2016 IEEE/WIC/ACM International Conference on Web Intelligence

A Basic Study on Spoiler Detection from Review Comments Using Story Documents

Kyosuke MAEDA
Graduate School of Engineering Science
Osaka University, Japan

Yoshinori HIJIKATA
Graduate School of Engineering Science
Osaka University, Japan
Email: hijikata@sys.es.osaka-u.ac.jp

Satoshi NAKAMURA
School of Interdisciplinary
Mathematical Sciences
Meiji University, Japan

***Abstract*—In many shopping sites such as Amazon.com it is possible to view and write reviews of items (products and content). Reviews of items including stories, such as novels, movies, and comics, include reviewers' opinions. Often, these reviews also include descriptions of the story. In some cases, these descriptions may spoil later reader's or viewer's enjoyment and excitement. Hereinafter, we call these descriptions spoilers. Spoilers may be related to the position in the story line. In this study we use story documents. Story documents are documents that record all of the details of the given story. Using the story documents, we investigate the location to which the content of the spoilers correspond in the story documents. Based on the result of the investigation, we consider how to detect spoilers in reviewers' comments.**

***Keywords*-opinion mining; spoiler detection; story document;**

I. Introduction

In recent years, people have been allowed write their opinions for items (products and content) on the Web (generally on online shopping sites or price-comparison sites). Generally, reviews written by reviewers who have read/watched the items are useful information for users who have not read/watched the items. However, reviews of items containing a story, such as novels, movies, and comics, might also include descriptions of the story elements. Some of these descriptions might refer to the ending of the story or contain details of the storyline (e.g., in a mystery novel, revealing the name of criminal or revealing how the trick is done). When people read novels or watch movies, they usually enjoy imagining what will happen in the next part of the story [1], [2]. Therefore, these descriptions might disappoint people, because they remove some of the fun people derive from the content [3]. Hereinafter, we call these descriptions spoilers.

It is useful for users that a system automatically detects spoilers from review comments. A review comment is a unit of submission, one user gives comments to one item in an online shopping site or a price-comparison site. Recently, there was a study into detecting review comments including descriptions of the story [4] and a study into detecting sentences including descriptions of a story from review comments [5]. However, not all descriptions of the story will disappoint users. Official website for the items and introduction pages at the shopping sites usually refer to the start (preface) of the story, which can interest users. Thus, descriptions of the start of the story are useful for users. By contrast, users usually want to enjoy the ending of the story by reading/watching the items. Therefore, descriptions of the ending of the story may disappoint users.

In this study, we use a story document which records the entire story of the item and examine the locations of spoilers in the story document. A story document is an entire body text of the item when it is a novel. A text of abstract that describes the entire story can also be a story document for other domains (comics and movies). We hypothesized that spoilers occur in the latter half of the story document. To test this hypothesis, we investigate the positions of words related to spoilers in the story document. Note that whether a given user regards descriptions as spoilers is different from person to person. In our investigation, we specifically address spoilers that most users regard as serious. For this investigation, we target Japanese novels for the domain and make a dataset of the words (in Japanese) related to spoilers (hereinafter, the spoiler-words dataset). Additionally, in this study, the story documents are the novels we described above.

Finally, based on the result of our investigation, we attempt to detect spoilers in review comments. The detection method might be a rule based approach or use machine learning algorithms. However, in this study, we detect words that occur with high frequency in the latter half of the story document in review comments. If those words are included in a review comment, we consider the comment includes spoilers. We manually check whether the detected review comments include actual spoilers.

The remainder of this paper is organized as follows. First, we introduce several related studies. Then, we describe how we created the spoiler-words dataset. Next, we introduce the investigation method and the results of the investigation. Then, we provide some discussion of the feasibility of spoiler detection. Finally, we present some conclusions and suggestions for future work.

II. Related work

In this section, we introduce several studies that detect plot elements (i.e., descriptions of the story) or spoilers regarding the story in review comments, and several studies that detect

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.97

spoilers regarding the result of sporting or other events from articles on social media.

A. Detecting plot points or spoilers in review comments

Studies of reviews on the Internet have been widely carried out in the field of text mining [6]. Studies of review comments for items including stories are the most closely related to our study. Recent years have seen studies of spoiler detection or plot point detection in review comments. Guo et al. detect plot points using latent Dirichlet allocation (LDA) [4]. Iwai et al. proposed a system judging whether a sentence in a review comment is a plot point or not using machine learning algorithms [5]. The above studies detect plot points, not spoilers. Plot points contain both useful information and spoilers. We do not detect plot points; rather, we detect spoilers that upset users.

Boyd-Graber et al. collected short sentences describing the story from TV Tropes [1], in which reviewers can determine whether the sentences are spoilers or not, and judged whether the sentences contain spoiler words or not using machine learning algorithms [7]. Their method uses words and sentence structures from review comments. By contrast, using story documents, we attempt to determine where the descriptions (word unit) in the review comments occur in the story document and use the result for spoiler detection.

B. Detecting spoilers from an article on social media

Some studies target social networking services (SNSs), rather than review sites. Klein et al. proposed a system that records how many episodes of items each user watches/reads (the degree of watching/reading progress) and gives a user a warning that the users whose degree of watching/reading progress is faster than the target user might give spoilers in their posts [8]. In addition to items included in a story, some studies detect spoilers for events in the real world. Golbeck detects the result of sports matches as spoilers from Twitter timelines [9]. Nakamura et al. detect the result of sports matches as spoilers, considering the time when the real world event opens and whether users view the event or not [10]. These studies target the result of sporting events. However, we intend to detect spoilers of the content of stories.

III. Creating a spoiler-words dataset

In this study, we investigate how spoilers occur in a story document. First, in this section, we show a policy for making a spoiler-words dataset. Then, we show the concrete dataset creation procedure. Finally, we show the characteristics of the spoilers that evaluators have described to create the spoiler-words dataset.

[1]http://tvtropes.org/pmwiki/pmwiki.php/Main/HomePage

Table I
Novels used in this study

	author name	title	text
item 1	Arthur Conan Doyle	The Red-Headed League	48KB
item 2	Keikichi Osaka	The Hangman of the Department Store	25KB
item 3	Kenji Miyazawa	The Night of the Milky Way Train	84KB
item 4	Souseki Natsume	Kokoro	366KB
item 5	Edgar Allan Poe	The Murders in the Rue Morgue	76KB

The title is translated in English

A. Policy of investigation

We collect spoilers to investigate where spoilers occur in the story document. Because the degree of unpleasantly by spoilers is assumed to be different for each user, we investigate the spoilers most people regard as serious. Therefore, we have asked several evaluators to describe the spoilers. All evaluators are Japanese university students and Japanese is their native language. In addition, the evaluators give the degree of spoiler for all spoilers that the evaluators described, including their own. This indicates how serious the spoiler is to the users. It is necessary to identify where the content of the described spoilers occurs in the story document. However, the text that evaluators have described is written in their own words. Thus, it is difficult to identify the position based on an exact match of the text with a sentence unit. Therefore, we extract words necessary to express the content of spoilers from sentences that the evaluators have described. We call this set of words the spoiler-words dataset.

There exist various kinds of items that include a story (e.g., novels, movies, and comics). In this study, we target Japanese novels from Aozorabunko [2]. The reason for this choice is that the entire text of items is available. In this study, we chose five items from Aozorabunko (see table I). The column of the right-side end of Table I is the amount of text (in KB) at the time of downloading.

B. Making procedure of dataset

We had six Japanese evaluators (three men, three women, with an average age of 19.5 years old) cooperate on collecting spoilers. Evaluators read five novels, which are shown in Table I and listed spoilers for the items in a short sentence in Japanese. We defined spoilers as “content that disappoints people who have not read the novel by revealing information”, and explain this definition to the evaluators.

Then, we ask the evaluators to determine the degree of all spoilers that they have described, including their own.

[2]At this website, the text of literary works that are out of copyright. http://www.aozora.gr.jp/

Table II
THE NUMBER OF DESCRIBED SENTENCES ABOUT THE SPOILERS

	user 1	user 2	user 3	user 4	user 5	user 6	ALL
item 1 (The Red-Headed League)	15	12	17	6	11	11	72
item 2 (The Hangman of the Department Store)	18	7	18	12	9	17	81
item 3 (The Night of the Milky Way Train)	34	19	31	17	7	14	122
item 4 (Kokoro)	44	19	91	14	17	22	207
item 5 (The Murders in the Rue Morgue)	15	17	29	11	13	13	98
ALL	126	74	186	60	57	77	580
mean of the number of the letters per one sentence	55.3	25.1	46.5	23.2	27.2	28.2	41.1

The degree of spoiler is rated on a five-point scale (1: small spoiler, 5: large spoiler). Because we target spoilers that most people regard as serious, we use only those sentences whose degree of spoiler is greater than 3 by majority vote (for more than three evaluators).

Finally, we extract the words necessary to express the content of the spoilers from the spoilers described above. Therefore, we had another five Japanese evaluators (five men, average age of 22.6 years old) participate. We showed the spoilers, in sentence units, to evaluators. A sentence is divided into several phrases. The evaluators choose the minimal phrases necessary to express the content of the shown sentences. The evaluators do not read the target novels. In this task, because evaluators should pay attention only to the meaning of the sentences, we believed that they could perform, regardless of whether they know the items' content. We use CaboCha [3], a Japanese dependency analyzer, to divide a sentence into phrases. We collect the phrases chosen by the majority of evaluators (for more than two evaluators). Then, we perform morphological analysis on the collected phrases and extract substantive information. Specifically, we extracted nouns, verbs, adjectives, and adverbs. This set of words constituted our spoiler-words dataset. Additionally, we changed the character name to a unified character name that we defined and changed all words into their base forms.

C. Characteristics of the Dataset

1) Tendency to answer of the evaluators: Table II shows the number of sentences that the six evaluators described and the mean number of letters per sentence. The average of the number of sentences that they described is 96.6. For all evaluators, the larger the amount of text in the novels (see Table I), the larger the descriptions. In addition, for evaluators who provided longer than average sentences, they described the entire story comprehensively, and thus, the length of every sentence tended to be longer.

2) Reliability of the evaluators: To measure the reliability of each evaluator's spoiler, we waited six months, then had the evaluators determine the degree of spoilers again for the same sentences that each evaluator described. In an evaluation of the first and the second evaluations, we checked whether an evaluation value changed. We use Ebel's intraclass correlation coefficient (ICC) [11] for this investigation. ICC can compute intra-rater reliability [ICC (1, 1)], one evaluator evaluating several times, and inter-rater reliability [ICC (2, 1)] (some evaluators evaluated only one time). We use ICC (1, 1) in this investigation. We show the values of ICC (1, 1) in Table III. In Table III, we also show the average and variance of the first and second evaluation values. Landis et al. said that if the ICC value is 0.6–0.8, it is significant [12]. Because the ICC (1, 1) values in Table III are 0.601–0.795, we believe that the first evaluations of the evaluators agree with the second evaluations.

Table III
THE DEGREE OF SPOILERS AND ICC VALUES

	first		second		
	mean	variance	mean	variance	ICC (1, 1)
user 1	2.65	1.90	2.97	1.83	.601
user 2	3.02	1.64	2.86	1.70	.710
user 3	3.58	1.72	3.33	1.49	.713
user 4	2.83	2.27	3.01	1.88	.760
user 5	3.00	2.14	3.21	2.45	.795
user 6	3.14	1.99	2.79	2.21	.692
ALL	3.11	1.99	3.08	1.85	.708

3) Reliability among evaluators: We show the number of sentences that each evaluator described ("♯ all sentences" in Table IV) and the number of sentences whose degree of spoiler is greater than 3 by majority vote (of more than three evaluators) ("♯ target sentences" in Table IV). From Table IV, nearly one quarter of ♯ all sentences are ♯ target sentences. In addition, we check the agreement about the degree of spoiler among all of the evaluators for the same sentence using ICC (2, 1), which indicates the inter-rater reliability. As a result, ICC (2, 1) became .591. The study in [12] indicates that this value corresponds to moderate agreement. Thus, there exist some differences in the judgment of how serious the content of the sentences is; however, general agreement among evaluators is achieved.

4) The number and characteristics of the extracted words: Table IV shows the number of words extracted by five evaluators ("♯ extracted words" in Table IV). This value is the number of words chosen by more than two evaluators. 98.2% of the extracted words are nouns and verbs (see

[3]http://taku910.github.io/cabocha/

Table IV
THE NUMBER OF SENTENCES DESCRIBED BY EVALUATORS, SENTENCES WHOSE DEGREE OF SPOILER IS GREATER THAN 3 BY MAJORITY VOTE (MORE THAN THREE EVALUATORS), AND EXTRACTED WORDS

	♯ all sentences	♯ target sentences	♯ extracted words
item 1	73	25	24 (24)
item 2	81	24	33 (33)
item 3	122	25	35 (35)
item 4	207	43	64 (63)
item 5	98	24	69 (66)
ALL	581	141	225 (221)

the numbers in brackets in Table IV). Therefore, after the investigation, we limit the words to nouns and verbs.

IV. THE INVESTIGATION METHOD AND THE RESULTS

In this section, first, we describe our investigation method. Then, we show the results indicating how the words in the spoiler-words dataset are distributed in the story document.

A. Investigation method

We present a method that analyzes word occurrence patterns (where the words occur in the story documents). We divide a story document, based on its number of letters, into several segments (hereinafter, we call each segment a part), and calculate a word's occurrence proportion (the number of occurrences of a word in each part, divided by the number of occurrences of a word in the entire story) for every word. Then, we sequentially add the occurrence proportion from the first part to the final part (for all eight parts) for every word (hereinafter, we call the sum of the occurrence proportion a cumulative occurrence proportion). In this investigation, we divide a story document into eight parts. We target the occurrence pattern with an occurrence proportion that is larger in the latter half than in the first half and whose cumulative occurrence proportion becomes 1.0 just for the final part (the eighth part) (hereinafter, we call this occurrence pattern the target pattern). We define a target pattern because of our hypothesis that content of the spoilers is inclined toward the latter half of the story (even occurring in the final scene).

B. Results of the investigation

To determine where the words related to spoilers occur in the story document, we compare the proportion of words fitting the target pattern (words whose occurrence pattern fits the target pattern) for the spoiler-words dataset and the proportion of words fitting target pattern for all words in the story document. Table V shows the number of words extracted from the story document and the proportion of words fitting the target pattern. Table VI shows the number of words that occur in the story document in the spoiler-words dataset and proportion of words fitting the target pattern. In addition, Figure 1 shows the distribution of words

Table V
THE NUMBER OF WORDS AND PROPORTION OF THE TARGET PATTERN IN THE STORY DOCUMENTS

	♯ words	% words fitting target pattern
item 1	1702	0.162
item 2	1084	0.145
item 3	1637	0.138
item 4	4629	0.175
item 5	1884	0.171

Table VI
THE NUMBER OF WORDS AND PROPORTION OF THE TARGET PATTERN IN THE SPOILER-WORDS DATASET

	♯ words *	% words fitting target pattern
item 1	20	0.65
item 2	26	0.461
item 3	25	0.44
item 4	58	0.534
item 5	51	0.411

* the number of words that occur in the story document in the spoiler-words dataset

in the spoiler-word dataset as an example of "The Red-Headed League (item 1)", and we perform qualitative analysis. Figure 1 shows 24 words in the spoiler-words dataset divided into 4 graphs; additionally, it shows a baseline (the occurrence pattern in which the words are equally distributed from the first part to the final part in the story document).

First, we perform quantitative analysis based on Table V and Table VI. For all items, the proportion of words fitting the target pattern in the story documents is lower than 0.2. By contrast, the proportion of words fitting the target pattern in the spoiler-words dataset is higher than 0.4 for all items: most of them are 0.4–0.7. Therefore, descriptions of the spoilers use words occurring toward the latter half of the story.

Then, we perform qualitative analysis. Some words in the spoiler-words dataset are character names or an unusual word (a word that occurs only in one story). However, these words do not always occur toward the latter half of the story. For example, in The Red-Headed League (item 1), the word "The Red-Headed League (the lower left in Figure 1)" occurs from beginning to end. Although some character names such as "Duncan Ross (the lower left in Figure 1)" and "Spaulding (the lower right in Figure 1)" are in the spoiler-words dataset, they rarely occur in the latter half of the story. Contrary to expectations, the distribution of words that might occur throughout the story, such as the stage of the story or the group name, are not in the latter half of the story. In addition, the mysterious words in the first half of the story have a similar distribution. By contrast, certain words (excluding character names and unusual words) occur toward the latter half of the story. For example, "clerk (the upper right in Figure 1)" and "bank (the lower left in Figure 1)" are the key of the demystification in The Red-Headed League. Thus, general words related to spoilers may occur

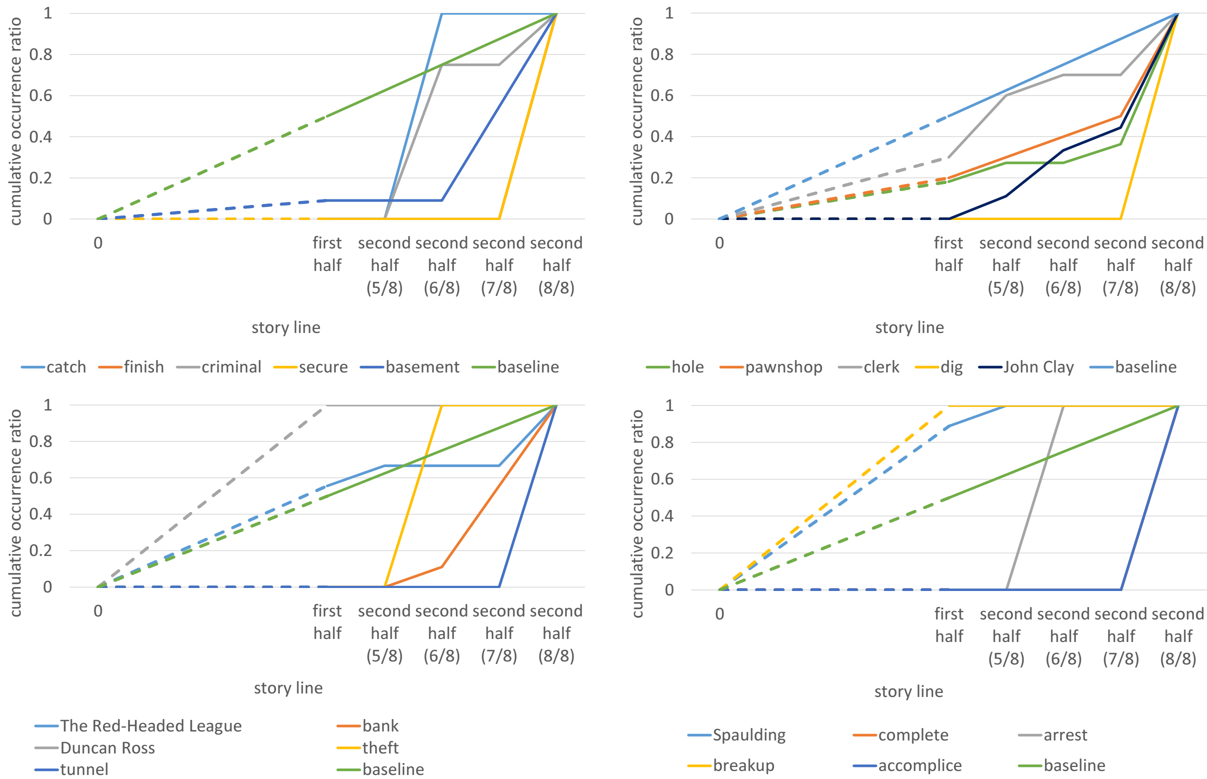

Figure 1. Cumulative occurrence proportion for each word in the spoiler-words dataset (for The Red-Headed League).

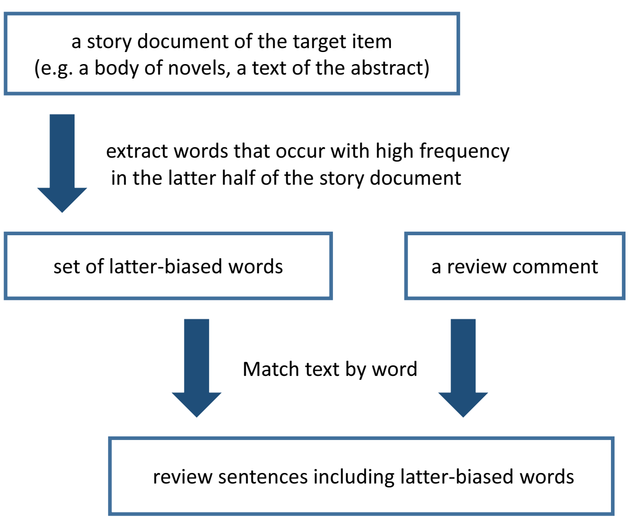

Figure 2. Workflow of our proposed method.

in the latter half of the story document (which are included in spoiler-words dataset).

V. Spoiler Detection from Review Comments

In this section, we propose a method of detecting spoilers from review comments, and consider the review comments that include words fitting the target pattern. Figure 2 shows the procedure of our proposed method detecting spoilers from review comments. Using story documents, this method can extract review sentences including latter-biased words, which may related to the item's spoilers. Important point of our method is that it may be possible to detect only spoilers, which disappoint users.

We collect the review comments from the Japanese online review site Booklog [4]. On this site, when reviewers contribute their reviews, they can attach labels as to whether the review contains spoilers. We explain the result using an example from "The Murders in the Rue Morgue (item 5)" (see "review comment 1" and "review comment 2" below, which are translated in English). Words fitting the target pattern are underlined. Review comment 1 has a label of spoiler. Although we detect important keywords such as "animal" and "window", we also detect general words such as "think" and " image", which are not important. Review comment 2 does not have the label of spoiler, despite it

[4]http://booklog.jp/

having many words fitting the target pattern. For example, "technique" and "reaching". However, in review comment 2, certain phrases are used, such as "closer to a prohibited technique (as story construction of novels)" and "reaching a totally unpredictable ending", which are different from their use in the story document.

As a result, the proposed method might be able to detect spoilers from the review comments. In addition, some of the words fitting the target pattern are used in a different manner in the review comments from the usage manner in the story document.

Review comment 1 ("The Murders in the Rue Morgue") ⟨spoiler⟩
At last, I have read it!!! I was surprised: I thought that I wanted to read it sometime and added it to my bookshelf, and one year and eight months have passed :-) A spoiler (and my opinions) are as follows: The Murders in the Rue Morgue... The criminal is an animal! Therefore, the word choice does not match. The location of the killing is described in minute detail, and I begin to feel sick when I imagine it. I don't quite understand the device of the window...

Review comment 2 ("The Murders in the Rue Morgue")
This marks the beginning of the modern detective story and is quite complete. The Murders in the Rue Morgue is like a monument. I was surprised at the unexpected ending. It is closer to a prohibited technique. I think there may be quintessential entertainment characteristics in a detective story of reaching a totally unpredictable ending that no one could have guessed is more important than how you deceive the reader. What Dupain says is difficult to understand because of his elevated vocabulary muddles the explanation. He begins with the essential facts, and then gets into the details, and moves on to a sermon, and (at last) advances to demystification.

VI. Conclusion

In this study, we proposed a method of detecting spoilers from review comments written about items including story. Using story documents, we investigate where the words related to spoilers occur in the story line. As a result of investigating spoilers in story documents, words related to spoilers tend to occur toward the latter half of the story documents. In addition, we detect words that occur toward the latter half of the story documents from the review comments and perform qualitative analysis on how the words are used in the review comments. Although the proposed method cannot support words that are not directly used in the story documents, it may be possible to detect spoilers if reviewers describe spoilers using words that occur in the story document. On the other hand, our target language is Japanese and target domain is novel. Therefore, our results may not be represented in other language and among other domain such as comic or movie. This may need to be examined.

Future work aims to perform an experiment to analyze quantitatively how our proposed method is useful when a user watches a review comment. In addition, we plan to detect words that are not directly used in the story document and improve the precision of the detection of spoilers.

Acknowledgments

This work was supported by a Grant-in-Aid for challenging exploratory research 15K12150.

References

[1] Loewenstein, G.: The psychology of curiosity: A review and reinterpretation. Psychological Bulletin, Vol.116, No.1, pp.75–98 (1994).

[2] Wilson, T., Centerbar, D., Kermer, D., and Gilbert, D.: The pleasures of uncertainty: Prolonging positive moods in ways people do not anticipate, Journal of Personality and Social Psychology, Vol.88, No.1, pp.5–21 (2005).

[3] Tsang, A.S. and Yan, D.: Reducing the spoiler effect in experiential consumption, Advances in Consumer Research, Vol.36, pp.708–709 (2009).

[4] Guo, S. and Ramakrishnan, N.: Finding the Storyteller: Automatic spoiler tagging using linguistic cues, *Proc. of COLING '10*, pp.412–420 (2010).

[5] Iwai, H., Hijikata, Y., Ikeda, K., and Nishida, S.: Sentence-based plot classification for online review comments, *Proc. of IEEE/WIC/ACM WI'14*, pp.245–253 (2014).

[6] Pang, B. and Lee, L.: Opinion mining and sentiment analysis, Journal Foundations and Trends in Information Retrieval, Vol.2, No. 1-2, pp.1-135 (2008).

[7] Boyd-Graber, J., Glasgow, K., and Zajac, J.: Spoiler alert: Machine learning approaches to detect social media posts with revelatory information. ASIST 2013, Vol.50, No.1, pp.1–9 (2013).

[8] Klein, D. and Jackson D.: Processing content spoilers, U.S. Patent Application 20140101244 (2014–4–10).

[9] Golbeck, J.: The Twitter mute button: A web filtering challenge, *Proc. of CHI '12*, pp. 2755–2758 (2012).

[10] Nakamura, S. and Komatsu, T.: Temporal filtering system to reduce the risk of spoiling a user's enjoyment, *Proc. of IUI '07*, pp. 345–348 (2007).

[11] Shrout, P.E. and Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability, *Psychological Bulletin 86*, pp.420–428 (1979).

[12] Landis, J.R. and Koch G.G.: The measurement of observer agreement for categorical data, Biometrics, pp.159–174 (1977).