Empathy Factor Mining from Reader Comments of E-manga

Eisuke Ito
Research Institute for IT
Kyushu University
Fukuoka, Japan 819-0395
Email: ito.eisuke.523@m.kyushu-u.ac.jp

Yuya Honda
Grad. School of Library Science
Kyushu University
Fukuoka, Japan 819-0395
honda.yuuya.128@s.kyushu-u.ac.jp

Sachio Hirokawa
Research Institute for IT
Kyushu University
Fukuoka, Japan 819-0395
hirokawa@cc.kyushu-u.ac.jp

***Abstract*—The digitization of manga (Japanese comics) is currently progressing. In the past, expressions of manga have evolved in a way that is suitable for printing on paper. With the spread of smartphones, the expression of e-manga is developing at present. In this research, we focus on reader comments on e-mangas in Comico. Comico is a popular e-manga service. Similar to other online contents services, such as YouTube, Comico implements a user comments system. Readers can post comments easily, and the comments quickly reach others, including the manga creator. Comico recognizes that reader comments may influence the creator and the story of e-manga. In this research, we try to mine the empathy factor of readers for the story and the characters of e-manga. We collect reader comments and apply feature selection of SVM (Support Vector Machine) to mine empathy factors among readers.**

Keywords–online contents; e-manga; user comments; interaction; feature selection; SVM.

I. INTRODUCTION

The digitization of manga (Japanese comics) is currently progressing. In the past, expressions of manga have evolved in a way that is suitable for printing on paper. With the spread of smartphones, the expression of e-manga is developing at present. In the past, most e-manga were digitized by image scanning from paper manga. At present, some smartphone-oriented expressions of e-manga are developing. Many e-manga applications for smartphones are being released.

The authors of the present paper are interested in online contents services. We analyzed a video recommendation method using comments for videos on nicovideo.jp [1], and proposed statistical analysis of video page views [2]. In addition, we have been studying the number of bookmarks and recommendation of novels using the link structure of bookmarks for the online novel site syosetu.com. Recently, we also analyzed the diversity trend of online novels [3].

In this paper, we focus on reader comments on e-mangas in Comico [4]. Comico is a popular e-manga service in Japan, and e-mangas in Comico are specialized for smartphone. Among the e-manga smartphone applications, the Comico app is the second most popular in Japan as of February 2017. Similar to other online contents services, such as YouTube, Comico implements a user comments system. Readers can post comments easily, and comments quickly reach to others including the manga creator. Comico recognizes that reader comments may influence the creator and the story of e-manga.

If a system can mechanically extract the factors readers empathizes with, it may become a good support tool for creation of e-manga. Toward realization of empathy factor extraction, we apply the SVM (Support Vector Machine) feature selection method [5] to reader comments, and extract important words.

The rest of this paper is organized as follows. Section II shows related work. In Section III, we describe the service and contents of Comico briefly, and show some statistics. Section IV shows comment crawler and preprocessing. In Section VI, we illustrate our feature selection method and some results of analysis. Section V presents in detail of "ReLIFE", which is an e-manga on Comico. We use comments for this e-manga as the first analysis. We describe the empathy factor extraction using feature selection of SVM in Section VI. Finally, Section VIII presents a brief summary and future work.

II. RELATED WORK

Sentiment analysis of a story and of readers response is a hot topic of text mining. Murakami et al. proposed a method to associate characters in the story with frequently appeared words which represent character's personality [6]. They manually input character names and many co-occurred sentences in manga. Then they apply co-occurrence analysis. The readers response is out of their target of the analysis.

Emotional analysis of manga and manga readers is another new genre gaining much attention. Writers and readers can use a new style of visual communication different from textual words. In [7], Cohn and Ehly analyze the visual vocabulary appearing in the Japanese manga. They extracted 73 visual expressions that have been used in Japanese manga. It shows that these are also used in 20 volumes selected from boys and girls manga. In our case, many emojes (pictograms) are included in comments, but we are not addressing emotional expressions in this paper. When analyzing emotions, Cohn's study can be used.

Instead of using words in comments, stamps are gaining much attention in casual communication such as WathsApp [8] and LINE [9]. For example, Dharma et al. are studying how to use manga (cartoon) for communication in SNS (Social Network System) [10]. SNS is the world's largest community. It realizes various kinds of two-way communication and various news distribution. Dharma et al. proposed manga picture in SNS to improve their satisfaction in hobby fields. The Comico reader comment covered by this paper unilaterally describes the thought of the reader, so it is not used for interaction between readers or between authors and readers.

III. COMICO

In this section, we describe our research target e-manga service "Comico".

 ISBN: 978-1-61208-620-0

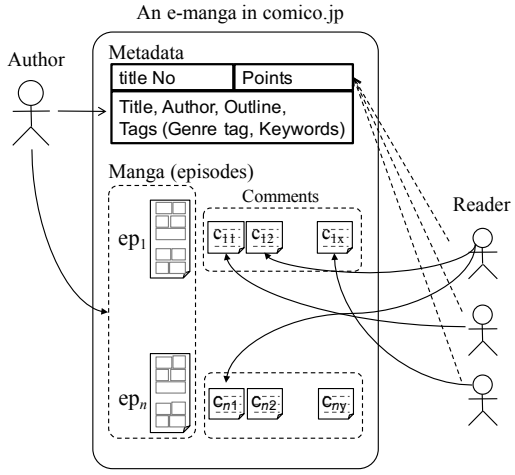


Figure 1. Data structure of an e-manga in Comico

A. Features of Comico

Comico [4] was established by "NHN comico" company in October 2013. According to a survey in February 2017 by Nielsen [11], Comico is the second largest user in Japan, and about 2.6 million people use it. (The first is LINE manga, 2.99 million users).

The most unique feature of Comico different from other e-manga services is that Comic only serves smartphone oriented original e-mangas. Almost all e-mangas on Comico are full-colored, and all e-mangas on Comico are able to browse just by swiping it from the top to the bottom. Comico does not serve e-manga which digitized past paper manga. For this reason, it does not need to move e-manga in the left-right direction, like a paper book.

Another feature is the use of reader comments. According to a news release from Comico, some e-manga creator used comments to measure reader's impression and used comments as a tool of interaction between readers and the creator.

Comico e-mangas are divided into three ranks: official, best_challenge and challenge. Newcomers' manga is serialized as a challenge work. When popularity comes out, then it is elected as best_challenge rank. When an e-manga in best_challenge rank is given quality and popularity, it arises to the official rank. Table I shows restrictions of comico e-manga in each rank, and Table II shows the number of e-mangas for each type.

TABLE I. RESTRICTIONS OF COMICO E-MANGA

Rank	Original	Min cuts	Color	New ep. freq.
official	Must	30	Must	Every two weeks
best_challenge	Should	15	Must	none
challenge	May	1	Should	none

B. Genre and tags

To improve searching and grouping, every e-manga on Comico is given tags by the manga creator. A tag represents genre, rank in Comico (official or challenge), and keywords. A tag is used like a hashtag in Twitter (a tag forms # and a word). Table III shows 12 genre tags used in Comico. Table IV shows the number of cartoon works with each genre tag attached. The total number of e-manga in Table IV is larger than the total number of works shown in Table II because most e-mangas are given multiple genre tags to hit a search query.

TABLE II. THE NUMBER OF E-MANGAS (OCT. 2017)

Type	Number
Continuing series	260
End series	188
Deleted series	9

TABLE III. 12 GENRE TAGS (Oct.2017)

Drama, Gag/Comedy, Common life, School, Love romance, Fantasy/SF, Horror/Mystery, Action, History, Sports, Essay, Omnibus

TABLE IV. NUMBER OF E-MANGA FOR EACH TAG. (Oct.2017)

Genre tag	Num. of e-manga
Drama	167
Gag/Comedy	180
Common life	154
School	120
Love romance	136
Fantasy/SF	140
Horror/Mystery	43
Action	54
History	10
Sports	7
Essay	25
Omnibus	3

IV. COMMENTS ANALYSIS

A. Comment crawler

To collect comments mechanically, we made a comment crawler program using Python. This crawler accesses the comments page of Comico by specifying title ID and episode ID of the e-manga series, and gets HTML of the web page. It extracts comment text using XPath from the obtained HTML page.

The collected comments are reformed and put into a search engine for later analysis. We extract the comment author name, the date and time of comment post, and words in a comment using the morphological analysis tool MeCab [12]. We also implemented e-manga character name extraction function. Characters are often described by nicknames or abbreviations, so we assigned candidates of nicknames or abbreviations for each character.

B. Limitation of comments

As shown in Figure 1, any user can post a comment for an episode, if the episode is open. So, the older the episode is, the more comments are posted. We limited comments to 7 days posted comment for fair analysis. Let $\boldsymbol{C}_i =< c_1, c_2, ..., c_n >$ be the set of comments posted to i-th episode (ep.i). Posting date of c_n is 7 days later than the date of c_1.

C. Attention to characters

We believe that the reader's empathy for e-mangas influences manga. We focus on reader's response towards the characteristic of hero and heroines of the manga. Kazuo Koike, who is a famous story creator, said at the beginning of his

book [13]. "Manga is character. If the character is catchy, the manga will be popular. If you create a catchy character, the character will be loved by many readers. And readers want to meet the character, then the manga will be sold and series of the manga will continue." We agree with his opinion and investigate whether the readers empathize with the characters of manga.

In this research, we consider that if a character name appears in a reader comment, then the commenter may have empathy with the character. In real world events such as sports, entertainments, and politics, people may show support for a person, team or party by using written placards which display their name. When a person empathizes with a person, a team, or a party, then they write their names on placard to show it. In other words, describing the name of an entity indicates that the descriptor may be interested in the entity.

D. Character frequency

We take $df(K)$ as the metrics of readers empathy to character K, where $df(K)$ is the document frequency of character K. In other words, $df(K)$ is the number of comments which include K's name at least one time in the comment. Most comments are short, and there are few comments which describes a character name multiple times. Also, there are few commenter who posts multiple comments to one episode.

We also count frequency of co-occurrence of character K and K', and we represent it as $df(K, K')$. In the story of manga (also novel or movie), the relation between two characters is important. There are cases where two characters become lovers, friends, or sports opponents. In manga dealing with love, the relationship between the two is important because readers are very interested and empathize with the progress of their relationship.

V. ReLIFE comments analysis

This section describes the analysis of ReLIFE comments. ReLIFE is an e-manga series on Comico, created by Yayoiso [14], and it is ongoing Japanese science fantasy high school drama [15]. ReLIFE is the most popular e-manga in Comico. Many users read it, and therefore many comments are posted. So, this is the best example for the first comments analysis.

A. Trend of the number of comments

Figure 2 shows the trend of $|\boldsymbol{C}_i|$ for each episode. $|\boldsymbol{C}_i|$ is the number of posted comments to ep.i $(i = 1..215)$ of ReLIFE series.

Figure 2 is marked A – E for remarkable spikes. Until ep.24, the number of comments is small. ReLIFE was not popular in the early stage of the series, then a few readers posted comments. In period A (ep.25–27), an event to present gifts to commenters was held. So, a lot of comments are posted. In period B (ep.39–40), the author notified that the paper book of ReLIFE will release. At D (ep.118), the number of comments is the highest. The story reached its first climax. At E (ep.144), the author notified that ReLIFE will be made a TV animation drama. At F (ep.187), the author notified that ReLIFE will be made into a movie. So, we found that ReLIFE readers reacted to the climax of the story, announcements of TV animation dorama and movie, or to the event which the creator / the company prepared.

B. Trend of characters

Table V shows the main 6 characters of ReLIFE. They are high school students in the same grade and in the same class. They look like 17 years old, but characters 1 and 3 are adults (27 years old), because they are rejuvenated with a special medicine 'relife'.

TABLE V. MAIN CHATACTOERS OF RELIFE

No	Name	Sex
1	Arata Kaizaki	M
2	Chizuru Hishiro	F
3	Ryo Yoake	M
4	Rena Kariu	F
5	Kazuomi Oga	M
6	An Onoya	F

Figure 5 shows $df(K)$ of character K for each ReLIFE episode (ep.1–127). To quantitatively measure readers' empathy for characters, we consider frequency (the number of appearances) of character name in comments. Character 1 "Kaizaki" is the main character, then he is generally frequent. In ep.27, sub-character Oga (character 5) is rapidly increasing. The reason will be clarified in Section VII-A.

Figure 6 shows $df(K, K')$ for each episode (ep.1–127). $df(K, K')$ is document frequency of co-occurrence of character K and K' in comments. There are 15 combination pairs for the characters 1 to 6 in Table V, and investigated co-occurrence frequency of all 15 pairs. Figure 6 shows the highest co-occurring eight pairs.

VI. Feature selection

In previous section, we understood what the reader empathized with, because we knew ReLIFE and checked the comments in detail. However, it is not efficient to apply manual analysis to all e-mangas. If a system can mechanically extract the factors of why reader empathizes, it may become a good support tool for creation of e-manga.

It is well known that SVM (Support Vector Machine) is a machine learning method that achieves good prediction performance compared with other methods. Moreover, SVM shows superior results when we combine it with feature selection. For example, Wariish et al. analyzed sales report using SVM, decision tree, and random forests [16]. The best result was obtained with SVM and feature selection.

Toward realization of the empathy factor extraction, we apply the SVM feature selection method [5] to reader comments and extract important words. We used SVM-light [17] with liner kernel, and with default parameters. We applied SVM assuming that all the comments containing a character K as positive data, and other comments are negative data.

A. Bag-of-words vectorization

To apply SVM, each comment must be vectorized. We vectorize comments using bag-of-words [18]. Figure 3 shows the outline of vectorization of documents (comments).

B. Extraction of feature words

We extracted feature words from the learning result of SVM [18]. SVM is a common method of machine learning for binary classification. Let input data be $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$,

 ISBN: 978-1-61208-620-0

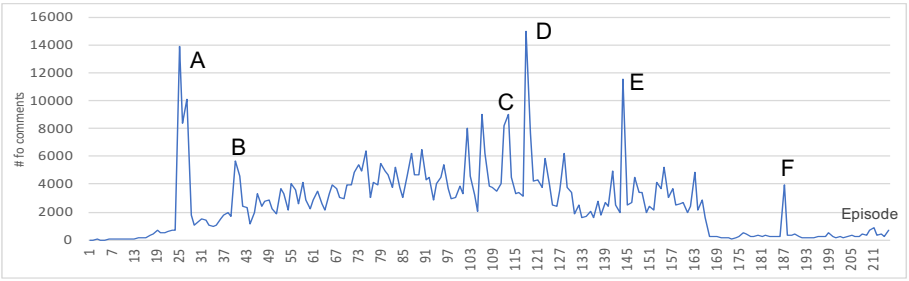


Figure 2. Number of comments for each ReLIFE episode (ep.1–ep.215)

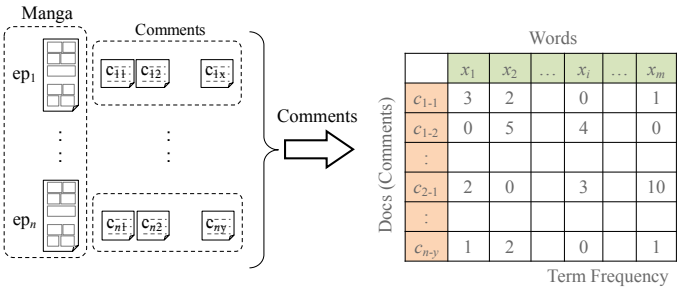
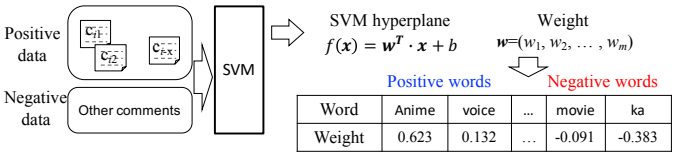


Figure 3. BoW vectorization of documents



Word	Anime	voice	...	movie	ka
Weight	0.623	0.132	...	-0.091	-0.383

Figure 4. Feature words selection using SVM

and two classes be +1 (positive) and -1 (negative). Then boundary function $f(\boldsymbol{x})$ of SVM is described as follows:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b, \tag{1}$$

where $\boldsymbol{w} = (w_1, w_2, \cdots, w_m)$ is weight vector and b is threshold. $\boldsymbol{w}$ and b are obtained by SVM machine learning.

Binary classification works as follows. Given a vector $\boldsymbol{x}$, and calculate $f(\boldsymbol{x})$ using obtained weight vector and threshold. If $f(\boldsymbol{x}) \geq 0$, then $\boldsymbol{x}$ is classified as positive, else classified as negative.

Next, we explain the method of feature words selection. At first, vectorize documents, and train SVM using training vectors. Next, give positive examples and negative example to the SVM. After that, calculate the importance of a word by using the weight for the word. The importance of word x_i is weight w_i in the weight vector generated by SVM training.

In [16], an experiment of feature selection is also performed. Let $S = (x_1, x_2, \cdots, x_m)$ be a set of words rearranged in descending order of the importance of words. m is the number of unique words which appear in documents. Then, the specific procedure is as follows:

I. Let $W' = \phi$ and $i = 1$.

II. While $i \leq N$, do the following steps.

- If i is even: $W' = W' \cup x_{m+1-i/2}$,
  If i is odd : $W' = W' \cup x_{(i+1)/2}$.
- Vectorize documets only using the words in W'.
- Train SVM and evaluate classification performance of SVM by using 5-fold cross validation.
- $i = i + 1$ and go back to II.

It is able to specify important words by comparing the classification performance in case of SVM of all words or a part of words.

VII. Extracted Empathy Factor of Characters

First, we vectorized all comments as shown in Figure 3. The number of all reader comments are 559,685, and the number of dimensions of a vector was 104,302. This means that there are 104,302 unique words in the set of comments.

SVM feature selection method needs to divide a document set into positive examples and negative examples. To extract the empathy factor of a character K, we set that positive examples are comments which include character K at least once, and other comments are negative examples. Extracting positive feature words are co-occurred with character K in reader's comments.

In [5], the score of a word is determined from the SVM model that separates the positive data the negative data. The score of a word represents the distance from the hyperplane. The characteristic words of positive data have positive scores, and those of negative data have negative scores. For feature selection, we choose top N positive words and bottom N negative words, as shown in Figure 4.

We applied feature selection method for reader comments of ReLIFE and extracted feature words from reader comments as the empathy factors for the character K.

A. Empathy factor by feature words selection

Table VI shows extracted top 20 positive words for each ReLIFE characters. Words in Table VI may illustrate what image or impression readers have for the character K. In Table VI, words starting with lower case letters are translated into English, but words beginning with capital letters remain in Japanese. Original results are Japanese single words because comments are written in Japanese.

 ISBN: 978-1-61208-620-0

Words for characters 2, 3 and 5 (Hishiro, Yoake and Oga) are easy to understand. These words express character's personality, appearance, and the role in the story of. In the row of character 5 (Oga), the word "Ogre" frequently appears because readers enjoy giving a nickname of "Ogre" to character 5 (Oga). On the other hand, characters 1, 4 and 6 (Kaizaki, Kariu and Onoya) are not easy. Character 1 (Kaizaki) is the main character, then he plays various roles in order to entered various situations. The role of character 6 (Onoya) is not yet decided in the story.

B. Performance of feature words selection

Table VII shows the results of classification performance of SVM. The 4th column in Table VII shows the number of words when F-measure becomes the highest. SVM classification accuracy may become highest when using all attributes. However, classification accuracy is higher when using limited effective attributes. So, we proposed a method to determine the optimum number of attributes in the SVM classification.

In Table VII, the number of words for max F-measure for characters 2 and 3 (Hishiro and Yoake) are both 8, and F-measure are 0.7900 and 0.8399. This means that characters 2 and 3 are able to present only 8 words. In other words, readers represent them using only 8 words. Actuary, the role of character 3 (Yoake) is facilitator of the story. Therefore, there are a few complicated topics for him. On the other hand, character 2 (Hishiro) is the main heroine. She is drawn as a cool beauty girl, and she does not act dynamically in the early part of the story. So, readers refer to her in simple expressions.

In case of character 4 (Kariu) in Table VII, the number of words for max F-measure is 20, and the value of F-measure is 0.8981. She is the highest F-measure in this analysis. Mr. Oga (character 5), his nick name is "Ogre" the number of words for max F-measure is 30, and the value of F-measure is 0.7924. In the first half of the story, their personality is clearly described, and the two will start dating as a lover. Readers were empathized in their life story and represented them in some words.

On the other hand, characters 1 and 6 (Kaizaki and Onoya) are not clear. Their F-measures are not high, and the number of words for max F-measure is 70.

VIII. Conclusion

In this paper, we focus on reader comments on e-mangas in Comico. We made a comment crawler program using Python, and collected a lot of comments from Comico. We showed statistical trend of comments and document frequency of characters. Human relations are important in stories, therefore we counted co-occurrence frequency of the two characters. We found that there are some spikes for the number of comments, and we estimated the reason for the rapid increase of comments.

Toward realization of the mechanical empathy factor extraction for support of e-manga creation of e-manga, we apply the SVM feature selection method to reader comments and extract important words. As the result of feature selection, we extracted words which represent each character. In other words, what the readers think about each character.

In the future, we want to expand empathy analysis to other e-mangas. More than 450 e-mangas exist on Comico. We hope to find a difference in empathy factors by genre. We also want to extract the transition of human relations.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15K00451.

References

[1] N. Murakami and E. Ito, "Emotional video ranking based on user comments," in Proceedings of iiWAS2011. ACM, December 2011, pp. 499–502.

[2] K. Noguchi, T. Iida, and E. Ito, "An analysis of cgm contents pageview using sir model and gbm," in Proceedings of ICCTD2017, March 2017, pp. 19–21.

[3] E. Ito and Y. Honda, "Keyword diversity trend of consumer generated novels," in Proceedings of ICESS2017, 2017, pp. 140–147.

[4] N. comico, "Comico," http://comico.jp [retrieved Dec. 2017].

[5] T. Sakai and S. Hirokawa, "Feature words that classify problem sentence in scientific article," in Proceedings of iiWAS2012. ACM, 2012, pp. 360–367.

[6] H. Murakami, R. Kyogoku, and H. Ueda, "Creating character connections from manga,," in Proceedings of ICAART 2011, 2011, pp. 677–680.

[7] N. Cohn and S. Ehly, "The vocabulary of manga: Visual morphology in dialects of japanese visual language," Journal of Pragmatics, vol. 92, 2016, pp. 17–29.

[8] "WhatsApp," https://www.whatsapp.com/ [retrieved Dec. 2017].

[9] "LINE," https://line.me/en/ [retrieved Dec. 2017].

[10] A. A. G. Dharma, H. Kumamoto, S. Kochi, N. Kudo, W. Guowei, C. Shu-Chuan, and K. Tomimatsu, "The utilization of social networking service and japanese manga in strategic user generated design," in Proceedings of ICEEI 2011, 2011, pp. 1–6.

[11] "Nielsen netrating," http://www.netratings.co.jp/news_release/2017/03/Newsrelease20170328.html, March 2017 [retrieved Dec. 2017].

[12] T. Kudou, "Mecab," http://taku910.github.io/mecab/ [retrieved Dec. 2017].

[13] K. Koike, Kazuo Koike's a new theory of characters. Goma Books (ISBN-10: 481491332X), August 2017.

[14] Yayoiso, "Relife," http://www.comico.jp/articleList.nhn?titleNo=2 [retrieved Dec. 2017].

[15] "Relife wiki," http://relife.wikia.com/wiki/ReLIFE_Wiki [retrieved Dec. 2017].

[16] N. Wariishi, S. Mitarai, T. Suzuki, and S. Hirokawa, "Text mining of daily sales reports," in Proceedings of AROB 2015, 2015, pp. 430–435.

[17] "SVM-Light," http://svmlight.joachims.org/ [retrieved Dec. 2017].

[18] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers Norwell, 2002.

 ISBN: 978-1-61208-620-0

TABLE VI. EXTRACTED TOP 20 POSITIVE WORDS FOR EACH RELIFE CHARACTERS

No	Name	Top 20 positive words
1	Arata Kaizaki	KUN, SAN, SENPAI (elder person), out, Mr., &, HANNOU (reaction), follow, trauma, spokesman, nice, adult, reaction, KAI, cool, same age, reader, liked, cool, boomerang
2	Chizuru Hishiro	SHIRO, N, a subject, HI, growth, insensitive, SAN, KAWAII(pretty, in-Kanji), hair, KAWAII (in-Hiragana), HISHIRON, KAWAI-, No., incorrect, communi-, NITA (laughing), gum, smile, SETSU, action
3	Ryo Yoake	end, girl, time, S (sadistic), KOTO, child, boy, SAN, elder sister, AKE, CHAN, revenge, AKE-YO, charge, love, KUDASAI (please), looks, plain clothe, job, moment
4	Rena Kariu	SHIRE, game, KIRE, injured, RIU, TAMARAI, TAMA, effort, RARE, NARE, poor, KARI, return, gentle, claim, a big meeting, TSUN-DERE, practice, retire, CHAN
5	Kazuomi Oga	GA, pure_ogre, ogre_handsome, gap, cv, KUN, pure, sensitive, tone-deaf, dull_ogre, family, awake, annoy_ogre, CHARA, faint_ogre, or, sport, insensitive, ogre_pure
6	An Onoya	CHAN, you, ANMA, training, charge, ONO, combination, support, cv, outside, &, anti, pair, or, convenience store, go out together, doubtful, plan, gal, be stuck on

TABLE VII. FS RESULTS: MAX WORDS AND CLASSIFICATION PERFORMANCE

No	Name	sex	# of comments	Num. of words max F-meatsure	Precision	Recall	F-measure	Accuracy
1	Arata Kaizaki	M	79,323	70	0.5379	0.8854	0.6690	0.6987
2	Chizuru Hishiro	F	96,921	8	0.7084	0.8990	0.7900	0.8250
3	Ryo Yoake	M	56,788	8	0.8092	0.9040	0.8399	0.7675
4	Rena Kariu	F	46,352	20	0.8591	0.9419	0.8981	0.8217
5	Kazuomi Oga	M	35,152	30	0.7073	0.9043	0.7924	0.7304
6	An Onoya	F	39,066	70	0.4637	0.7487	0.5724	0.7587

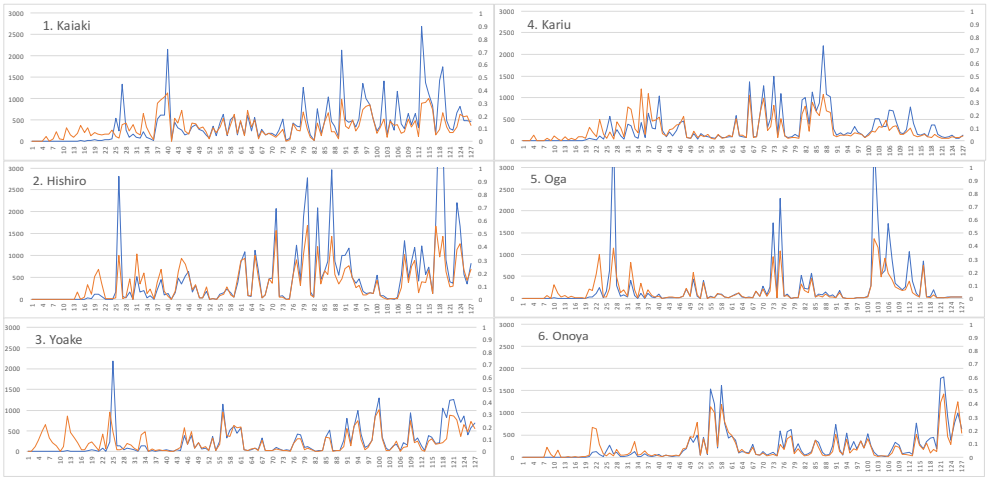


Figure 5. $df(K)$ of character K for each ReLIFE episode (ep.1–127)

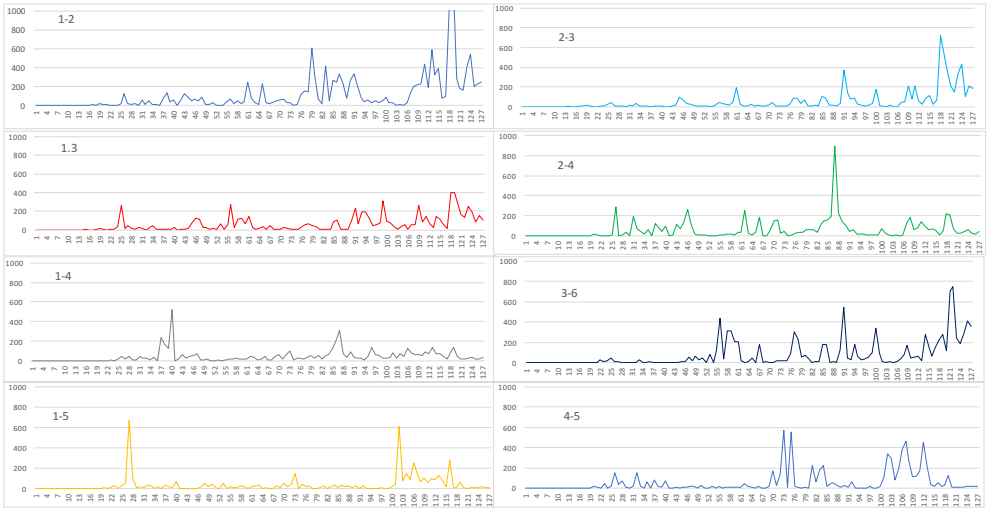


Figure 6. $df(K, K')$ of two charactors K and K' for each ReLIFE episode (ep.1–127)