2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

Sentence-based Plot Classification for Online Review Comments

Hidenari IWAI, Yoshinori HIJIKATA, Kaori IKEDA, Shogo NISHIDA
Graduate School of Engineering Science
Osaka University, Japan
Email: hijikata@sys.es.osaka-u.ac.jp

***Abstract*—Many shopping sites provide functions to submit a user review for a purchased item. Reviews of items including stories such as novels and movies sometimes contain spoilers (undesired and revealing plot descriptions) along with the opinions of the review author. In this paper, we propose a system that helps users see reviews without seeing such plot descriptions. This system classifies each sentence in a user review as plot-related or non-plot-related and hides plot descriptions from user reviews. We tested five common machine-learning algorithms to ascertain the appropriate algorithm to address this problem. We also proposed a method of generalizing people's names, which we think is strongly related to the plot description. We verified its contribution to the classification results. Finally, we implemented a display interface of user reviews in which users can control the level of plot hiding.**

I. Introduction

Online shopping has become increasingly popular in recent years. People gather information related to items (products in online shopping sites) when they are considering buying them. They decide whether or not to buy an item based on that gathered information. Some websites (e.g. Amazon.com and eBay) provide users with a function by which they can write and read reviews to items easily. Reviews are useful for evaluating a certain item because users can learn many other users' opinions from them.

However, reviews of items that include stories (e.g. novels, movies, and computer games) might come with *spoilers* (undesired plot descriptions). Here is an example.

⟨Review Example⟩ *Half-Blood Prince is easily one of the better books in the Harry Potter series. Several of the chapters are particularly well-written, with great suspense and imagery. After completing this book, I was in a state of total shock. Before reading this book, if I had to make a list of impossible things that could never happen, Snape killing the Headmaster and fleeing the school with a bunch of Death Eaters would have been right at the top of the list. However, I'd have been wrong. I had a very strong feeling that Dumbledore would be the one to die in this book. However, I never saw it coming the way it happened. The disturbing ending leaves you frustrated in anticipation of the next book.* (Review from Amazon.com "Harry Potter and the Half-Blood Prince")

Users can understand that this book is a good one of this series from this review. The opinion is useful for people making a decision about whether or not to buy this book. However, readers are forced to be informed of a crucial plot twist, and one climactic event in the whole Harry Potter saga that they would not have known before reading the book: 'Snape killing the Headmaster'. After reading this review, they might be less satisfied and much less surprised when reading the book.

As described in this paper, we propose a system that helps users see reviews without seeing plot descriptions. Here, plot descriptions mean the descriptions of items' stories. The system hides a part of a review that is regarded as a plot description and presents the remaining part of the review to users. A sentence is a unit that can be shown or hidden from users. For the decision of hiding, the system classifies each sentence as plot description or not (hereinafter *plot classification*). Plot classification is realized using machine learning algorithms. Furthermore, generalization of people's names and peculiar words is incorporated to the detection method for improving the performance. Our system targets English-language reviews.

Here, we do not directly classify sentences as a spoiler or not. A spoiler sentence includes a description regarding a plot of the story. However, the seriousness or sensitivity for knowing the story (the level or threshold for judging as a spoiler) differs among users (The details will be shown in Subsection V-B). For example, knowing a fact that a character will be murdered later in the book is not critical for some users. However, other users do not want to know even the fact that a character will die in the book. Therefore, we detect plot descriptions and assign them the score of the likelihood of a plot description.

This study contributes to the literature of this field by accomplishing the following.

- Classification of sentences in reviews as plot descriptions or not is achieved for the first time.
- Five common machine-learning algorithms were tested for use in the classification.
- Some methods of improving classification performance were proposed based on the review characteristics.
- A display interface was developed that hides plot descriptions from user reviews.

The remainder of this paper is organized as follows. First, we introduce some related works. Then, we evaluate the performance of the plot classification realized by the five machine-learning algorithms. We also propose some methods

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.40

IEEE computer society

for improving the classification performance and evaluate them. Subsequently, we show that users' judgment on spoiler differs among users and propose a system that hides plot descriptions at the user's favorite hiding level. Finally, we present some conclusions and future works.

II. Related work

A. Spoiler filtering

Recently, studies of spoiler detection or plot detection has been done for review documents or articles in social media.

Golbeck tried to identify and block every tweet on a given topic [1]. She especially tried to block tweets on TV programs and sporting events. For detecting spoiler tweets on TV programs, she extracted actor and character names from the IMDB(Internet Movie Database). She considered that a tweet is a spoiler if it includes the above names. For detecting spoiler tweets on sporting events, she made a blacklist of words that are mostly names of player, team and stadium. This is actually the simple method and works well for a short message like tweets in Twitter. However, if a sentence is in a longer-review comment, the above proper names are not necessarily included even if it is a spoiler.

Nakamura et al. developed a filtering system of spoilers for sports news [2]. In this system, users should input several words related to their favorite sports team or their favorite sporting event like "baseball" and "Yankees", and "Olympic" and "figure skating". Then the system displays news on the Web hiding the results of the related sporting event. They used a rule-based method for detecting spoilers. They also compared several interfaces such as deletion and blacking out for hiding the results of sporting events from Web contents [3]. Keywords used for reporting sporting event results are limited compared to those used in a spoiler of items including story. A rule-based method cannot be applied to the review comments to these items.

Guo and Ramakrishnan considered that spoilers of items including story are related to plot descriptions like we do [4]. If the review comment (document) includes more descriptions on the item's story content, it may be a spoiler in higher probability. They ranked review comments according to the spoiler probability score. They used bag-of-words representation and LDA-based model. For calculating the spoiler probability score, they calculated the similarity between synopsis obtained from the item description page on IMDB(Internet Movie Database) website and comment. Because this method predicts the related topics from the words, it is difficult to apply this method to sentence-level spoiler (or plot) detection.

B. Opinion classification

Although identifying a plot or a spoiler is relatively new research theme, the idea itself is similar to opinion classification, which is popular research theme in natural language processing. We introduce some major studies and recent trends here. The following are the basic studies. Pang et al. classified sentences in reviews as subjective or objective using SVM and Naive Bayes [5]. Yu et al. separated opinions from facts at either document level or sentence level using Naive Bayes [6]. Riloff et al. [7] proposed a bootstrapping process to enable classifiers to learn from unlabeled texts.

Classifying opinions as positive or negative is of interest lately. This classification is also called polarity classification. Dave et al. [8] proposed a polarity classification method at the sentence level using Naive Bayes. Wilson et al. [9] proposed a method for classifying each sentence as positive or negative considering a priori polarity and contextual polarity of words in the sentence.

Recently, research orientations of three kinds are becoming popular. One is aspect-based sentiment analysis (or opinion mining) [10], [11]. This type of study extracts entity (product name) and aspects (product feature) for opinion analysis. Nouns are usually recognized as aspect candidates [10], frequencies are used for evaluating aspect candidates [12]. A user's evaluations to aspects are predicted using special word classes such as a sentiment word, a modifier and a negator [13].

Another is cross-domain sentiment classification that builds a classifier in one domain and applies it to another domain with little work on collecting additional learning data [14]. Semi-supervised learning [15], joint-topic modeling [16] and graph-based method with the label propagation [17] are used for bridging the two domains.

The last orientation is cross-language sentiment classification that exploits English review data and sentiment analysis tools for analyzing non-English reviews [14]. Usually reviews written in the target language (language other than English) are translated to the source language (English) [18], [19]. In this case, English tools and analysis methods are used for sentiment classification. Some researchers took an approach of translating reviews in the source language to the target language [20], [21], [22], which is intended to increase the data size for learning the classification model.

Our study aims at detecting plot descriptions from user reviews on items including stories. Opinions are easy to classify as positive or negative using dictionaries of evaluation expressions. Plot descriptions include the characters' actions or feelings. Nevertheless, no dictionary about characters' actions or opinions exists. Apparently classifying sentences as plot descriptions or not is rather more difficult than polarity classification is.

III. Plot classification by Machine Learning

Plot classification is realized using a machine learning technique. Machine-learning algorithms of five kinds are tested for plot classification. First, we make a classifier using each kind of machine-learning algorithm. Then, we classify each sentence of unknown reviews as a plot description or not. The classification results are evaluated using ground-truth data.

A. Approach

To develop a system that helps users read reviews without learning important parts of stories, we must consider how to detect plot descriptions from a review. We checked 100 reviews in Amazon.com to ascertain clues to support a judgment of whether a sentence is a plot description or not. Some specific words are likely to appear in plot descriptions; other words seldom appear in plot descriptions. For example, the words 'kill' and 'island' seldom appear in sentences, except in plot descriptions. However, the words 'think' and 'frustrated' are frequently used to represent reviewers' opinions, but these words seldom appear in plot descriptions. However, it is difficult to check all of the words manually to ascertain which one appears in plot descriptions but which seldom appears outside of plot descriptions and which one seldom appears in plot descriptions but which often appears outside of plot descriptions. It is also difficult to infer rules for detecting plot descriptions from those checked words.

For this study, we use machine-learning algorithms using words as features for identifying sentences as plot descriptions or not. This method judges each sentence of new user reviews using a learned classifier as a plot description or not. We call this method *sentence-level plot classification method (SP method)* in this section. Sentence p is represented in a bag-of-words for conducting machine learning as

$$p = \langle w_1, w_2, ..., w_M \rangle. \tag{1}$$

Therein, w_m stands for a word. The number of occurrences $x_{n,m}$ of each word w_m in sentence p_n is also recorded. For creating bag-of-words model, we did not remove stop words because some stop words may be related to plot (Actually, personal pronouns are related to plot. The details are explained in Subsection IV-C). The Porter stemming algorithm is conducted to remove morphological and inflectional endings from words.

We adopt machine-learning algorithms that produce models in which we can obtain each sentence's score of the likelihood of a plot description. We therefore adopt the following five algorithms: Naive Bayes (NBayes), Support Vector Machine (SVM), Logistic regression (Logistic), decision tree (D-tree), and the k-nearest neighbor algorithm (k-NN). We use C4.5 as a learning algorithm of the decision tree.

B. Dataset and evaluation metric

We compiled a dataset for use with the evaluation in the following way. First, we selected the comic, novel, and DVD categories from the category list provided by Amazon.com. Those categories were selected because they included items with stories. Second, 100 items were selected randomly from each category. We collected five reviews randomly per item. In all, we collected 500 reviews per category. We asked three human evaluators to find plot descriptions among the sentences in the collected reviews and label them as 'plot'. We treated that sentence as belonging to 'plot' class if two or more of three people regarded a certain sentence as plot. We treated the sentences which were not assigned to 'plot' class as 'non-plot' class. Table I shows the number of words, the number of words occurring more than once in the data set, the number of sentences assigned 'plot' class, and the number of sentences assigned 'non-plot' class in each category. The number of words means the number of kinds of words that correspond to elements of the bag-of-words.

TABLE I
STATISTICS OF GROUND TRUTH.

	comic	novel	DVD
#word	6414	6334	6966
#word (> 2)	3603	3539	3761
#plot sentences	1523	1602	1357
#non-plot sentences	3484	3225	3445
#plot documents	304	279	256
#non-plot documents	193	219	239

We calculated the kappa coefficient [23] to ascertain the degree of accordance of labels among the three evaluators. The results are 0.612 for the comic category, 0.544 for the novel category and 0.466 for the DVD category. Generally, the value of kappa coefficient the accordance is regarded as low in 0–0.4, as medium in 0.4–0.6, as high in 0.6–0.8 and extremely high in 0.8–1.0 [24]. We can say that the labels acquired in our data set preparation have medium or high accordance.

We measure the classification performance using the F-value on the plot class (F-value). The F-value reflects both the precision and recall of a classification. Given the set of sentences P that are classified as 'plot' class by the classifier and the set of sentences G that are assigned 'plot' class in the ground truth data, the F-value is calculated as follows.

$$Precision = \frac{|P \cap G|}{|P|} \tag{2}$$

$$Recall = \frac{|P \cap G|}{|G|} \tag{3}$$

$$F\text{-}Value = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

C. Learning condition

We evaluated the five machine learning algorithms using ten-fold cross validation to avoid data bias. Generally, for classification problem, imbalanced data might have a negative influence on classification performance. As shown in Table I, the number of sentences with 'plot' class is smaller than those with 'non-plot' class in our data set. The following three measures are well-known against imbalanced data [25]: (i) conduct over-sampling of the data with a class of fewer data, (ii) conduct subsampling of the data with a class of larger data and (iii) ignore one of the two classes using a recognition-based instead of a discrimination-based inductive scheme. We used subsampling because our dataset is sufficiently large.

Generally, the machine learning performance depends on the number of attributes (the kinds of words in our study). We changed the number of attributes for use in machine learning from 10 to 2000 (and the maximum number of attributes) for the evaluation. We selected words used as features based on the mutual information, as Glover et al. did [26]. We did not use words appearing only once in the data set. For SVM, polynomial kernel was used in the experiment. To obtain the appropriate probability, the Logistic regression model is used for the output of the SVM.

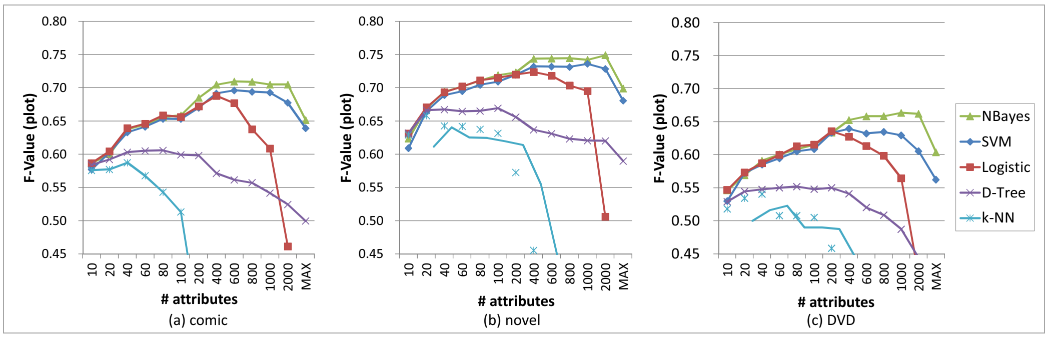


Fig. 1. F-values of each algorithm for the SP method changing the number of attributes.

D. Baseline

The objective of this evaluation is to clarify which machine learning algorithm is appropriate for the plot classification. We also compare our method (SP method) with a standard classification method.

First, we checked the difficulty of the classification problem we treated. We randomly classified sentences to plot or non-plot description. We tested random methods of two kinds. The first one decides plot or not in 50% to 50% probability. The second one decides an original percentage of plot ratio (see Table I). The following are the classification results (F-value) of the random methods: 0.378 (comic), 0.399 (novel) and 0.361 (DVD) in 50% to 50%; and 0.304 (comic), 0.332 (novel), and 0.283 (DVD) in original plot ratio.

Classifying a sentence is a more difficult task than classifying a document (a set of sentences) because it cannot use much word information. Therefore, a method that classifies a review document to plot or not and considers all the sentences in the document as the same classified class may outperform a sentence classification. We call this method *document-level plot classification method (DP method)*. For example, when a document is classified to plot, all the sentences in the classified document are assigned to plot. This method is also realized using machine-learning techniques. The five machine learning algorithms explained in Subsection III-A are also tested in the DP method. It is necessary to prepare ground truth data in which each document is given a ‘plot’ class or ‘non-plot’ class. We asked three human evaluators to judge each document in our dataset as plot or not plot. We treated that document as belonging to ‘plot’ class if two or more of three people regarded a certain document as plot. We treated the sentences which were not assigned to ‘plot’ class as ‘non-plot’ class. Table I shows the number of documents assigned to the ‘plot’ class and those assigned to the ‘non-plot’ class in each category. We calculated the kappa coefficient [23] to elucidate the degree of accordance of labels among the three evaluators. The results are 0.588 for the comic category, 0.582 for the novel category and 0.510 for the DVD category. Table II shows the F-values of the process that classifies a review document to plot or not in NBayes, SVM and Logistic in each category.

TABLE II
THE F-VALUES OF THE PROCESS THAT CLASSIFIES A REVIEW DOCUMENT TO PLOT OR NOT.

	comic	novel	DVD
NBayes	0.804	0.829	0.810
SVM	0.800	0.831	0.817
Logistic	0.808	0.827	0.807

E. Results

Figures 1-(a), (b), and (c) show the F-values of five machine learning algorithms used for the SP method in comic, novel and DVD category (the results of the DP method are shown later). We tested k as 1, 3, ..., 39, 41 in k-NN. Only the best results are shown ($k = 19$ in comic, $k = 5$ in novel, and $k = 23$ in DVD). Little difference is found among NBayes, SVM, and Logistic until the number of attributes becomes 200 in each category. When it exceeds 200, the F-values of NBayes become higher than those of other algorithms. When comparing F-values at the highest case, NBayes achieves the highest value, SVM becomes the second highest, and Logistic becomes the third highest in every category. The k-NN and D-tree achieved worse values than the three algorithms above.

The decrease of F-values of Logistic begins earlier than NBayes and SVM when the number of attributes increases because logistic regression is influenced by multicollinearity. Generally in machine learning, the generalization performance of the classifier becomes worse when the data have high dimensionality. This phenomenon is known as the curse of dimensionality. However, NBayes and SVM have a feature relaxing the curse of dimensionality. Therefore, the timing of the decrease of F-value becomes later than that of other algorithms.

Table III shows the number of attributes in each combination of algorithm and category when F-values become the highest (Hereinafter, the ONA (optimal numbers of attributes)) for the SP method. The column ‘Sentence-level classification’ in Table IV shows the F-values under the ONA of the SP method. We see whether statistically significant difference exists among algorithms under the condition of the ONA. We conducted t-

TABLE III
OPTIMAL NUMBER OF ATTRIBUTES OF EACH ALGORITHM FOR DOCUMENT-LEVEL PLOT CLASSIFICATION.

Algorithm	Category		
	comic	novel	DVD
NBayes	600	2000	1000
SVM	600	1000	400
Logistic	400	400	200
D-tree	80	100	80
k-NN	40	20	40

TABLE IV
SENTENCE-LEVEL PLOT CLASSIFICATION VS. DOCUMENT-LEVEL PLOT CLASSIFICATION.

Category	Algorithm	Sentence-level classification	Document-level classification
comic	NBayes	**0.709**	0.537
	SVM	**0.696**	0.521
	Logistic	**0.688**	0.510
novel	NBayes	**0.749**	0.573
	SVM	**0.736**	0.565
	Logistic	**0.724**	0.570
DVD	NBayes	**0.663**	0.511
	SVM	**0.639**	0.510
	Logistic	**0.636**	0.503

test for this statistical validation. The results show a significant difference was found between NBayes and SVM, Logistic, k-NN, and D-tree in every category ($p < 0.05$). We can see that NBayes is the best for the sentence-level plot classification from this result.

Because the SP method calculates the score of the likelihood of a plot description, the threshold for plot classification can be changed. Figure 2 shows the precision and recall curve of NBayes, SVM, and Logistic of the SP method. The results are the same tendency in every category, therefore we shows the result in the comic category. The precision of NBayes is better than that of SVM and Logistic under every recall value.

We also tested the DP method for comparison. D-Tree and k-NN did not exhibit high performance. Therefore, we only tested NBayes, SVM and Logistic in this study. For acquisition of the ONA, we changed the number of attributes from 10 to 2000 (and the maximum number of attributes). Table V presents the ONA in each combination of algorithm and category. Table IV shows results when comparing the SP method (the proposed method) with the DP method under their ONAs. The results show that the F-values of the DP method are lower than the SP method. We found statistically significant differences between the above two methods when conducting t-test on each algorithm ($p < 0.01$).

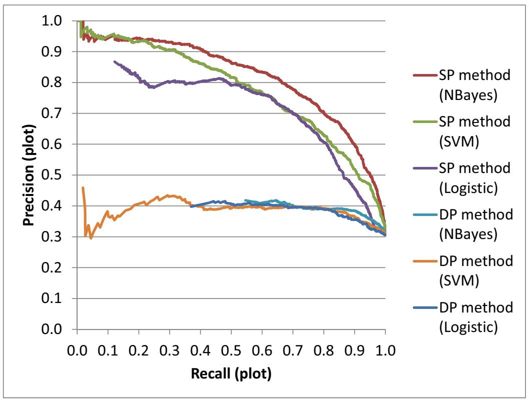


Fig. 2. The precision and recall curves of the sentence-level classification and docment-level classification using Naive Bayes and SVM.

TABLE V
OPTIMAL NUMBER OF ATTRIBUTES OF EACH ALGORITHM FOR DOCUMENT-LEVEL CLASSIFICATION.

Algorithm	Category		
	comic	novel	DVD
NBayes	200	100	60
SVM	80	80	80
Logistic	80	20	40

The conclusions obtained from the above results is that NBayes achieved the highest F-values among other algorithms. NBayes tends to achieve high F-values steadily when the number of attributes to use is changed. Naive Bayes is the most appropriate algorithm for plot classification. The proposed sentence-level plot classification method (SP method) achieved better results than the document-level plot classification method (DP method). When class information is assigned in sentence level, the sentence-level plot classification achieves better results than the document-level plot classification.

IV. GENERALIZING WORDS OF PERSONAL NAMES AND PECULIAR WORDS

In this section, we improve our sentence-level plot classification method proposed in Section III. In detail, the new method generalizes people's names and peculiar words to avoid zero frequency in other items. We evaluate the degree of improvement that it achieves.

A. Method

We found out that sentences that include character names tend to be plot descriptions. We further found that sentences including author names tend not to be plot descriptions. Here, a character name is the name of a character that appears in an item; an author name is a creator's name. However, character names and author names are of a huge variety. For that reason, the same words seldom appear in reviews of other items. Low scores are assigned to sentences including those words.

If we generalize those words and train the classifier using the generalized words as features, then we can use the existence of character names and author names in the sentences for plot classification. When a character name and an author name are found, they are respectively replaced by a tag representing a character name and a tag representing an author name. For example, "Peter finally got married with his rival Sally." changes into "⟨character⟩ finally got married with his rival ⟨character⟩." Here, ⟨character⟩ is the tag denoting that the replaced word is a character name. We generalize author names similarly (⟨author⟩ is used for this tag). In the DVD category, reviewers write actor and actress names in their reviews. Therefore, we generalize those names in the DVD category (⟨actor⟩ is used for this tag). Reviewers might also use character names that do not appear in the content summary. We sought to generalize words representing people's names (⟨other⟩ is used for this tag). We designate tags of the above types as 'generalization tags'.

To generalize character names and author names, we used the database of names of people provided by The U.S. Census Bureau[1], the online dictionary (a general English dictionary) provided by ALC Press Inc.[2], and item description pages provided by Amazon.com. We can obtain a content summary from an item description page in Amazon.com. First, the words that were listed as general names (words aside from personal names) in the online dictionary were removed from the database of names of people. We used the list of the remaining words as a dictionary of personal names. If a word in a review appeared in both the content summary and the dictionary of personal names, then the word was generalized to a character name. We extracted author names and actor names from each item description page. They were easy to extract because each author name and actor name appeared in a uniform manner in the item description page.

The generalization method described above can identify and generalize popular human names (mainly European names). However, the method does not cover human names of other regions such as Asia and Africa. Furthermore, names that are not used in real life occur in some comics such as 'Doraemon' and 'Pokémon'. These words are not recorded in the dictionary of names of people. New names are coming into the world one after another. It is difficult to incorporate these names adequately into this system.

Therefore, we generalize words that have only occurred in one item as peculiar words for that item. These peculiar words might include the above character names. For finding peculiar words, we first find words that appear in more than one item in our dataset. These words are recorded in the list of non-peculiar words. We checked all the words in the dataset and generalized them as peculiar words if they are not recorded in the list of non-peculiar words (⟨peculiar⟩ is used for this tag).

B. Evaluation

We evaluated our improvement methods using the dataset explained in Subsection III-B. Before evaluating the performance of plot classification, we evaluated the accuracy of the word generalization. The precision for extracting author names and actor names become 100% because they are written in a uniform manner in the item description pages. However, our extraction method of people's names sometime cause false-negative detection because of the existence of general words that are also used for people's name like 'White', 'Page' and 'Card'. These general words are deleted from the dictionary of personal names. Table VI shows the number of false-negative detections (FN), false-positive detections (FP), true-positive detections (TP), the precision and the recall. Our generalization method achieves high precision in any category. However, its recall is not high. We designed our extraction method so that its extraction precision becomes high. If it extracts words that are not actually people's name, the performance of plot classification may largely decrease because description of the character seems highly-related to plot.

TABLE VI
PRECISION AND RECALL OF PEOPLE NAMES EXTRACTION.

	false-negative	false-positive	true-positive	Precision	Recall
comic	48	3	52	0.945	0.520
novel	60	2	58	0.967	0.492
DVD	12	0	44	1.000	0.786

Next, we evaluated the performance of plot classification. We compared the case that does not incorporate word generalization (hereinafter 'ORG'), the case that generalizes personal names (character name, author name, actor name and other personal names) (hereinafter 'CAN'), the case that generalizes peculiar words (hereinafter 'PEC'), and the case that generalizes both personal names and peculiar words (hereinafter 'CAN+PEC'). We used NBayes, SVM, and Logistic as learning algorithms here because they achieved better results than those of D-tree and k-NN. The ONA was found for each combination of the above cases of three kinds and the above algorithms of three kinds by changing the number of attributes to use for learning the model in the same manner as in Section III-C. Results under the ONAs are shown in Figure 3.

When examining CAN, one finds that the F-values of CAN are better than those of ORG in the comic and novel category. When comparing the comic category and the novel category, the effect of word generalization of the novel category is better than that of the comic category because many popular personal names exist in the novel category although some people's names are not registered in the dictionary of personal names in the comic category.

For clarification of the degree of generalization effect to plot classification, we examined mutual information of each word and the 'plot' label. We obtained the ranking of words according to mutual information. Table VII shows the ranks of generalization tags and their frequency (generalization of peculiar words was not conducted). In the novel category, ⟨character⟩, which stands for the generalized tag of personal names, has high mutual information. However, it is not so high in the comic category, as shown by the result in which the improvement of F-value is small in the comic category.

Compared to the comic and novel category, almost no effect of the generalization of people's name exists in the DVD category. Character names and actor names were obtained in the DVD category. However, the users used character names and author names for both presenting the story of the item and criticizing the actor's performance. Therefore, the effect of generalization of names is lower than that in the comic and novel category.

[1] http://www.census.gov, 1995 edition
[2] http://www.alc.co.jp

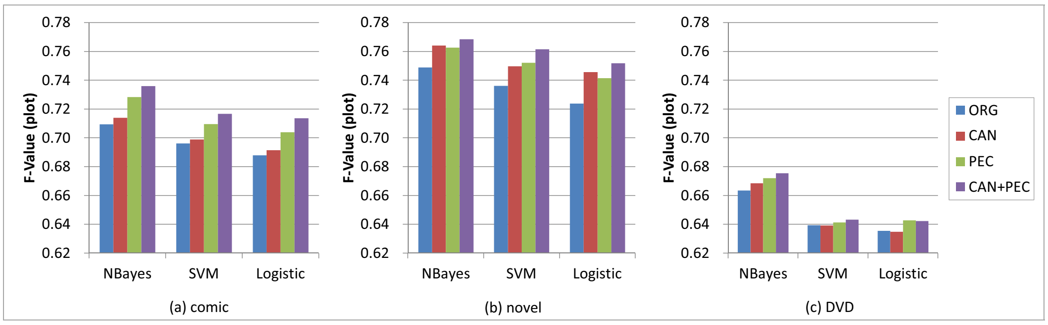


Fig. 3. *F*-value for each algorithm with generalization (under the optimal number of attributes).

TABLE VII
RANK ON MUTUAL INFORMATION AND THE NUMBER OF OCCURRENCES OF GENERALIZATION TAG.

	comic		novel		DVD	
	Rank	# tags	Rank	# tags	Rank	# tags
⟨character⟩	9	196	2	623	15	263
⟨author⟩	34	320	18	427	2575	198
⟨actor⟩	-	-	-	-	8	522
⟨other⟩	11	304	158	256	11	382

TABLE VIII
WORDS PECULIAR FOR AN ITEM AND HIGHLY RANKED IN MUTUAL INFORMATION (MI).

(a) comic ($<$ 100 rank of MI)

Rank	Word	MI
16	Negi	0.0067
31	Husky	0.0042
35	Tasuku	0.0040
42	Cooro	0.0033
54	Shun	0.0029
56	Ryouta	0.0028
57	Katchoo	0.0028
62	Owly	0.0027
63	Anima	0.0027
68	Grumply	0.0026
76	Cain	0.0024
84	Wormy	0.0023
93	Hatsumi	0.0022
96	Asuna	0.0021
97	Avatar	0.0021

(b) novel ($<$ 100 rank of MI)

Rank	Word	MI
47	Sarah	0.0039
57	AJ	0.0032
73	Uthred	0.0027
93	Tempe	0.0021
99	Cord	0.0019

(c) DVD ($<$ 150 rank of MI)

Rank	Word	MI
86	Corso	0.0019
90	surgery	0.0018
116	Wilder	0.0016
123	Domon	0.0015
126	booker	0.0015
136	Rome	0.0014
146	Carver	0.0013

The effectiveness of the generalization of peculiar words was also examined (see Figure 3). Although the *F*-value of PEC is higher than that of ORG in the novel category, there is little difference between PEC and CAN. The *F*-value of PEC is higher than that of CAN in the comic category. This result illustrates that the generalization of personal names contributes to the classification sufficiently in the novel category (also see Table VII). Many names are not recognized as personal names in the comic category (we will examine this point later). The case that generalizes both personal names and peculiar words (CAN+PEC) achieved the highest *F*-value in each category.

We examined whether a statistically significant difference on *F*-values exists or not when the generalization of personal names and peculiar words is incorporated. *t*-test is used for that examination. Although no significant difference is found in the DVD category, it exists between ORG–PEC, ORG–CAN+PEC, CAN–CAN+PEC, PEC–CAN+PEC in the comic category and it exists between ORG–CAN, ORG–CAN+PEC in the novel category ($p < 0.01$).

Results show that the generalization of personal names is effective for the novel category. The generalization of peculiar words is effective for the comic category. When we generalize both personal names and peculiar words, the *F*-value became the best in the entire category. Although we find no significant difference in the DVD category, the *F*-value increased when we introduced the generalization in all the algorithms. Therefore, our improvement method is effective for plot classification.

C. Consideration

We examined the actual words recognized as peculiar words by the system. Table VIII shows the words generalized as peculiar words and ranked high in mutual information. These words are actual characters' names that are not registered in the dictionary of personal names. They do not contribute to plot classification when the generalization of peculiar words is not done because they are not registered in the dictionary of personal names and appear only in a specific item. When considering the words recognized as peculiar words and ranked higher in mutual information, the number of generalized words is the highest in the comic category. From this, we can understand the reason why the generalization of peculiar words worked the best in the comic category.

Finally, we examined words with high mutual information in each category. Table IX shows the top 20 words with high mutual information in each category (generalization of personal names and peculiar words is conducted). Here, the words are processed through the Porter stemming algorithm. Generalization of peculiar words is highly-associated to the 'plot' class in the comic and DVD category. Generalization of character names is highly-associated to the 'plot' class in the novel category.

Here, another interesting finding comes out. The most of the highly-ranked words are generally considered as stop words like 'I', 'he' and 'who'. These words do not generally present the meaning or the content of a document. Therefore, they are usually deleted for many NLP applications like search and

TABLE IX
TOP 20 WORDS IN MUTUAL INFORMATION.

Rank	comic	novel	DVD
1	⟨peculiar⟩	i	⟨peculiar⟩
2	i	⟨character⟩	i
3	he	book	hi
4	hi	read	he
5	her	⟨peculiar⟩	dvd
6	she	her	who
7	book	he	my
8	read	she	him
9	who	thi	⟨actor⟩
10	⟨character⟩	hi	thi
11	him	famili	their
12	⟨other⟩	son	⟨other⟩
13	with	their	movi
14	my	seri	her
15	fight	ha	with
16	comic	who	⟨character⟩
17	into	you	watch
18	thi	mother	find
19	friend	⟨author⟩	woman
20	power	my	she

classification. When these words are written in the review document for an item with story, they are mostly used for the characters' actions. This fact improves the performance of plot classification. This shows that the plot classification task is inherently different from other NLP problems.

V. IMPLEMENTATION OF A PLOT HIDING SYSTEM

A. *System structure*

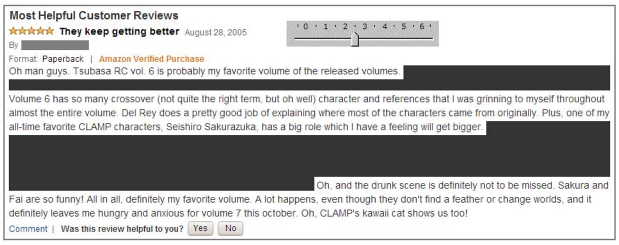
We implemented a system in which users can browse user reviews without seeing plots. The system was implemented in a Web application. The program on the server is implemented in a Java servlet. The client-side program is JavaScript program embedded in a HTML document. The server-side program embeds the likelihood score of plot in each sentence by adding ⟨span⟩ tag. The score is recorded in an attribute value to the tag. We targeted Amazon.com as e-commerce site for this implementation. The likelihood score is calculated using the NBayes algorithm with the generalization of personal names and peculiar words.

The learned classifier calculates the score of the likelihood of a plot description in each sentence. The client-side program hides sentences with a score higher than a threshold when displaying user reviews. Blacking out (filling with black color) is used for hiding sentences that are judged as plot descriptions. The screenshot of our system is depicted in Figure 4.

B. *User's control*

As explained in Section I, a spoiler is different from a plot description. A plot description is a description of an item's story. Which plot description becomes a spoiler is dependent on users. When a user considers the plot description as the one he/she does not want to see, it becomes a spoiler. In other words, the seriousness or sensitivity for knowing the story (the level or threshold for judging as a spoiler) differs among users.

We examined the diversity of spoiler judgments among users. We asked seven people to judge each sentence a spoiler or not in the dataset explained in Section III-B. The ratio of spoilers in each user are 23.5%, 11.1%, 10.6%, 9.0%, 6.3%, 5.5% and 3.2%. The highest ratio is 23.5% and the lowest ratio is 3.2%. We can see that the degree of judging as a spoiler is different among users.


Most Helpful Customer Reviews
★★★★★
0 1 2 3 4 5 6
Comment
Was this review helpful to you?
Yes
No

They keep getting better August 28, 2005

By [redacted]

Format: Paperback | Amazon Verified Purchase

Oh man guys. Tsubasa RC vol. 6 is probably my favorite volume of the released volumes. [redacted]

Volume 6 has so many crossover (not quite the right term, but oh well) character and references that I was grinning to myself throughout almost the entire volume. Del Rey does a pretty good job of explaining where most of the characters came from originally. Plus, one of my all-time favorite CLAMP characters, Seishiro Sakurazuka, has a big role which I have a feeling will get bigger. [redacted] Oh, and the drunk scene is definitely not to be missed. Sakura and Fai are so funny! All in all, definitely my favorite volume. A lot happens, even though they don't find a feather or change worlds, and it definitely leaves me hungry and anxious for volume 7 this october. Oh, CLAMP's kawaii cat shows us too!

Fig. 4. Reviews displayed by our proposed system by Amazon.com.

Therefore, we let users to control the level of plot by the slider (see Figure 4). Users can adjust the threshold to hide sentences. They set the threshold on several user reviews in advance. Here, they should select items in which they do not care to see the items' stories (items the users has already read or watched). After that, they browse other items' reviews under the set threshold. We expect that users can read user reviews without seeing spoilers to some extent.

VI. CONCLUSION

As described in this paper, we proposed a method for classifying review sentences as plot description or not. Five common machine-learning algorithms were tested: Naive Bayes, SVM, Logistic Regression, Decision Tree and k-Nearest Neighbor Algorithm. We evaluated the classifiers trained using the five algorithms in terms of the F-value on the plot class. Results show that the performance of Naive Bayes was the best of the five algorithms. Furthermore, we proposed improvement methods by generalizing words of personal names and peculiar words. We evaluated the degree of improvement using the improvement methods. Results suggest that generalization of personal names and peculiar words serve as effective features for use in plot classification. We also implemented an interface that hides plot descriptions. In this interface, users can adjust the hiding level of plot descriptions. We evaluated that our plot classification method can detect sentences with plot descriptions. However, we have not evaluated whether it can hide sentences with a spoiler for real users. As a subject for future work, we intend to conduct a user experiment on whether the system can hide spoilers.

REFERENCES

[1] J. Golbeck, "The Twitter Mute Button: a Web Filtering Challenge," *Proc. of CHI '12*, pp. 2755–2758, 2012.

[2] S. Nakamura and T. Komatsu, "Temporal Filtering System to Reduce the Risk of Spoiling a User's Enjoyment," *Proc. of IUI '07*, pp. 345–348, 2007.

[3] S. Nakamura and T. Komatsu, "Study of Information Clouding Methods to Prevent Spoilers of Sports Match," *Proc. of AVI '12*, pp. 661–664, 2012.

[4] S. Guo and N. Ramakrishnan, "Finding the Storyteller: Automatic Spoiler Tagging Using Linguistic Cues," *Proc. of COLING '10*, pp. 412–420, 2010.

[5] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. of ACL '04*, pp. 271–278, 2004.

[6] H. Yu and V. Hatzivassiloglou, "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences," *Proc. of EMNLP '03*, pp. 129–136, 2003.

[7] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," *Proc. of EMNLP '03*, pp. 105–112, 2003.

[8] K. Dave, S. Lawrence, and D.M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. of WWW '03*, pp. 519–528, 2003.

[9] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proc. of HLT '05*, pp. 347–354, 2005.

[10] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. of KDD '04*, pp. 168–177 2004.

[11] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *Proc. of WWW '05*, pp. 342–351, 2005.

[12] M. Hu *et al.*, "OpinionBlocks: A Crowd-Powered, Self-Improving Interactive Visual Analytic System for Understanding Opinion Text," *Proc. of INTERACT '13*, pp. 116–134, 2013.

[13] L. Qu, G. Ifrim, and G. Weikum, "The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns," *Proc. of COLING '10*, pp. 913–921, 2010.

[14] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, 2012.

[15] A. Aue and M. Gamon, "Customizing Sentiment Classifiers to New Domains: A Case Study," *Proc. of RANLP '05*, 2005.

[16] Y. He, C. Lin, and H. Alani, "Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification," *Proc. of ACL '11*, pp. 123–131, 2011.

[17] W. Qion, S. Tan, and X. Cheng, "Graph Ranking for Sentiment Transfer," *Proc. of ACL-IJCNLP '09*, pp. 317–320, 2009.

[18] X. Wan, "Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis," *Proc. of EMNLP '08*, pp. 553–561, 2008.

[19] M. Bautin, L. Vijayarenu, and S. Skiena, "International Sentiment Analysis for News and Blogs," *Proc. of ICWSM '08*, 2008.

[20] R. Mihalcea, C. Banea, and J. Wiebe, "Learning Multilingual Subjective Language via Cross-Lingual Projections," *Proc. of ACL '07*, pp. 976–983, 2007.

[21] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual Subjectivity: Are More Languages Better?," *Proc. of COLING '10*, pp. 28–36, 2010.

[22] J. Brooke, M. Tofiloski, and M. Taboada, "Cross-linguistic Sentiment Analysis: From English to Spanish," *Proc. of RANLP '09*, pp. 50–54, 2009.

[23] S. Siegel and N.J. Jr. Castellan, "Nonparametric Statistics for the Behavioral Sciences," McGraw-Hill, 1988.

[24] J.R. Landis and G.G. Koch, "The measurement of observer agreement for categorical data," International Biometric Society, Vol. 33, No. 1, pp. 159–174, 1977.

[25] N. Japkowicz, "Learning from Imbalanced Data Sets: A Comparison of Various Strategies," AAAI Press, Vol. 68, pp. 10–15, 2000.

[26] E.J. Glover *et al.*, "Using Web Structure for Classifying and Describing Web Pages," *Proc. of WWW '02*, pp. 562–569, 2002.