `Dp` and `Pr` were low. This is because the 3 categories tend to be assigned to the same feature vectors as shown in Fig. 7. Another observation is that there were relatively strong dependencies among the 3 categories `PM`, `Tc` and `Wr`. Similar to the relationship between `LT`, `PM` and `Tc`, technology can have strong relationships with `PM` and `Wr` related events.

> **Q.** Were there dependencies between feature vectors in the same category?
> **A.** The `Pr`, `St` and `LT` categories have strong event dependencies.

We next investigate the dependencies of feature vectors by MI in the same category. Table 9 shows the scores of 3 categories `St`, `LT` and `Tc` have relatively strong dependencies between feature vectors in their categories compared with other categories. In addition, we apply Euclidean distances to measure the inner-category similarity. Interestingly, feature vectors of `Tc` are relatively similar to each other since its

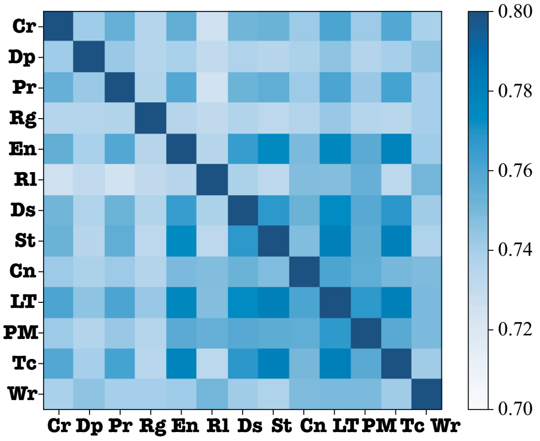

Fig. 10 Inter-category dependency. MI of feature vectors between different categories

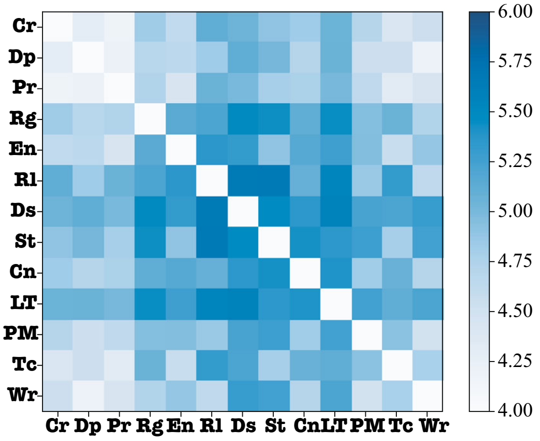

Fig. 11 Inter-category similarity. Euclidean distances between different categories

Table 9 Inner-category dependency and similarity. MI scores and Euclidean distances of feature vectors in the same categories

Category	MI	Euc. Dist.
Cr	0.751	4.275
Dp	0.739	4.167
Pr	0.756	4.061
Rg	0.750	5.002
En	0.774	4.591
Rl	0.758	4.908
Ds	0.766	5.430
St	0.780	4.972
Cn	0.731	5.162
LT	0.784	5.445
PM	0.764	4.611
Tc	0.782	4.394
Wr	0.758	4.025
Ave.	*0.761*	*4.69*

The italic-faced number indicates the average score of all categories

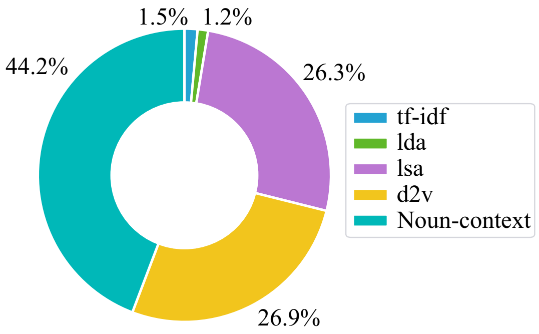

Fig. 12 Ratio of important feature vectors

score is low. This result indicates that this category's events describe similar topics or use the same words.

5.3.4 Analyzing importance of feature vectors

Q. How was each feature important in this study?
Q. How well were scores of micro-average F_1 for NB and SVM-RBF equipped with the important features?
A. The Noun-context-based feature was the most used feature type to create vectors.
A. Doc2Vec's importance score was the highest of all feature type.
A. The Noun-context + SVM-RBF had better scores for En than ALL+NB although the total score of Noun-context + SVM-RBF was weaker than that of ALL+NB.

Figure 12 shows ratios of the numbers of feature types used in the result of applying the proposed method. We can see that the noun-context type is the most used feature type. This feature type occupies approximately 45% of the combined feature vector.

In Fig. 13, we show the average importance values (blue bars) and standard deviations (black lines) of our features. We can see that the Doc2Vec and other semantic-based features were the most important in event MLC. The noun-context feature is the lowest in this figure.

Tables 10 and 11 show scores of six measurements (micro- and macro-average F_1, multi-label accuracy, hamming loss, log loss and ranking loss) for NB and SVM-RBF equipped with Doc2Vec and Noun-context-based feature vectors. Looking at the result of SVM-RBF's micro-average F_1, total scores were weaker than the ones of All+NB; however, there was a category (En) where SVM-RBF was better than ALL+NB. In addition, three loss scores for SVM-RBF were better than the ones for NB. On the other hand, comparing results of Doc2Vec and Noun-context-based feature vectors for NB, the scores of Doc2Vec were better than ones for the other more than 10% for 10 categories: Pr, Rg, En,

Table 10 Micro-average F_1 scores for the top important feature. F_1 scores for NB obtained when using d2v and Noun-context settings for each class

Category	All+NB		All+SVM-RBF	
	Doc2Vec (%)	Noun-context (%)	Doc2Vec (%)	Noun-context (%)
`Cr`	64.7	69.7	24.0	44.2
`Dp`	71.6	73.8	20.8	52.0
`Pr`	53.4	38.9	0.0	3.2
`Rg`	20.1	7.6	0.0	0.0
`En`	46.8	5.5	30.0	49.1
`Rl`	50.1	26.0	10.3	30.3
`Ds`	60.4	28.6	26.7	22.5
`St`	36.6	13.5	14.7	19.2
`Cn`	31.0	27.5	0.0	8.3
`LT`	36.2	25.3	2.9	4.0
`PM`	38.5	0.0	5.7	11.9
`Tc`	52.5	14.9	6.3	18.8
`Wr`	53.2	14.8	4.7	14.5
Total	52.9	43.1	15.4	30.4

Table 11 Scores for the top important feature. Scores of macro-average F_1 (MF), multi-label accuracy (MA), hamming loss (HL), log loss (LL), ranking loss (RL) for NB and SVM-RBF

		MF (%)	MA (%)	HL	LL	RL
NB	Doc2Vec	47.3	38.4	0.274	10.668	0.258
	Noun-context	26.6	28.9	0.174	44.741	0.212
SVM-RBF	Doc2Vec	11.2	9.3	0.185	5.333	0.148
	Noun-context	21.4	18.5	0.167	5.084	0.133

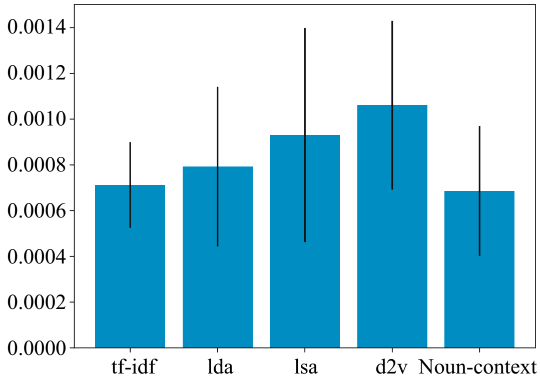

Fig. 13 Feature importances

`Rl`, `Ds`, `St`, `LT`, `PM`, `Tc` and `Wr`. One possible reason for the difference was the sizes of the feature vectors since the noun-context-based feature type uses all words to create the feature vector. The sizes of Doc2Vec and Noun-context feature types are 100 and 24,594, respectively.

Q. How well did combinations of feature vectors improve micro-average F_1 scores?

Table 12 Feature combinations. Micro-average F_1 scores for NB equipped with incrementally added feature vectors

	Simple (%)	Dimensional reduction (%)
v_5	43.1	54.2
$v_5 + v_2$	44.4	57.5
$v_5 + v_2 + v_3$	45.0	58.8
$v_5 + v_2 + v_3 + v_4$	45.0	59.1
All	36.5	60.2

A. Adding Doc2Vec to noun-context resulted in the most improvement. Adding other feature types with dimensional reduction also improved the score by approximately 1%.

A. Applying the dimensional reduction method was important as combining feature types without the method resulted in lower scores than the case where the method was used. In particular, adding TF-IDF without the method reduced the score.

We then investigate which features truly contribute to improving the micro-average F_1 scores. We incrementally combined Noun-context-based feature vectors with

Doc2Vec, LSA, LDA and TF-IDF in descending order of numbers of used feature type in combined all feature types and results are contained in Table 12. The simple column means that we combine feature vectors without applying dimensional reduction methods and train NB on them whereas the dimensional reduction column lists the score for NB trained on the results of applying L1 dimensional reduction. The dimensional reduction column indicates that it is able to linearly improve accuracy by combining feature vectors with applying a dimensional reduction. The simple column also shows that adding feature types makes the improvement excluding TF-IDF.

6 Conclusions

Understanding categories of events can have many applications including support for building historical analogy models, across-time connection of events/entities, or structuring longer text collections such as Wikipedia (e.g., year related articles). In this paper we introduce a classification technique for multi-labeled documents of events. We showed that our technique could improve micro-average F_1 scores by approximately 10%. For this evaluation, we created a new ground truth dataset, and have made it available on a public repository.

Future work will (a) *investigate a novel feature selection algorithm that is robust for training data*. It is considerable that the feature selection process also utilizes correlations; however, one of the trends in multi-label classification (MLC) studies is how to train classifier on an incomplete dataset whose labeled data have wrongly assigned categories or missed suitable ones. This trend indicates that it is problematic for feature selection to simply use the correlation that leads to incorporating the wrong correlation. One of the solutions is to find implicit semantic intermediate labels from feature vectors; however, this is essentially classification. Thus, although it might be possible to incorporate the findings of implicit semantics to feature vectors, it might not be straightforward. We believe that this study can be useful as a baseline to facilitate designing the novel feature selection method. We also plan to (b) *propose a novel and effective learning system specialized to history*. This system will bridge past and present events by estimating how well each event is relevant to event categories.

Acknowledgements This work was partially supported in part by MEXT Grant-in-Aids (#17K12792, #19K20631 and #26750076). We express our gratitude to all the reviewers for their thoughtful comments.

References

1. Au Yeung, C.M., Jatowt, A.: Studying how the past is remembered: towards computational history through large scale text mining. In: CIKM ’11, pp. 1231–1240. ACM, New York (2011)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Boix-Mansilla, V.: Historical understanding: beyond the past and into the present. In: Stearns, P.N., Seixas, P., Wineburg, S. (eds.) Knowing, Teaching, and Leaning History: National and International Perspectives, pp. 390–418. New York University Press, New York (2000)
4. Chapman, A., Facey, J.: Placing history: territory story identity-and historical consciousness. Teach. Hist. **116**, 36–41 (2004)
5. Chen, W., Yan, J., Zhang, B., Chen, Z., Yang, Q.: Document transformation for multi-label feature selection in text categorization. In: ICDM ’07, pp. 451–456. IEEE Computer Society, Washington, DC (2007)
6. Chew, M.M., Bhowmick, S.S., Jatowt, A.: Ranking without learning: towards historical relevance-based ranking of social images. In: SIGIR ’18, pp. 1133–1136. ACM, New York (2018)
7. Clavert, F., Majerus, B., Beaupré, N.: #ww1. twitter, the centenary of the first world war and the historian. Twitter for Research (2015)
8. Cong, G., Lee, W., Wu, H., Liu, B.: Semi-supervised text classification using partitioned em. In: Lee, Y., Li, J., Whang, K.Y., Lee, D. (eds.) Database Systems for Advanced Applications. Lecture Notes in Computer Science, vol. 2973, pp. 482–493. Springer, Berlin (2004)
9. Cook, J., Das Sarma, A., Fabrikant, A., Tomkins, A.: Your two weeks of fame and your grandmother’s. In: Proceedings of the 21st International Conference on World Wide Web, WWW ’12, pp. 919–928. ACM, New York (2012)
10. Creecy, R.H., Masand, B.M., Smith, S.J., Waltz, D.L.: Trading MIPS and memory for knowledge engineering. Commun. ACM **35**(8), 48–64 (1992)
11. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)
12. Doquire, G., Verleysen, M.: Mutual information-based feature selection for multilabel classification. Neurocomputing **122**, 148–155 (2013). (Advances in cognitive and ubiquitous computing)
13. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: NIPS’01, pp. 681–687. MIT Press, Cambridge (2001)
14. Ferron, M., Massa, P.: Collective memory building in Wikipedia: the case of north African uprisings. In: WikiSym ’11, pp. 114–123. Mountain View, California (2011)
15. Garcia-Gavilanes, R., Mollgaard, A., Tsvetkova, M., Yasseri, T.: The memory remains: understanding collective memory in the digital age. Sci. Adv. **3**(4), e1602368 (2017)
16. Ghani, R.: Combining labeled and unlabeled data for multiclass text categorization. In: ICML ’02, pp. 187–194. Morgan Kaufmann Publishers Inc., San Francisco (2002)
17. Gopal, S., Yang, Y.: Multilabel classification with meta-level features. In: SIGIR ’10, pp. 315–322. ACM, New York (2010)
18. Halbwachs, M.: La Memoire Collective. Les Presses universitaires de France (**in French**) (1950)
19. Harris, R., Rea, A.: Making history meaningful: helping pupils to see why history matters. Teach. Hist. **125**, 28–36 (2006)
20. Hoerl, C., McCormack, T.: Time and Memory: Issues in Philosophy and Psychology. Oxford University Press, Oxford (2001)
21. Huet, T., Biega, J., Suchanek, F.M.: Mining history with le monde. In: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction. In: AKBC ’13, pp. 49–54. ACM, New York (2013)

22. Ikejiri, R.: Designing and evaluating the card game which fosters the ability to apply the historical causal relation to the modern problems. Jpn. Soc. Educ. Technol. **34**(4), 375–386 (2011). (in Japanese)
23. Ikejiri, R., Fujimoto, T., Tsubakimoto, M., Yamauchi, Y.: Designing and evaluating a card game to support high school students in applying their knowledge of world history to solve modern political issues. In: ICoME '12. Beijing Normal University (2012)
24. Ikejiri, R., Sumikawa, Y.: Developing a mining system to transfer historical causations to solving modern social issues. In: WHA '16 (2016)
25. Ikejiri, R., Sumikawa, Y.: Developing world history lessons to foster authentic social participation by searching for historical causation in relation to current issues dominating the news. J. Educ. Res. Soc. Stud. **84**, 37–48 (2016). (in Japanese)
26. Jacoby, R.: Social Amnesia: A Critique of Contemporary Psychology. Transaction Publishers, Piscataway (1997)
27. Jatowt, A., Duh, K.: A framework for analyzing semantic change of words across time. In: JCDL '14, pp. 229–238. IEEE Press, Piscataway (2014)
28. Jatowt, A., Kawai, D., Tanaka, K.: Digital history meets Wikipedia: analyzing historical persons in Wikipedia. In: JCDL '16, Newark, New Jersey, USA, pp. 17–26 (2016)
29. Jatowt, A., Kawai, D., Tanaka, K.: Predicting importance of historical persons using Wikipedia. In: CIKM '16, pp. 1909–1912. ACM, New York (2016)
30. Jatowt, A., Kawai, D., Tanaka, K.: Timestamping entities using contextual information. In: SIGIR '17, pp. 1205–1208. ACM, New York (2017)
31. Jatowt, A., Kawai, H., Kanazawa, K., Tanaka, K., Kunieda, K., Yamada, K.: Multi-lingual analysis of future-related information on the web. In: Culture and Computing'13, pp. 27–32 (2013)
32. Kanhabua, N., Nguyen, T.N., Niederée, C.: What triggers human remembering of events?: a large-scale analysis of catalysts for collective memory in Wikipedia. In: JCDL '14, London, United Kingdom, pp. 341–350 (2014)
33. Kosmerlj, A., Belyaeva, E., Leban, G., Grobelnik, M., Fortuna, B.: Towards a complete event type taxonomy. In: WWW '15 Companion, pp. 899–902. ACM, New York (2015)
34. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis. In: EMNLP '04, pp. 230–237
35. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML'14, Bejing, China, pp. 1188–1196 (2014)
36. Lee, J., Kim, D.W.: Feature selection for multi-label classification using multivariate mutual information. Pattern Recognit. Lett. **34**(3), 349–357 (2013)
37. Lee, J., Kim, D.W.: Fast multi-label feature selection based on information-theoretic feature ranking. Pattern Recognit. **48**(9), 2761–2771 (2015)
38. Lee, P.: Historical literacy: theory and research. Int. J. Hist. Learn. Teach. Res. **5**(1), 25–40 (2005)
39. Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: WWW '05, pp. 391–400. ACM, New York (2005)
40. Lieberman, E., Michel, J.B., Jackson, J., Tang, T., Nowak, M.A.: Quantifying the evolutionary dynamics of language. Nature **449**, 713–716 (2007)
41. McCallum, A.K.: Multi-label text classification with a mixture model trained by EM. In: AAAI 99 Workshop on Text Learning (1999)
42. Mikolov, T., Kai, C., Suchanek Greg, C., Dean, J.: Linguistic regularities in continuous space word representations. In: ICLR'13 Workshop (2013)
43. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS'13, 2013, Lake Tahoe, Nevada, United States, pp. 3111–3119 (2013)
44. Mikolov, T., Yih, W.t., Zweig, G.: Efficient estimation of word representations in vector space. In: NAACL'13 (2013)
45. Ministry of Education Culture, Sports, Science and Technology: Japan Course of Study for Senior High Schools (2009)
46. Miyazaki, T., Sumikawa, Y.: Label propagation using amendable clamping. In: IUI'18 Workshop on WII (2018)
47. Nie, L., Wang, M., Zha, Z., Li, G., Chua, T.S.: Multimedia answering: enriching text QA with media information. In: SIGIR '11, pp. 695–704. ACM, New York (2011)
48. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Mach. Learn. **39**(2–3), 103–134 (2000)
49. Noreen, E.W.: Computer-Intensive Methods for Testing Hypotheses. Wiley, New York (1989)
50. Odijk, D., de Rooij, O., Peetz, M.H., Pieters, T., de Rijke, M., Snelders, S.: Semantic document selection. In: TPDL'12, pp. 215–221. Springer, Berlin (2012)
51. Ogata, I., Kato, T., Kabayama, K., Kawakita, M., Kishimoto, M., Kuroda, H., Sato, T., Minamizuka, S., Yamamoto, H.: Encyclopedia of Historiography. Koubundou, Minamiuonuma (1994)
52. Pargel, M., Atkinson, Q.D., Meade, A.: Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature **449**, 717–720 (2007)
53. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: WWW '08, pp. 91–100. ACM, New York (2008)
54. Radinsky, K., Davidovich, S., Markovitch, S.: Learning causality for news events prediction. In: WWW '12, pp. 909–918. ACM, New York (2012)
55. Radinsky, K., Horvitz, E.: Mining the web to predict future events. In: WSDM '13, pp. 255–264. ACM, New York (2013)
56. Singh, J., Nejdl, W., Anand, A.: History by diversity: helping historians search news archives. In: HIIR '16, pp. 183–192. ACM, New York (2016)
57. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: SIGIR '10, pp. 841–842. ACM, New York (2010)
58. Staley, D.J.: A history of the future. Hist. Theory **41**, 72–89 (2002)
59. Sumikawa, Y., Jatowt, A.: Classifying short descriptions of past events. In: Advances in Information Retrieval, ECIR '18, pp. 729–736. Springer, Berlin (2018)
60. Sumikawa, Y., Jatowt, A., Düring, M.: Digital history meets microblogging: analyzing collective memories in twitter. In: JCDL '18, pp. 213–222. ACM, New York (2018)
61. Sun, X., Wang, H., Yu, Y.: Towards effective short text deep classification. In: SIGIR '11, pp. 1143–1144. ACM, New York (2011)
62. Takahashi, Y., Ohshima, H., Yamamoto, M., Iwasaki, H., Oyama, S., Tanaka, K.: Evaluating significance of historical entities based on tempo-spatial impacts analysis using Wikipedia link structure. In: HT '11, pp. 83–92. ACM, New York (2011)
63. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multi-label classification of music by emotion. EURASIP J. Audio Speech Music Process. **2011**(1), 4 (2011)
64. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data, pp. 667–685. Springer, Boston (2010)
65. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: ICWSM'10, Washington, DC, USA (2010)
66. van Drie, J., van Boxtel, C.: Historical reasoning: towards a framework for analyzing students' reasoning about the past. Educ. Psychol. Rev. **20**(2), 87–110 (2008)
67. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)

68. Wang, B., Tu, Z., Tsotsos, J.K.: Dynamic label propagation for semi-supervised multi-class multi-label classification. In: 2013 IEEE International Conference on Computer Vision, pp. 425–432 (2013)
69. Wang, F., Zhang, C.: Label propagation through linear neighborhoods. In: ICML'06, pp. 985–992. ACM, New York (2006)
70. Yang, Y.: Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In: SIGIR '94, New York, NY, USA, pp. 13–22 (1994)
71. Zelikovitz, S., Marquez, F.: Transductive learning for short-text classification problems using latent semantic indexing. Int. J. Pattern Recognit. Artif. Intell. **19**(2), 146–163 (2005)
72. Zhang, M.L., Pea, J.M., Robles, V.: Feature selection for multi-label naive Bayes classification. Inf. Sci. **179**(19), 3218–3229 (2009)
73. Zhang, M.L., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. Pattern Recognit. **40**(7), 2038–2048 (2007)
74. Zhang, Y., Jatowt, A., Bhowmick, S., Tanaka, K.: Omnia Mutantur, Nihil Interit: connecting past with present by finding corresponding terms across time. In: ACL/IJCNLP, pp. 645–655. ACL (2015)
75. Zhang, Y., Jatowt, A., Tanaka, K.: Temporal analog retrieval using transformation over dual hierarchical structures. In: CIKM '17, pp. 717–726. ACM, New York (2017)
76. Zhu, X.: Semi-supervised learning with graphs. Ph.D. thesis, Pittsburgh, PA, USA (2005). AAI3179046