Understanding the Users and Videos by Mining a Novel *Danmu* Dataset

Guangyi Lv, Kun Zhang, Le Wu *Member, IEEE*, Enhong Chen, *Senior Member, IEEE*, Tong Xu *Member, IEEE*, Qi Liu *Member, IEEE*, and Weidong He

Abstract—Recent years have witnessed a successful rise of the time synchronized *gossiping comment*, or so-called danmu combined with online videos. This new business mode has enriched communication among users by sending users' feelings through danmus and sharing these danmus on time synchronized videos. Can danmu communication be helpful for better user behavior modeling or video analyzing? To this question, in this paper, preliminary attempts are made on analysis of users and videos by introducing a *Danmu* dataset which is collected from a real-world danmu-enabled video sharing platform. The dataset contains 1.7 TB of videos and danmus in total across 8 video categories. With a focus on the 7.9 million danmu records and 4.8 million video frames, we first perform the basic statistic analysis and high-level semantic analysis. After that, we show some of the previous work on this area, including user behavior modeling, fine-grained video understanding and labeling, video plot generation and image-enhanced semantic understanding. For each application, we also propose its possible future directions. We hope this new dataset will inspire new ideas in areas among language, multimedia and user understanding.

Index Terms—Danmu, comment, video, user-generated.

1 Introduction

Recent years have witnessed a boom of online video-sharing sites [1], which play an important role in people's daily life. According to the forecast by Cisco, online video will hold more than 80% of consumers' internet traffic by 2020 [2]. Consequently, considering the fierce competition among video websites, improving the watching experience becomes the most important strategy to attract and retain users. Meanwhile, thanks to the emergence of the novel time-sync comments, or so-called *Danmu* [3], real-time comments on video shots have become more and more popular, e.g., niconico[1] in Japan and Bilibili[2] in China. Taking Bilibili as an example, it is revealed that it has a total number of 1.4 billion danmus, and it is still growing at an incredible rate of three million per day. Different from traditional online reviews which are displayed in a separate space outside the video (e.g., Youtube[3]), danmu is a new type of comment that is overlaid directly on the top of videos by synchronizing the comment with specific playback time. Besides, danmus are also different from subtitles. Subtitles are provided by the video publisher, while danmus are generated by viewers and the amount of text and the number of topics are far more larger than that of subtitles. As shown in Fig.1, the word *Danmu* literally means "bullets on the screen", because it shows a similar scene of comments flying across the screen [4].

Fig. 1. The danmu-enabled video and the gossiping behavior.

As a matter of fact, by sharing danmus directly on the time synchronized videos, this new type of business mode has increased user activity, such as comments and views, on these platforms. For example, *Bilibili* said that it was just the special experience of danmus that led the platform to have 71.8 million monthly active users in 2017, which is 2.5 times the number in 2016.[4] Intuitively, through reading danmus, audiences can involve directly and share their opinions with others. Therefore, whether a user watches a video is not only determined by the video content, but also affected by other audiences' previous interactions with the video. In other words, as audiences have got accustomed to danmu-styled interactions, a video itself is no longer the only factor that attracts users. In addition, audiences are more eager to see others' opinions or even to gossip about several typical plots or characters, as shown in Fig. 1.

In fact, though danmus are regarded as comments by audiences, they are more than ordinary text data. They have several characteristics that help build relationships among language, multimedia and human behavior. First, there are many special expressions in danmus, for example, "2333"

G. Lv, K. Zhang, E. Chen (Corresponding author), T. Xu, Q. Liu, and W. He are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China. (email: gylv, zhkun, hwd@mail.ustc.edu.cn, tongxu, qiliuql, cheneh@ustc.edu.cn).
L. Wu is with the School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230029, China. (email: lewu@hfut.edu.cn)

1. http://www.nicovideo.jp
2. http://www.bilibili.com
3. http://www.youtube.com
4. https://www.ithome.com/html/it/349567.htm

represents laughing[5] while "*high energy*"(literal translation) indicates a climax[6] ahead. These special expressions are associated with users' emotions and interest. Second, though danmus are time synchronized comments on videos, these language expressions are not limited to the Question Answering pattern. They are style-free user opinions which are correlated to episodes, music, several specific actors or real-time events. Last but not least, danmus have several characteristics when they are presented to users, e.g., special color, appearance mode and several special emojis, which help to reveal a user's or a community's behaviors. In summary, those characteristics show danmus ' potential in domains such as Natural Language Processing (NLP), media understanding and user modeling in academic and industry. Several prior studies have made preliminary attempts to apply danmus to various domains, such as language analyze [5], [6], semantically comments understanding [7], comment based video understanding [4], or user behavior modeling [8], [9]. These studies focused on a unique aspect of danmus— how to comprehensively understand danmus, but their potential research and application values is sill being explored.

To this end, in this paper, we put forward a novel danmu dataset [7] that provides a possibility of deeper understanding to users and videos. The dataset is collected from the large danmu-enabled video platform *Bilibili*. It contains 1.7 TB of videos and danmus in total across 8 kinds of video categories, including 7.9 million danmu records and 4.8 million video frames. To understand this dataset comprehensively, we first give a detailed description to data types, followed by statistical analysis. Then, we focus on the textual expressions of danmus with high level semantic analysis. After that, we provide several research tasks of the dataset, including: user behavior analysis in danmu-enabled platforms, video analysis and modeling with danmus, and the image-enhanced sentence understanding. For each research task, we also introduce key ideas and the possible future research directions along this line briefly. We hope the release of this dataset could shed light on more future research works on it.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work briefly. Section 3 presents basic information of the dataset, including data collecting, pre-processing and its format. Sections 4 and 5 show deeper analysis of danmu data on statistics, semantic and human behavior levels, which is expected to be useful in inspiring new researches. Next, we report and analyze typical applications of danmu and danmu-enabled videos in Section 6. Finally, we conclude this paper in Section 7.

2 Related work

In this section, we review the related works from two aspects: 1) Existing Vision & Language datasets; 2) Applications of Time-sync Video Comments.

5. http://www.top-news.top/news-12505087.html
6. http://www.szdaily.com/content/2015-11/23/content_12510781.htm
7. Available on http://bigdata.ustc.edu.cn/dataset/Danmus

2.1 Vision & Language Datasets

Bringing vision and language elements into an intelligent system has long attracted a large group of researchers. A proper dataset is of vital importance to modern models [10], [11] which provide end-to-end methods, i.e., combining deep convolutional networks and recurrent neural networks as autoregressive models, and currently dominate this area. Most available datasets contain both vision and language information that mainly focus on the task called Image Caption [12], [13], [14] or Video Description [15], [16], [17], which aims to generate "meaningful" descriptions from images or episodes, where the descriptions are strictly based on objects or attributes. Those vision & language datasets often provide one or multiple captions per image. The captions of these datasets are either the original photo title and descriptions provided by online users, or are generated by crowd workers for existing images. In contrast, video datasets [18], [19] aligned with descriptions, generally represent limited domains and small lexicons, which is due to the fact that video processing and understanding is a very compute-intensive task. In addition, there are also some video datasets [20], [21] containing tag/label information. They do not include text data and usually have a dedicated usage, such as video classification [22], [23] and human motion/activity recognition [24], [25].

However, no matter the above datasets are collected from social networks or labeled by crowd workers, they are more concerned with the objective relationship between vision and language other than the users' subjective opinions or behavior. In contrast, since the time-sync comments are sent to a certain video by millions of real audiences at any time, they always indicate users' immediate interest better. And, this immediate interest can be used to model the relationships between users and media. Moreover, some of the unique properties (e.g., frequency, position, color, or emojis) of danmu data can also reveal users' special behaviors, which are useful in some important searches on social media [26], [27], [28]. Above all, *Danmu* dataset tends to be larger in size and contain more contextual information, which is bringing media, language and user together.

2.2 Applications of Time-sync Video Comments

Time-sync comment, which has received growing research interests, is a new interactive mode and provides a new source of information regarding the video. Early works mainly focused on statistics for comments and the co-relation between comments and videos. Lin *et al.* [5] analyzed comments sent by Chinese users when they were watching Japanese movie. They proposed a statistical method to identify whether a Chinese word is a loanword from Japanese or not based on danmus. In [6], the author investigated co-relation between emotional comments and popularity of a video.

In contrast, recent works tend to process time-sync comments on a more complex semantic level. A Temporal and Personalized Topic Model (TPTM) [7] is used to generate time-sync video tags for videos using comments only. Furthermore, with the success of deep learning technologies in various domains, modeling time-sync comments with neural networks is considered to be a new fashion. In [4],

a semantic embedding technology is involved to perform temporal-labeling for video shots. They first represent time-sync comments into semantic vectors, then a video splitting framework is designed to extract and label meaningful segments based on mapping the semantic vectors to predefined labels in a supervised way. Based on the fact that time-sync comments are semantically related to the corresponding frame, danmus are further leveraged to learn image-enhanced model to perform Natural Language Inference (NLI) tasks [29].

Above all, time-sync comments or danmus, as a new interaction mode for online videos, are gradually supporting the needs of various researches. They have great potential value in both academia and industry.

3 DATASET

In this section, we make a brief introduction to the *Danmu* dataset, which will cover these three following aspects: 1) The source and methods for data collections; 2) Preprocessing of the the videos and danmus; 3) The structured data format.

3.1 Data Collections

To understand videos and user behaviors comprehensively, we collect data from Bilibili[8], which is one of the largest danmu-enabled video sharing platforms in China. We crawl videos and danmus through web pages that are available to the public. The original data contains 1.6 TB of videos which were published in the last 7 years from 2011 to 2018. Those videos have been also watched and commented with $7,924,272$ danmus by $2,097,079$ different users. In order to satisfy various research requirements, those data are crawled from the following 8 categories: *Anime, Movie, Dance, Music, Play, Technology, Sport*, and *Show*.

- **Movie:** This category includes classic movies from all over the world. As the eighth art form, movie shows us stories with rich plots by depicting different scenes and showing different relations between characters. Moreover, a movie usually lasts 1 or 2 hours, during which there are a large number of scene changes and plot fluctuations.
- **Anime:** This category contains Japanese animations (a style of hand-drawn and computer animation) which shows more exaggerated plots and personifies many objects. Moreover, as a typical representative of the ACG (i.e., Anime, Comic and Games) culture, anime contains plenty of domain knowledge, which reveals current popular contents. This kind of information is highly diverse in various aspects (complexity, expression, etc), posing plentiful challenges to language and images.
- **Dance:** This category refers to a special channel in Bilibili, where the videos are user-uploaded ones with content of dances accompanied by ACG related music. This kind of videos do not contain specific plots or stories. Most of them intend to show current popular ACG dances that are originally shown by animation characters.
- **Music:** This category is mainly composed of animated songs or pure music, and is accompanied by user generated MV extracted from a specific video.
- **Play:** This category mainly focuses on user-generated instrumental videos, including piano, violin, and other niche musical instruments. Its scene barely changes.
- **Tech:** This category includes science and technology experiments presented in a simple and straightforward way. It mainly explains several common or unusual phenomena in the real world. Most of the videos in this category are less than 20 minutes.
- **Sport:** This category consists of different kinds of sports playback videos or sports related commentary videos. Part of these videos are complete sports events. Others are the clips of the exciting parts of the sports events.
- **Show:** This category mainly consists of different variety shows. As an important part of TV shows, variety shows draw plenty of people's attention, leading the current trend of fashion. This kind of videos include tremendous contents, such as stars, popular games, songs and so on.

There are still millions of videos that are quite different in various aspects. In order to model and evaluate danmus, videos, and user behaviours better, we crawl videos with higher danmusdensity, which can be defined as the number of danmus divide by the video's duration. Next, we will introduce the data pre-processing. The final version of the dataset can be downloaded on: "http://bigdata.ustc.edu.cn/dataset/Danmus".

3.2 Pre-processing

To protect the privacy of the original users who sent these danmus, all the user IDs are hashed. Besides, the raw video files are not released due to their copyrights[9]. If you need the raw videos for academic purpose, please contact us.

3.2.1 Danmu Text Translation

In order to adapt to more application, we intend to translate the danmus, which are Chinese comments, to English. Since our main focus are sentiment analysis and multimodal research, we did not focus on the translation of danmus. Thus, we take advantage of different available translation tools, such as Google Translation[10], Baidu Translation[11], as well as Youdao Translation[12], to translate the original Chinese danmus to English version.

Moreover, danmus contain plenty of domain-specific words, which cannot be easily translated with online tools. Thus, we first utilize Name Entity Recognition tool, i.e. Stanford CoreNLP[13], to label all the name entity in danmus. Next, we translate these name entities to pinyin. Then, these modified danmus are sent to translation tools to get their English version. Along this line, most of the domain-specified name entities are kept in the translation process, which is beneficial to semantic understanding and further research about danmus.

8. https://www.bilibili.com

9. We do not own the copyrights to the videos and the danmus. Their use is restricted to non-commercial research and educational purposes.

10. https://translate.google.cn/

11. https://fanyi.baidu.com/

12. http://fanyi.youdao.com/

13. https://stanfordnlp.github.io/CoreNLP/

3.2.2 Font Color Shrinking

We have noticed that every danmu has a color specified by its sender. The original danmu color is represented in 24-bits RGB form, which allows 16,777,216 color variations. The color space is too large for some statistical applications. To simplify the usage of color, we shrink the color space to 64. We achieve this by reducing the number of bits that are used by color channels. For example, suppose we have a 24-bits color denoted as $c_{23}c_{22}...c_0$, and the 3 channels are $c_{23}c_{22}...c_{16}$ for red, $c_{15}c_{14}...c_8$ for green and $c_7c_6...c_0$ for blue. To reduce the color space, for every channel, we take the highest two bits to form a new channel, i.e., $r = c_{23}c_{22}$ for red, $g = c_{15}c_{14}$ for green and $b = c_7c_6$ for blue. Note that the new channels (each has only 2 bits) can have at most 64 combinations. Then we repeat the new RGB channels 4 times to recover the 24-bits format $rrrrggggbbbb$ as the shrinked color.

3.2.3 Key Frame Extraction

As mentioned before, our raw data are collected from the Internet. To protect the copyright of the videos, the original video files are not available in this dataset. Instead, we provide video frames (images) to support computer vision related research. To be specific, first we split every video into images by 1 frame per second. Then, the frame will be re-sized with its height being 480 pixels (keep the original ratio). Finally we have $4,816,133$ frames in total.

Since the total number of frames is too large and comments are sparse (there is even no comment in many frames), key frame extraction is carried out to reduce the amount of data and increase the density of comments. To be specific, we first extract features for frames by constructing the scalable color descriptors (SCD) [30]. Then, based on these features, an affinity propagation algorithm [31] is performed to cluster the frames, and the kernels are collected as our key frames.

3.3 Data Format

Finally, we provide our dataset in three collections: Video-Meta, Danmus and Frames.

3.3.1 Video-Meta

This collection provides meta information of 4,435 videos, which include *category*, *video ID*, *title*, *description*, etc. There are also some rating attributes, e.g., the number of being *viewed* or *shared*, which are potentially helpful for researches devoted to recommendation or popularity prediction. A JSON formatted sample and the details of attributes are shown below.

```
{"category" : "play",
"video_id" : 10025115,
"title" : "[白衣少侠]核爆神曲Penbeat aLIEz",
"desc" : "希望大家喜欢，对朋友们表示真心的感谢",
"duration" : 298,
"pubdate" : 1492872098,
"favorite" : 12344,
"coin" : 19645,
"view" : 386245,
"share" : 6024}
```

- **category**: The category that the video belongs to. It can be one of {"*movie*", "*anime*", "*dance*", "*music*", "*play*", "*tech*", "*sport*", "*show*"} which separately indicates the eight types mentioned in Section 3.1.
- **video_id**: The video ID. The ID is in fact an integer comes from the original video URL of Bilibili[14]. Please note that many URLs may expire due to the update of the website.
- **title**: The video's title which is a UTF-8 charset string.
- **desc**: The video description which is written by the uploader. It is in string type with UTF-8 character set.
- **duration**: The video duration (in seconds).
- **pubdate**[15]: Publication date (unix time) of the video.
- **favorite**[15]: The number of times the video was added to user's collection.
- **coin**[15]: The number of coins the video obtained. When viewers like a video, they can give a coin to the video as a positive feedback.
- **view**[15] [16]: The number of times the video was played.
- **share**[15] [16]: The number of times the video was shared.

3.3.2 Danmus

Danmu texts are the main component of the dataset. There are a total of $7,242,272$ records in this collection. Below is a sample in JSON format:

```
{"category" : "dance",
"video_id" : 6683560,
"time" : 35.503,
"timestamp" : 1476515039,
"font_color" : "0xFFFFFF",
"font_color_64" : "0xFFFFFF"
"font_size" : 25,
"mode" : 1,
"text" : "啊，好可爱！",
"text_trans" : "Oh, so lovely!",
"user_id_hash" : "a29e273b"}
```

- **category** and **video_id**: Have the same meaning as the Video-Meta collection.
- **time**: Time in second the danmu shows in the video. It is a float number, e.g., $time = 35.503$ means this danmu appears on the screen at $35.503s$.
- **timestamp**: Date time (unix time) that the danmu was sent by a user.
- **font_color** and **font_color_64**: Font color represented by a hexadecimal string. A hexadecimal color is specified with: 0xRRGGBB. RR (red), GG (green) and BB (blue) are hexadecimal integers between 00 and FF specifying the intensity of the color. "Font_color_64" is the shrinked version as is described in section 3.2.2.
- **font_size**: Font size.
- **mode**: The appearance mode of the danmu. It can be one of {1, 4, 5, 7}, which indicates "rolling" (rolling from right side to left side), "bottom-most" (shows like traditional subtitles), "top-most", and "special" respectively.
- **text** and **text_trans**: The text of danmu. Both of them are UTF-8 strings. "Text_trans" is the English translation.
- **user_id_hash**: The hashed user ID. To protect users' privacy, the original user ID is not included. This attribute is mainly used to indicate whether some danmus are sent by the same user. We cannot get detailed user information.

14. "10025115" means the URL is https://www.bilibili.com/video/av10025115
15. Videos of "movie" category don't have this attribute.
16. Videos of "anime" category don't have this attribute.

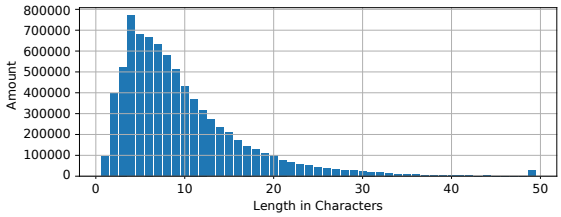


Fig. 2. The Length distribution of danmus.

3.3.3 Frames

This collection contains 4,816,133 frames in total. Each frame is zoomed out as 480 pixels height image. We can obtain the frame's corresponding danmus by the "time" attribute. A data sample and attributes details are listed as follow.

```
{"category" : "movie",
"video_id" : 1153046,
"time" : 38,
"file_name" : "00000018.jpg",
"data" : BinData(),
"key_frame" : true,
"start_time": 37,
"end_time": 39}
```

- **category** and **video_id**: Have the same meaning as the Video-Meta collection.
- **time**: Time of the frame in the video.
- **file_name**: A string that indicates the original file name of the frame image. The extension of the file name indicates the image format.
- **data**: Binary data of the image file.
- **key_frame**: A boolean value indicates if the frame is a key frame or not. See Section 3.2.3 for key frame extraction.
- **start_time** and **end_time**: If **key_frame** is "true", these two values indicate the range of the segment which is represented by the frame.

4 Statistics Analysis

In this section, we perform a series of statistics and analysis on various attributes of danmus, which consist of basic statistics and special attributes. By introducing details of this information, we can have a clearer look on the characteristics of this new type of comments.

4.1 Basic Statistics

First of all, we will introduce the basic statistics of danmus, including the number of distribution in each category over display time and the length distribution over display step.

Fig. 2 shows the length (number of characters) distribution of danmu sentences. From this figure, we can observe that their length are mainly between 3 to 11, which is roughly the length of one single sentence. Taking the format of danmus and audiences into consideration, it is natural that audiences only send one sentence at one time, and we found danmu sentences usually do not contain line breaks. This phenomenon is quite different from the traditional comment system. Moreover, we can find that short sentences hold a relatively high proportion. This phenomenon is closely related to the form of danmu-enabled videos. As mentioned in section 1, audiences send danmus while watching videos. Thus, short sentences are preferred to express their feelings about what they are watching. Furthermore, Fig. 2 shows that there are several sentences whose length in characters are all more than 40. This type of sentences can even form a paragraph. When checking these specific sentences, we find that most of them are only repeating the last word of the meaningful part. For example, the audiences tend to repeat "*3*" in "*2333*"(which has similar meaning to "*lol*") many times to emphasis that they think the videos are very interesting.

Besides the length of danmus, we also analyze the number distrubition of danmus over the playback time-line. Since the playback time of different types of videos are of great disparity, the numbers of danmus in different videos are not directly comparable. Therefore, we first normalize all danmus' display time by dividing them by the corresponding videos' duration, so that danmus' display time can be represented by the percentage of playback time. Then, we split them by 1% into 100 parts and count the number of danmus in each part. Fig. 3 demonstrates the number distribution of danmus in each category.

From these figures, we can obtain many interesting features of danmus. One common feature of these figures is that the number of danmus is the largest in the beginning. Then, it experiences a severe drop and paces down slowly. This can be treated as one common phenomenon in the online video playing. The audiences are attracted by various reasons (e.g., the title, the cover image) to watch the videos. Then, on one hand, part of them feel bored and turn off the videos. On the other hand, some viewers are immersed in and start to enjoy the videos. Thus, as a result, many of them stop sending comments, so we can find that danmus become fewer and smoother in the middle of the videos. This phenomenon shows us that the characteristics of danmus are closely related to user behaviors, on which we will conduct further investigation in section 5.4.

With the consideration of numbers of danmus in the beginning part, we can divide these figures into four parts. The *movie* and *show* categories have the most danmus. We can observe that they have more than 100 danmus in the beginning. The *movie* can even has more than 300 danmus. The following are *anime* and *sport* categories, which have more than 35 danmus in the beginning. The last are *tech* and *music* categories, which have less than 25 danmus in the beginning. This phenomenon reflects the attractiveness of different types of videos. Since the *movie* and *show* videos both have famous actors or actresses, there must be plenty of fans watching the videos. Moreover, *movie* often has interesting plots and *show* has interesting games or attractive anecdotal stories, therefore, it is natural that these types of videos attract much more audiences. On the contrary, the *tech* and *music* categories are short of attractive spots, thus their audiences are far less than the other categories.

When arriving at the end part of videos, the danmus of most videos keep still except for the *tech* category, which has a sudden rise and then drop quickly. As mentioned in section 3.1, videos in *tech* category often introduce different experiments to reveal some phenomena in the real world, so the process of the experiments are rather boring. When arriving the end of the experiments, the remaining audiences will not only comment the results, but also celebrate their persistence to the end, such as "*survival to the end*". Thus,

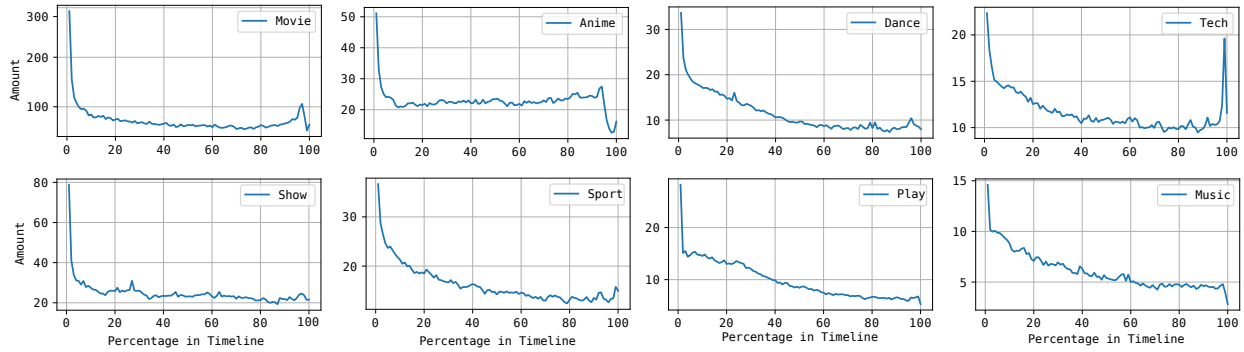


Fig. 3. The number distribution of danmus over display time.

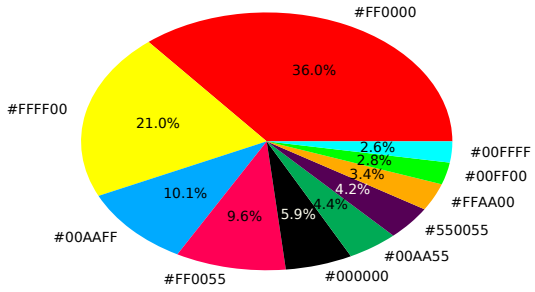


Fig. 4. The color distribution of danmus.

we can observe that danmus in this category arise suddenly in the end part. After that, the videos come to an end, so the danmus drop rapidly again.

4.2 Special Attributes of Danmus

As mentioned before, danmu , as a special comment on the videos, contains several special attributes. In this section, we will introduce three unique characteristic: 1) Color; 2) Display Type; 3) Emoji.

4.2.1 *Color*

Different from the traditional comments, audiences can choose a specific color when sending danmus. We collect all the colors of danmus and there are more than 300 different colors. In order to demonstrate different meanings of different colors clearly, we transfer these colors to the final 64 colors which we have introduced in Section 3.2. Then, we select the top 10 colors according to their number and visualize the results in Fig. 4 for better analysis. We have to note that we do not take color white into consideration because white is the default color which does not contain any special meaning.

As shown in Fig. 4, we can observe that color red and yellow are the two most commonly used colors. In the real world, color red is always used for hints or warnings. It can draw people's attention at first sight, so most people are very sensitive to this color. When it comes to videos, we find that most of the danmus whose color are red are particular type of comments such as "*high energy*" and "*forward warning*". This type of danmus are often used to remind plot changes, especially in horror films. Moreover, we observe that the red danmus appear more in the *movie* and *anime* categories which have rich plots, and appear less in *music* or *play* categories which are short of plots.

For color yellow, we find that audiences are likely to comment on several actors with this color. To be more specific, when watching the videos that are about several actors, such as Han Lu[17] and TFboys[18], audiences would like to praise or defend their idols with this color. This phenomenon inspired us that fans of different actors or actresses might be in favor of a specific color to distinguish their comments from others. Moreover, the color of danmus provides an opportunity for audiences to find those people who have similar interest to them quickly. At the same time, the colors of danmus also allow us to group similar users precisely. We can utilize these danmus with the same color to analyze the behaviour and interest of specific groups.

4.2.2 *Display Type*

Apart from the color, audiences can also choose different display types to express their opinions. Since the danmus are displayed upon the video frames, there are four different display types: 1) *Rolling danmus*: displaying the comments across the video frames from right to left; 2) *Top-most danmus*: displaying the comments by staying at the top of the videos for a period of time; 3) *Bottom-most danmus*: displaying the comments by staying at the bottom of the videos for a period of time; 4) *Special effect danmus*: displaying the comments with several special effects (e.g., fade in, fade out). Fig. 5 shows the proportions of different types of danmus in different categories.

Since the default display type is rolling, we can observe that it accounts for roughly the same proportions in different categories. When taking the top-most danmus into consideration, we can observe that the top-most danmus account for the largest proportion in *anime* category. As mentioned before, top-most danmus would stay at the top of the videos for a period of time. Thus, this type can attract audiences' attention more quickly and express owners' views more directly. This phenomenon often occurs when the scene in videos are far different from the reality. For example, the anime videos can be expressed in an extremely exaggerating way to attract audiences. Thus, we can observe that *anime* videos have the largest proportion of top-most danmus. On the contrary, most of the videos in other categories are based on the real world. This kind of scene is relatively few, and so are the top-most danmus in these videos.

17. https://en.wikipedia.org/wiki/Lu_Han
18. https://en.wikipedia.org/wiki/TFBoys

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2019.2950411, IEEE Transactions on Big Data

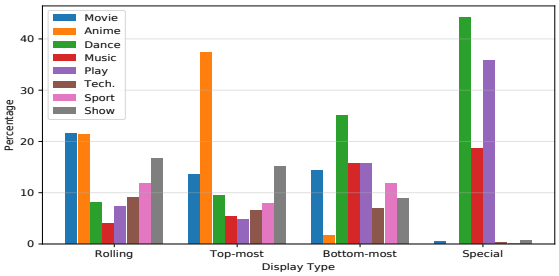


Fig. 5. The proportions of danmu types in different categories.

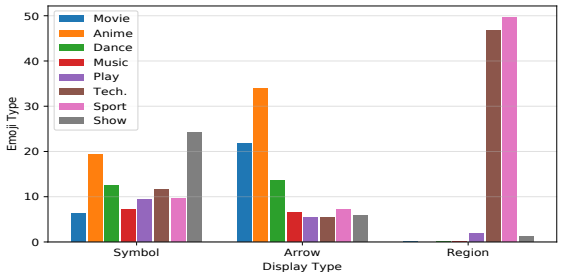


Fig. 6. The proportions of emojis types in different categories.

The bottom-most danmu is very similar to traditional subtitles, which is also displayed at the bottom of videos. In fact, the bottom-most danmus play the role of subtitle or lyrics to some extent. Meanwhile, without affecting the main content of videos, they provide necessary explanation for better video content understanding. Thus, we can observe that videos in *anime* and *movie* categories have the smallest proportion of the bottom-most danmus, because these videos have already been subtitled. Adding extra contents at the bottom may cover the subtitles and has a bad influence on the videos. However, the videos in *dance*, *music* and *play* categories do not have subtitles. The bottom-most danmus are capable of explaining dance movements, background music, and other contents in a detailed perspective. Therefore, we can observe that this type of danmus takes up the largest proportion in these videos.

As for the special effect danmus, they basically appear only in *dance*, *play* and *music* categories. We have mentioned that these videos usually have no plot and most of them are connected to specific music. At the same time, special effect danmus have exaggerated forms of expression, which can properly set off the atmosphere and enhance the appeal of music. However, the way they display may have a bad influence on the videos in *anime* or *movie* category because audiences need to pay attention to the plots. Thus, we can find they rarely appear in such videos. Generally speaking, these four different types of danmus reveal several special features of different videos. We can utilize these features to analyze and understand these videos better.

4.2.3 *Emoji*

With the development of social media, people are no longer satisfied with using only text to express their feelings. Images (or emojis) can be helpful for intuitive idea expression and easier understanding. Thus, more and more emojis appear on social media, and it is true in danmus as well. In order to analyze the emojis in danmus better, we divide them into three categories: 1) *Symbol*: utilizing traditional symbol characters to form a face-like word (e.g., [•_•]); 2) *Arrow*: utilizing different arrows to indicate the objects of their comments; 3) *National or regional flags*: utilizing different flags to express their love and support to a specific country or region. Fig. 6 demonstrates the proportions of different types of emojis in different categories.

Similar to rolling danmus, symbol emojis, which consist of traditional emojis, almost have the same proportions in different categories. Compared with the text, emojis can mobilize the atmosphere better and motivate the audiences to comment on the same objects. Since the density of climax plots in movies are less than the other videos, we can observe that symbol emojis account for a smaller proportion in *movie* category.

On the contrary, we can observe that videos in *anime* and *movie* categories have the largest proportions of arrow emojis in Fig. 6. As mentioned before, these two categories have plenty distinctive characteristics. Different audiences may be interested in different characters. Thus, they would send their comments when the corresponding characters shows. Moreover, danmus appear directly on the videos, so audiences can easily use arrows to point out the roles that they are interested in. Therefore, they would like to use arrows to comment on the roles they are interested in when watching videos of *anime* or *movie* categories. Videos in other categories do not have so many characters, so we can find these videos have smaller proportion of arrow emojis.

As for region emojis, they have a similar function to arrow emojis, which is to indicate audiences' love and support to specific objects. Different from the arrow emojis that aim at characters in the videos, region emojis, or so-called flag emojis, each demonstrates a specific nation. Thus, we can find that the videos in *sport* or *tech* categories have the largest proportion of this type of emojis. Different flags represent audiences' love and support to a specific team in *sports* videos or specific technology in *tech* videos. They can display different characteristics of videos very well.

5 Semantic Analysis

In this section, we focus on analyzing high-level semantics of danmus from three parts: 1) Styles and Topics, 2) Special Expressions, and 3) Danmus and User Behaviours.

5.1 The Style and Topics of Danmus

In this section, we apply topic models on the dataset to analyze danmu text in coarse-grained semantic level better. To be specific, we utilize the Gensim Tool[19] to generate topic information on danmus comments in each category. The parameters we select are as follows: The number of topics is set as 20. The number of topic words in each topic is set as 50. The iteration of model learning is set as 50.

After generating the topics of danmus, we extract three topics and several corresponding topic words to verify the semantic information of danmus. Table 1 shows part of the topic words in each topic of different video categories. We can observe that most of the videos contain the topic "*23333, lol*", which indicates that most of the videos include funny

19. https://radimrehurek.com/gensim/

TABLE 1
Part of topic words in several categories.

Category	Topic	Topic words	Category	Topic	Topic words
Movie	Topic 1	2333, Haha, cry with joy, in a trance	Anime	Topic 1	2333, Laugh, cry with joy, Kiss, Love
	Topic 2	High energy, Front, Pay attention		Topic 2	Throwing Flowers, The end, Bilibili, Headquarters
	Topic 3	Very cute, Beautiful, Voice, Aaa		Topic 3	Human, Vampire, Magician, Knight
Music	Topic 1	Pleasant to hear, Divine tune, Aaa	Play	Topic 1	Leave my name, Very good, Come on
	Topic 2	Faith, Railgun, Electro Master, Forever		Topic 2	Contract, Confession, Support,National anthem
	Topic 3	Unicorn, Royal Highness, Knights		Topic 3	Pipa, Guzheng, Piano, Flute
Tech	Topic 1	2333, Haha, cry with joy, in a trance	Show	Topic 1	Chaoyue Yang, Yangmu, Yihan Chen (Estelle), Xingjie Zhu
	Topic 2	High energy, Cell, Viruses, Microscope		Topic 2	TF boys, Yuan Wang, Lei Zi, Fall in love
	Topic 3	China, America, Japan, Nation		Topic 3	23333, hahaha, Degang Guo, 666

parts and these parts draw attention of audiences greatly. It is natural that audiences want to be pleased when watching videos. In *movie* category, there are more emotional words, such as "*high energy*", "*beautiful*" and "*Aaa*". This kind of phenomenon demonstrates that audiences have been immersed in the plot of movies and they are very interested in the climax of spots. Meanwhile, there are more interesting phenomena. As shown in Table 1, the third topic of *anime* shows several special words, such as "*Human, Vampire, Magician, Knight*". The second topic of music shows several anime characters, such as "*Railgun, Electro Master*", which are from the famous anime "A Certain Magical Index"[20]. We can observe that these two categories have several implicit connections. Since most audiences of these videos are fond of ACG and most music videos are about anime, it is natural that these two categories have several common topics.

Different from the above categories, videos in *tech*, *show* and *play* have several unique features. We can observe that the third topic in *play* reveals several instruments, the second topic in *tech* reveals several scientific terms, and the first and second topics in *show* demonstrate several names of celebrities. These topics can help audiences have a more detailed understanding about the videos at the first sight, which means, they will provide better opportunities for precise video recommendations.

Moreover, we can observe that the first and second topics in *show* display the name of male celebrities and female celebrities respectively. Recalling the topic model mechanism, we can conclude that audiences will comment on actors and actresses with different ways of expression, including diction and manner of speaking. After a deeper analysis, we obtain much more interesting things. First, the surrounding words around the names of actors or actresses are the shows they attended or their companies. Second, there are more laudatory words around the actors, such as "*handsome*" and "*cute*". On the contrary, there are more names of other actresses around the actresses. For example, "*Meiqi Meng*" and "*Renyu Liu (Reyi)*", which are both names of actresses, appear around actress "*Yihan Chen (Estelle)*". Moreover, when talking about the actors, the audiences (or fans) will praise and defend their idols. Thus, we can find plenty of comments like "*Those who slander Kai Wang can get out and turn off the videos*". However, when talking about the actress, there are much more comparison. For example, "*In terms of beauty, I think Renyu Liu should be the first, Yihan Chen (Estelle) is the second*" or "*I might like this group if Chaoyue Yang and Jing Fu were not included*". More details about user behaviours will be discussed in section 5.4.

20. https://en.wikipedia.org/wiki/A_Certain_Magical_Index

5.2 Special Expression

As mentioned in section 4.2, danmus contain plenty of special attributes compared with traditional comments. In this section, we start to study danmu's unique characteristics in a semantic level.

First, we focus on analyzing the special expression which is different from the traditional usage. During years of our research on *Danmu*, we have investigated hundreds of ACG[21] users who are familiar with danmu-styled content. People who often watch danmu-enabled videos tell that the expression of danmus are quite different from our daily language. Since it is impossible to make people to enumerate all the special phrases of danmus and new expressions that will continue to emerge as time goes by, we utilize a simple but effective method to automatically find out the difference between danmu text and ordinary text.

Our method is based on word embedding, in which words or phrases from the vocabulary are mapped to vectors of real numbers. We now give a formal description. Given an ordinary corpus (which may have multiple options, such as Chinese Wikipedia[22]) $\boldsymbol{T}$ and a danmu corpus $\boldsymbol{D}$ with each vocabulary $\{\boldsymbol{w}_i^T\}$ and $\{\boldsymbol{w}_j^D\}$, where $\boldsymbol{w}_i^T$ and $\boldsymbol{w}_j^D$ indicate the i^{th} and j^{th} word respectively, a word $\boldsymbol{w}$ is called *special expression word* if: 1) $\boldsymbol{w} \in \{\boldsymbol{w}_i^T\} \cap \{\boldsymbol{w}_j^D\}$, and 2) the meaning of $\boldsymbol{w}^T$ is quite different from $\boldsymbol{w}^D$, i.e., there is a large margin between their embeddings. We aim to find out words with special meaning through the following steps:

1) Representing all words in vocabulary $\{\boldsymbol{w}_i^T\}$ into word vectors $\{\boldsymbol{v}_i^T\}$ using corpus $\boldsymbol{T}$. Then, $\{\boldsymbol{v}_i^T\}$ is used as initialization for words in $\{\boldsymbol{w}_k\} = \{\boldsymbol{w}_i^T\} \cap \{\boldsymbol{w}_j^D\}$.
2) Based on the initialization, performing word embedding again on the *Danmu* dataset $\boldsymbol{D}$ to obtain vectors $\{\boldsymbol{v}_j^D\}$.
3) For every word in $\{\boldsymbol{w}_k\}$, calculating a score $s_k = \mathrm{dist}(\boldsymbol{v}_k^T, \boldsymbol{v}_k^D)$ and then ranking the words based on the score. We extract the top n words as the *special expression words*, which obviously have relatively larger semantic margin in ordinary language and danmus.

21. The ACG is an abbreviation of "Anime, Comic and Games", used in some subcultures of Asian countries.
22. https://en.wikipedia.org/wiki/Chinese_Wikipedia

TABLE 2
Special Expressions in several categories. For each special expression entry, the upper line is the Wiki neighbor words and the bottom line is the danmus neighbor words.

Category	Words	Neighbor	Category	Words	Neighbor
Movie	嘿嘿嘿 HeiHeiHei	请原谅，好开心，吓呆了 Please forgive, So happy, Stunned	Anime	嘿嘿嘿 HeiHeiHei	好开心，请原谅，听不清 So happy, Please forgive, Can't hear
		233333, gg, 打卡 233333, gg, Sign in			FFFFFFF, yooooooooo, hhhhhhhh FFFFFFF, yooooooooo, hhhhhhhh
	葫芦娃 Calabash Brothers	白龙马，沙悟净 White dragon horse, Friar Sand		打脸 Smack in the face	叫骂，出拳，失去机会 Shouting, Punch, Lose chance
		陈意涵，郭采洁，林青霞 Yihan Chen, Caijie Guo, Qingxia Lin			黑化，要命，捅 Deprave, Terribly, Stab
	狗粮 Dog food	熟肉，猫粮，喂猪 Cooked meat, Cat food, Feed pigs		小司 Xiao Si	小咖，碎碎念，马小虎 Xiao ka, Self-talking, Xiaohu Ma
		果汁，鲜肉，棉花糖 Juice, Fresh meat, Marshmallow			闺女，老奶奶，正太控 Daughter, Granny, Shota Complex
Music	想哭 Want to cry	催人泪下，唱出来 Tear-jerking, Singing out	Dance	护眼 Eye protection	手术灯，红外摄像机 Opearting-lamp, Infrared camera
		少女心，太天真，圆脸 Girlish heart, Naive, Moonface			盲降，niconiconi，截屏 ILS, niconiconi, Screen shot
	表白 Bare one's heart	鼓足勇气，喜极而泣 Take courage, Tears of joy		哇哦 wow	DATSUN, NAOKI, harmony, Gigantic DATSUN, NAOKI, harmony, Gigantic
		心疼，吃醋 Heartache, Be jealous			dd, cool, nb, yy, boom, gg, kk dd, cool,nb, yy,boom,gg,kk
	便当 Bento	蛋炒饭，叉烧饭 Egg fried rice, Char siew rice		讲道理 Reasoning	责斥，就事论事 Reproof, Fact-oriented
		狗粮，香香，主角 Dog food, Fragrant, Leading role			说真的，说实话，胡扯 Indeed, To be frank, Bullshit
Shows	zz	Metric, CIELAB, TLV, Dst Metric, CLELAB, TLV, Dst	Sports	EOS	尼康，APS-C，佳能 Nikon, APS-C, Canon
		NB, 233333333, 截图成功 Amazing, 233333333, Got screenshot			hhhhhhhh, 66666, 2333333 hhhhhhhh, 66666, 2333333
	撒花 Throwing flowers	放凉，麻织品，调匀 To cool, Linen, Stir evenly		截屏 Screenshot	媒体文件，剪贴板，选项卡 Media files, Clipboard, Tabs
		全剧终，菠萝包，眼睫毛 The end, Pinapple bun, Eyelash			全屏，FLAG，PPT Full screen, FLAG, PPT
	可怕 Dreadful	可悲，超乎想象 Miserable, Beyond imagination		丁宁 Ning Ding	张楠，徐晨，吴敏霞 Nan Zhang, Chen Xu, Minxia Wu
		可悲，要命，吓人 Miserable, Terribly, Scary			刘国梁，王皓，张怡宁 Guoliang Liu, Hao Wang, Yining Zhang

To be specific, the first two steps can be achieved by an auto-encoder based on structure as follows:

$$\begin{aligned} \boldsymbol{v}_t &= \boldsymbol{E}\boldsymbol{w}_t, \quad t \in 1, 2, ..., n, \\ \{\boldsymbol{h}_1, \boldsymbol{h}_2, ..\boldsymbol{h}_n\} &= \mathrm{RNN}(\{\boldsymbol{v}_1, \boldsymbol{v}_2, ..\boldsymbol{v}_n\}), \\ \mathrm{P}(\boldsymbol{w}'_t) &= \mathrm{softmax}(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}), \end{aligned} \tag{1}$$

where the inputs are one-hot representations of words in a length n sentence $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..\boldsymbol{w}_n\}$, and the matrix $\boldsymbol{E}$ is initialized by a pre-trained 300 dimensions word2vec model on Wiki dataset[23]. The notation $\mathrm{RNN}()$ means to re-train the word vectors through a Recurrent Neural Network (RNN) which have been widely adopted to deal with sequential data and embedding tasks. In our case, we choose Gated Recurrent Units (GRU) [32] as the base of the embedding tasks. It is formulated as [33]:

$$\begin{aligned} \boldsymbol{z} &= \sigma(\boldsymbol{U}_z\boldsymbol{v}_t + \boldsymbol{W}_z\boldsymbol{h}_{t-1}), \\ \boldsymbol{r} &= \sigma(\boldsymbol{U}_r\boldsymbol{v}_t + \boldsymbol{W}_r\boldsymbol{h}_{t-1}), \\ \boldsymbol{h'} &= \tanh(\boldsymbol{U}_h\boldsymbol{v}_t + \boldsymbol{W}_h(\boldsymbol{r} \odot \boldsymbol{h}_{t-1})), \\ \boldsymbol{h}_t &= (1 - \boldsymbol{z}) \odot \boldsymbol{h'} + \boldsymbol{z} \odot \boldsymbol{h}_{t-1}, \end{aligned} \tag{2}$$

where all $\boldsymbol{W}$ and $\boldsymbol{U}$ are the trainable parameters, $\boldsymbol{v}_t$ is the t^{th} input word vector, $\boldsymbol{z}$ and $\boldsymbol{r}$ represent the update and reset gate respectively, and $\boldsymbol{h}_t$ is the GRU cell's t^{th} state whose size is 500. The output word $\boldsymbol{w}'_t$ in Eq. (1) is sampled from the distribution $\mathrm{P}(\boldsymbol{w}'_t)$. Intuitively, the whole procedure is to feed a word sequence $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..\boldsymbol{w}_t\}$ to the RNN and output another one $\{\boldsymbol{w}'_1, \boldsymbol{w}'_2, ..\boldsymbol{w}'_t\}$. The word embedding is achieved by minimizing the sequence reconstruction loss which is formulated as the negative log-likelihood:

$$L = -\sum_{t=1}^{n} \log\ \mathrm{P}(\boldsymbol{w}'_t = \boldsymbol{w}_t | \boldsymbol{h}_{t-1}, \boldsymbol{w}_0, ..., \boldsymbol{w}_{t-1}). \tag{3}$$

In our experiments, we separately run the above algorithm on danmus for these 8 categories and sample several results of each one. Table 2 shows the sampled results.

From the table, we can obtain many interesting phenomena. We can observe that there are onomatopoeias in most categories. For example, "*HeiHeiHei*" in *Movie* and *Anime* categories and "*wow*" in *Dance* category. We can observe that the Wiki neighbors of former are often related to the "*happy*" emotion. Meanwhile, its danmu neighbors are some onomatopoeias, such as "*yoooooo*" or "*hhhhhh*", which express the happy emotion of audiences. Moreover, there are more words having very different meanings from their traditional ones. "*Dog Food*" is a very typical example. We can observe that its Wiki neighbors are often about the food, such as "*boiled meat (熟肉)*" and "*cat food (猫粮)*", which indicates

23. https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2

that its original meaning is something to feed dogs. On the contrary, we can observe that its danmus neightbors are quite different. For example, "*Handsome guy (鲜肉)*" means young and pretty actors. "*Marshmallow (棉花糖)*" means very sweet. All its danmu neighbors demonstrate its special meaning here, i.e, *public display of affection*. Meanwhile, we can observe the same phenomenon in *Sports*. "*EOS*" originally represent camera equipment, like its Wiki neighbors shows. However, when comes to danmus, it represents some happy and interesting emotions like "*hhhhhh*" and "*666666*". Furthermore, when talking about names, we can find out that danmus contain more domain knowledge and precise meanings than ordinary ones. For example, "*Ning Ding*", name of a famous ping-pang player in China. We can observe that its Wiki neighbors are some sports players, such as "*Minxia Wu*", name of a Chinese diver; "*Chen Xu*", name of a badminton athlete. All of these neighbors are athletes, but they belong to different sports activities. When comes to danmus neighbors, we can observe that these names all belong to ping-pang players, such as "*Guoliang Liu*", "*Hao Wang*" and "*Yining Zhang*". Based on these phenomena, we can conclude that danmus contain more domain knowledge and have more precise meanings when expressing specific meanings or specific objects.

5.3 Danmu Semantic Embedding

In previous section, we have analyzed some special expressions in *Danmu* dataset based on word embedding. In this section, we are going to perform some analyses on the whole danmu sentences. It is known that when processing text data, Bag-Of-Word or TF-IDF [34] are often used as features, and they are usually good at modeling long documents. While, as we discussed before, danmus can be regarded as a special type of short text. Moreover, since deep learning technologies are widely adopted in many NLP tasks, we tried to perform deep semantic embedding for danmus.

To be specific, we utilize character-level RNNs as auto-encoder which consists of an encoder ϕ:

$$\begin{aligned} \boldsymbol{u}_t &= \boldsymbol{E}\boldsymbol{c}_t, \quad t \in 1, 2, ..., n, \\ \boldsymbol{h}_t &= \phi(\boldsymbol{u}_t|\boldsymbol{h}_{t-1}), \quad \boldsymbol{h}_0 = \boldsymbol{0}, \end{aligned} \tag{4}$$

and a decoder ψ:

$$\begin{aligned} \boldsymbol{u}'_t &= \boldsymbol{E}\boldsymbol{c}'_t, \quad t \in 1, 2, ..., n, \\ \boldsymbol{h}'_t &= \psi(\boldsymbol{u}'_{t-1}|\boldsymbol{h}'_{t-1}), \quad \boldsymbol{u}'_0 = \boldsymbol{0}, \boldsymbol{h}'_0 = \boldsymbol{h}_n, \\ \mathrm{P}(\boldsymbol{c}'_t) &= \mathrm{softmax}(\boldsymbol{W}\boldsymbol{h}'_t + \boldsymbol{b}). \end{aligned} \tag{5}$$

where $\boldsymbol{c}_t$ and $\boldsymbol{c}'_t$ are one-hot vectors of the input characters, $\boldsymbol{u}_t$ and $\boldsymbol{u}'_t$ are the corresponding character representations, $\boldsymbol{E}$ is a randomly initialized embedding matrix. Similar as Eq. 2, $\phi()$ and $\psi()$ are chosen as the GRU cell, in our case. The auto-encoder is learned in a pair-wise manner. At each time, a pair of danmu sentences $\{\boldsymbol{c}^a_1, \boldsymbol{c}^a_2, ..., \boldsymbol{c}^a_n\}$ and $\{\boldsymbol{c}^b_1, \boldsymbol{c}^b_2, ..., \boldsymbol{c}^b_m\}$ are sampled from the same frame and then put into the encoder ϕ by characters separately to get the corresponding embedding vectors $\boldsymbol{h}^a$ and $\boldsymbol{h}^b$ through Eq. (4). In the decoder, $\boldsymbol{h}^a$ and $\boldsymbol{h}^b$ are inputs to the decoder ψ and the reconstructed sentences $\{\boldsymbol{c}'^a_1, \boldsymbol{c}'^a_2, ..., \boldsymbol{c}'^a_n\}$ and $\{\boldsymbol{c}'^b_1, \boldsymbol{c}'^b_2, ..., \boldsymbol{c}'^b_m\}$ are returned. We optimize the parameters based on reconstruction loss, which is the sum of the negative log-likelihood of correct characters at each step:

$$L_{rec} = -\sum_{t=1}^{n} \log\ \mathrm{P}(\boldsymbol{c}'_t = \boldsymbol{c}_t|\boldsymbol{c}'_0, ..., \boldsymbol{c}'_{t-1}). \tag{6}$$

More importantly, to involve more danmu semantic information, we also add a semantic loss formulated as:

$$L_{sem} = \mathrm{dist}(\boldsymbol{h}^a, \boldsymbol{h}^b), \tag{7}$$

in which we take "temporal correlation" assumption [4] , i.e., comments appear in the same frame hold the similar topics (relevant to the frame, but can be still regarded as diverse) compared with those not in the same frame. The distance function $\mathrm{dist}()$ here is chosen as cosine distance. Finally, the overall loss function for the embedding is given by:

$$L_{rec} = L^a_{rec} + L^b_{rec} + L_{sem}. \tag{8}$$

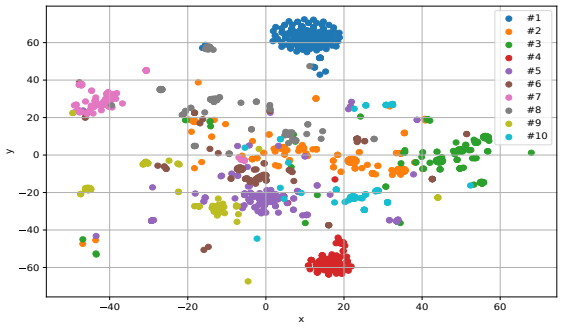


Fig. 7. The t-SNE of top 10 biggest clusters of danmu representations in one episode of videos in *anime* category.

After getting the representations of danmu semantic, we tend to perform the cluster analysis. However, there are too many danmus to be visualized. In order to demonstrate the performance of danmu semantic embedding better, we select one episode in *anime* category as an example. We generate the semantic representations of all danmus in this video and do t-SNE analysis. Fig. 7 demonstrates the t-SNE results of the top 10 biggest clusters. We can obtain several interesting phenomena. In the following parts, we will give a detailed analysis about several important clusters, i.e., blue, red, pink and purple clusters.

The first cluster we analyze is the blue cluster, which is far away from the other clusters. After deeper analysis, we figure out that danmus in this cluster are all about the end of the anime videos. Specifically, this cluster contains plenty of danmus, such as "*The End*", "*The End, Throwing Flowers*", as well as "*Seal Released*". Like old movies, "*The End*" means the anime comes to an end and the audiences are reluctant to part with it. "*The End, Throwing Flowers*" is a humorous expression and usually appear at the last epoch of one anime because audiences would like to express their remembrance or spoof to this anime[24].

The second cluster that attracts our attention is the red one. This cluster contains plenty of funny comments, such as "*233333*", "*hahaha*", as well as "*show up with BGM*". Moreover, these danmus appear at the funny part of the videos

24. https://zh.moegirl.org/%E5%AE%8C%E7%BB%93%E6%92%92%E8%8A%B1

TABLE 3
User behaviours and corresponding examples.

Behaviour	Examples	Behaviour	Examples
Sign in and Contract	感谢承包商 Thank the contractor.	Throwing Flowers	完结撒花 Finishing and Throwing Flowers
	2016.8.4 其他6个小伙伴 2016.8.4 Another 6 Buddies		完结，好棒 It's the end. Very Good
	照例，感谢承包商 By convention, thank the contractor.		完结撒花** Finishing and Throwing flowers**
Danmu Eye Protection	老天，这背景音乐... OMG! This BGM...	High Energy	前方高能 High energy ahead
	对不起，无意冒犯 Sorry, mean no offence		前方弹幕高能！非战斗人员快速撤离 High energy ahead! Non-combatants please retreat quickly!
	先暂停，盖好被子 Have a break, let me cover the quilt first		老天老天老天 OMG OMG OMG OMG
Counter	啪x20 Pa*20	Stars	给敖子逸打call！！！ Put my glow stick up for Ziyi!!!!!!
	计数君，你够了 That's enough Mr. counter		源哥的侧脸真的无敌了！ Yuan's side face is really invincible!
	计数君来啦 Mr. counter is coming		蔡徐坤表情真的笑死我了哈哈哈哈哈哈 I was really amused by Cai Xukun's facial expression, hahahahahaha

in almost every category, especially these funny music or songs. This phenomenon demonstrates that most audiences watch videos for relaxation and fun. The interesting part is more attractive to them.

The following pink cluster is about "*hougong (Harem)*", which describes the situation in which the hero attracts attentions of many female characters, and has interaction and close relationships with them in ACG culture[25]. For example, *Sword Art Online*[26] indicates the hero Kirito's female admirers, such as Asuna, Leafa, and Sinon. Even the artificial intelligence Yui is fond of him. Therefore, we can observe plenty of similar danmus, such as "*Rhythm of hougong*" and "*hougong +1*".

The last but not least cluster is the purple one. The danmus in this cluster often have specific meanings. When characters in videos do something particularly bold and dangerous at the same time, this kind of danmus will appear on the screen. To be specific, this type of danmus are treated as a sign, or so-called "*flag*", which means something special will happen[27]. Thus, we can observe the danmus, such as "*I respect you as a man*" or "*The flag has already been set up*".

Meanwhile, we analyze these clusters in a more detailed perspective. As described above, the blue, red, and pink clusters are not only far away from others, but also have closer ties inside than others. Moreover, we have observed that most of danmus in each of these clusters are either the same or expressing the same meaning in a similar way. For example, most of the danmus in blue cluster celebrate the end of a anime or movie. The audiences always prefer to utilize "*Throwing Flowers*", "*It is the end, so happy*", or "*Throwing the flowers for the end*" to express their feelings. This kind of phenomenon draws our attention. What are the relations between danmus and user behaviours? Whether danmus can reveal the pattern of user behaviours and be helpful for us to understand user behaviours. Thus, in the next part, we intend to investigate their relations and try to explain them.

25. https://en.wikipedia.org/wiki/Harem_(genre)
26. https://en.wikipedia.org/wiki/Sword_Art_Online
27. https://zh.moegirl.org/Flag

5.4 Danmu and User Behaviours

As mentioned in section 5.3, we find out that there may be relations between danmus and user behaviours. In order to investigate the relations better, we further discuss the "*semantic concentration*" through the video time-line in this section. The semantic concentration score of sentence group i is defined as:

$$score_i = \frac{1}{\mathrm{C}_2^{|\boldsymbol{G}_i|}} \sum_{\boldsymbol{s}^a, \boldsymbol{s}^b \in \boldsymbol{G}_i} \mathrm{dist}(\boldsymbol{h}^a, \boldsymbol{h}^b), \tag{9}$$

where $\boldsymbol{s}^a$, $\boldsymbol{s}^b$ are two different sentences from $\boldsymbol{G}_i$ and $\boldsymbol{h}^a$, $\boldsymbol{h}^b$ are the corresponding embedding vectors obtained in Section 5.3. The notion $\mathrm{C}_2^{|\boldsymbol{G}_i|}$ here indicates the number of all possible pair combinations in sentence group $\boldsymbol{G}_i$ and $\mathrm{dist}()$ here is also chosen as the cosine similarity.

Based on this formulation, we sort danmu groups by their scores and extract several typical user behaviors for case study as shown in Table 3. At the same time, in order to investigate those behaviors more deeply, we also perform the cluster analysis for danmu embeddings along the time-line of videos. Since the cluster method we use is the same as 3.2.3, we do not set the number of clusters. Based on the above two steps, we observe a very interesting phenomenon. All the danmus along the time-line are naturally clustered into three categories: 1) The beginning; 2) The end; 3) The middle. This phenomenon demonstrates that there must be some connections between danmus and user behaviours. Therefore, in order to better analyze these user behaviours, we try to group them into three categories and analyze them with corresponding video clips.

5.4.1 The Beginning.

The "*Sign in and Contract*" behavior always appears in the beginning of the videos. As shown in Fig. 8, when recommending a video that has been watched before, audiences would love to leave some notes, such as "*2016.8.4 Another 6 little partner*", to express that they have watched this video, no matter they finished watching the video or not. Thus, there are many danmus in the beginning of the videos, which have been observed in subsection 4.1. Moreover,

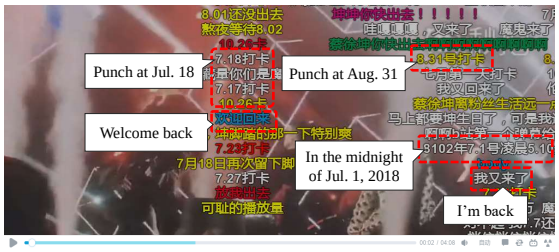


Fig. 8. Viewers like to send date or "Punch" to perform *sign in* at the beginning of the videos. The example was chosen from a concert.

some videos need to be paid for their copyright. The facilitators provide the service so that part of the audiences vote to pay for the videos and they can watch the videos earlier than others. This service is called "contract". Thanks to these contractors, the facilitators are able to predict the popularity of videos that audiences intend to buy. Meanwhile, other audiences can benefit from these contractors. Therefore, we can observe the comments, such as "*Thank the contractor*" in the beginning of the videos.

5.4.2 The End.

When a series of anime or TV Series finished, audiences always send comments such as "*Throwing Flowers*" to celebrate the end of the videos, which can be shown in Fig. 9. We have analyzed this phenomenon in sections 5.2 and 5.3. Moreover, they like to express their feelings with different emojis, such as the flower and firework emojis for "*Throwing Flowers*". Thus, we can observe more emojis in this kind of videos, as mentioned in subsection 4.2.3.

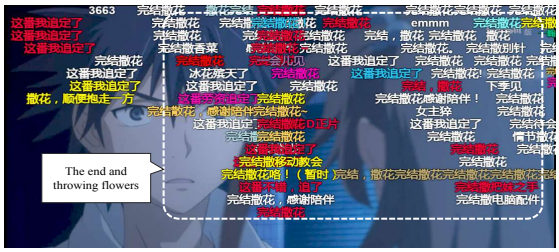


Fig. 9. "Throwing flowers" at the end of the last anime episode.

5.4.3 The Middle.

Despite the obvious behaviors at the beginning and the end, there are more special behaviors in the middle of the videos. The behavior "*Danmu Eye Protection*" always occurs before the horrible part of a movie, especially in horrible movies. They utilize plenty of danmus to cover the scenes of videos, so that latter audiences will not be frightened. Meanwhile, this behavior also indicates the climax of videos. The behavior "*High Energy*" is similar to "*Danmu Eye Protection*" behavior. The difference is that "*High Energy*" always occurs in the climax of the videos, as shown in Fig. 10. The audiences who have watched the video would like to notify the location of climax to latter audiences. Moreover, videos that have more climaxes parts will attract more audiences. The number of danmus is a very suitable signal that reflects the popularity of videos.

The behavior "*Counter*" is also an interesting behavior. When watching a video, part of the audiences would like

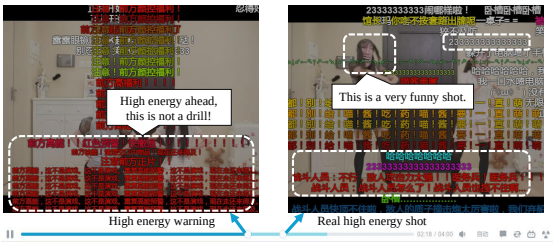


Fig. 10. "High energy warning" always occurs before the really interesting shots. The example was chosen from a dance video.

to count some specific actions in the videos, such as kiss and famous cars. They would continue counting the number until the end of the videos. Sometimes, when a video has very big number of some specific objects, it will attract more audiences. As for the last behavior "*Stars*", it is similar to "*Counter*" behavior except that its target is a specific celebrity. Fans of these stars would express their love to the idols. In order to demonstrate that their idols are very famous, they always post lots of same comments when their idols appear in the videos.

In conclusion, these behaviors give us better access to the analysis of user behaviors when watching videos and the improvement of their watching experiences, which traditional comments do not contain. Despite of these unique characteristics, danmus also attract the researchers' interests. They have made preliminary attempts from different perspectives. In the following section, we will give a brief introduction to these works and reveal some potential directions of *Danmu* dataset.

6 Applications and Future Directions

In order to better demonstrate that *Danmu* dataset has a wide range of applications in the fields of text, images, multimedia, etc., and has potential research value, we will give brief introductions to early attempts on 1) User Behavior Modeling; 2) Video Semantic Understanding; 3) Generative Applications; 4) Application in Other Domains in the following parts. These potential directions are shown in Fig. 11. For each direction, we select one representative work to reveal the role of *Danmu* dataset.

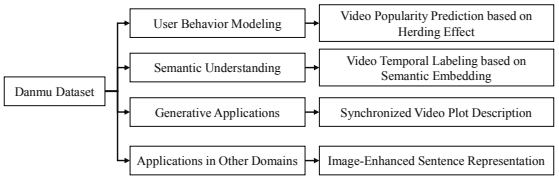


Fig. 11. Four potential directions and the typical applications.

6.1 Video Popularity Prediction

In the modern society, it is possible to record user behaviour digitally for better service providing, e.g., user profiling [8], user pattern learning [35], and recommendation [36], [37]. As mentioned before, danmus are closely related to the behaviours of audiences. Thus, *Danmu* dataset can be helpful to model behaviours of audiences and predict videos'

popularity. Next, we will introduce one work to show how danmus are used to model user behavior.

Considering the video popularity prediction problem, it is natural that video popularity is directly related to user behavior. However, fine-grained characterization of this relation is still challenging. The early attempt is to use danmu information [3]. As mentioned in section 5.4, danmus and user behaviors are closely related. Researchers try to tackle this problem by leveraging a phenomenon called "herding effect", i.e., an audience could directly see other audiences' interaction (e.g., views and comments) with videos, which makes whether the audience watches a particular video is easily affected by others' previous interactions with this video.

To involve the herding effect into prediction model, they define a measurement to quantify the herding effect, which can be formulated as follows:

$$h_{i,t} = \prod_k (1 + h_{i,t}^k)^{-\theta_{k,t}}, \quad h_{i,t}^k = \frac{|\bar{n}_t^k - n_{i,t}^k|}{\bar{n}_t^k}, \tag{10}$$

where the aspect k can be one of $\{$*view*, *danmu*, *upload*$\}$, which means, the herding effect is considered from 1) the number of views, 2) the number of danmus, and 3) the uploaded date. $n_{i,t}^k$, $\bar{n}_t^k$ and $\theta_{k,t}$ here respectively stands for the i-th video's quantity, the popular videos' average quantity and the parameter of aspect k. Once they have the measurement of herding effect, a prediction model is then utilized to combine $h_{i,t}$ with other factors that may influence video popularity. Then, the parameters are learned automatically through an efficient estimation method. The experimental results on a subset of *Danmu* dataset show that the proposed model improves the prediction accuracy by 47.19% compared to the baselines.

The above work proved that it is feasible to involve danmus information into user behavior modeling. While, due to the complexity and diversity of human behavior, further studies are needed. First, as discussed in Section 4, danmus' quantity distribution over playback time, font colors, display type or even number of emojis often reflect a video's features, an audience's emotion and interest. How to take advantage of these multiple attributes to describe one's behavior is worth exploring. Second, some of the behaviors in danmus also reflect characteristics of human psychology, which means the application of this dataset is definitely not limited to the field of computer science, but can also be beneficial to academic research in other disciplines, such as sociology, psychology or human culture. Thus, how to understand users from the perspective of psychology based on danmus is another problem worth paying attention to.

6.2 Video Temporal Labeling

Since *Danmu* dataset contains plenty of danmu texts and corresponding images, semantic understanding about videos is another important application scenarios. Semantic understanding is known as a fundamental and yet challenging task [4]. It requires the agent to model and represent the semantic of texts or images as comprehensive as possible, and it can be applied in many areas, e.g., Visual Question Answering [38], Visual Dialog [39], as well as dialog system [40]. In the following part, we will introduce the work that focuses on temporal labeling videos based on danmus.

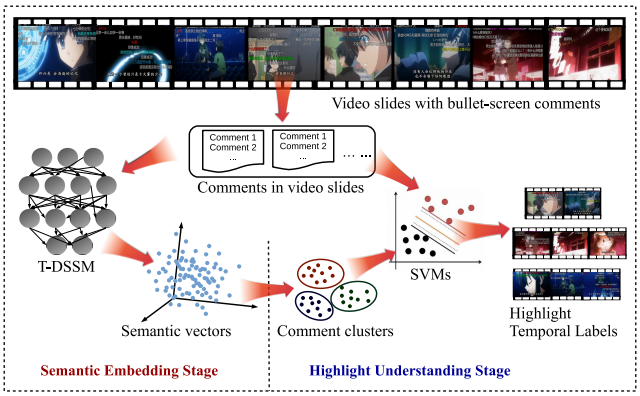


Fig. 12. The architecture of labeling framework.

The motivation of "temporal labeling" lies in the limitations of conventional video recommendation/retrieval systems. Before danmus emerge, those systems mainly rely on video titles or human annotations which are usually insufficient for fine-grained requirements. Though a large amount of efforts have been made on automatic labeling to enrich the meta-data, most of them focus on the whole video other than labeling precisely with timestamp on video shots. Fortunately, danmus give us the opportunity to "go inside" the video to understand the content across the playback time. In our work [4], we try to model video shots based on the semantics of danmus. We design a danmu embedding framework to assign temporal labels on highlighted video shots, as shown in Fig. 12. Considering the informal expression and latent meanings of danmus (as shown in Section 5.2), danmu text is not involved directly in using the traditional NLP methods (e.g., n-gram or Bag-of-Words features). Instead, the labeling procedure is divided into two stages: 1) *Semantic Embedding Stage*: generating semantic representations for danmus via a deep neural network called **T**emporal **D**eep **S**tructured **S**emantic **M**odel (T-DSSM); 2) *Highlight Understanding Stage*: constructing features for video shots based on danmu representations, and then recognizing and labeling video highlights in a supervised way. Experiments on a real-world dataset prove that this framework could effectively label video highlights with a significant margin compared with baselines.

Although it is just a simple attempt, we are among the first researchers who try to understand multimedia content through crowdsourcing data and reveal that it is attractive to conduct follow-up studies along this line. Based on the *Danmu* dataset, there are still challenges for further studies. First, danmus are rich of user's emotional information (e.g., color and emojis, mentioned in Section 4.2 and 4.2.3), while there are few researches around emotion analyses of danmu-enabled videos. Therefore, understanding multimedia semantics from the perspective of user emotion will be an important research direction. Second, the semantic relations we used in the above work is relatively rough. However, we know that danmus is not a simple (objective) description of the video content. Usually their topic is "gossiping" (joking or subjective review) for the content in a particular area of the video frame. Thus, we not only need to know which video segments are related to a specific danmu, but also to find out which visual elements in the frames are relevant. In

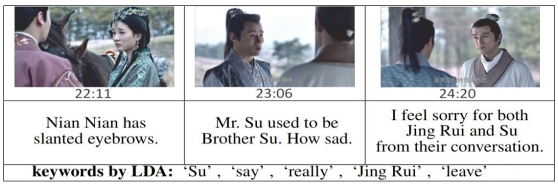


Fig. 13. The example of selected comments and corresponding frames.

addition, this gossiping behavior is often closely related to video shot and the context of danmus sent by other viewers. The contextual characteristics make the challenge even more critical. Last but not least, due to the rapid growth of danmus' amount, the expression of danmu language is being updated constantly. How to capture new semantics and semantic migration dynamically in an online learning way is an inevitable problem.

6.3 Synchronized Video Plot Description

Danmu dataset contains huge amount of text data and rich multimedia information. Therefore, we are also very interested in generative applications [41]. As an important component of machine learning, generative applications focus on generating human readable contents and have may downstream tasks, e.g., Image Captioning [42], [43], Visual Question Answering [38], [44], and Video Description [45], [46]. Next, we will introduce the work that utilizes danmus to generate video description in a synchronized way, in which the descriptions are easy to be read and understood.

Traditional video descriptions are usually generated by humans. However, with the rapid growth of online sharing videos, effective management and annotation of videos become into urgent and indispensable challenges. Automatic video tagging draws more and more attention which mainly focuses on detecting concepts and associating tags of the entire videos. Since danmus are time-sync video comments that are closely related with the videos, researchers try to generate temporal tags or labels based on danmus. Linli et al. [46] proposed a novel temporal summarizing model with the consideration of danmus, in which the representative and interesting danmus of a video are selected and highlighted along the timeline. Moreover, a novel temporal summarizing model based on the data reconstruction principle is proposed to resolve the issues of informal, noisy and redundant information contained in danmus. Based on this framework, the video can be described in a synchronized way, conveying relevant, important and non-redundant information of the video, which is easy to read and comprehend. Experimental results on real-world data demonstrate the effectiveness of this framework and prove that danmus could be the bridge to understand and describe videos. Fig. 13 demonstrates the performance of the proposed method. We can observe that the selected comments are very informative and provide consistent description to the video content in a time-sync manner compared with the topic words generated by LDA model.

The work introduced above can be regarded as a typical study of synchronously describing video plot by danmu text. However, it is only a simple beginning of this type of work. First, although the real-time description for video is accomplished, it only achieves the selection of the existing danmus. For the cold-start problem and some danmu-less videos which may exist in practical applications, further explorations still needs to be carried out. Second, plots of the same video shot may be diversified, and the same is true for users' interest. Thus, it is worth studying to obtain personalized video plot descriptions from different users. Third, solution in that work is not really generating sentences, while automatic generation for danmu-styled content can be interesting in many scenes such as live streaming, Kara OK, video game real-time review, etc. However, through our pilot studies, we found existing encoder-decoder based methods (which are widely used in Image Caption) are quite difficult to generate danmu-styled content given multimedia data due to the complex semantic relations. In other words, danmu related generative task is another valuable direction supported by this dataset.

6.4 Image-Enhanced Sentence Representation

In addition to works directly related to videos, *Danmu* dataset may have broad application in other domains like Natural Language Processing (NLP) [47] or Computer Vision (CV) [48]. Next, we will introduce the work that utilizes danmus and corresponding frame images to do Natural Language Inference (NLI).

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), requires an agent to determine the semantic relationship between premise sentence (p) and hypothesis sentence (h) among *entailment* (if the semantics of hypothesis can be concluded from the premise), *contradiction* (if the semantics of hypothesis cannot be concluded from the premise) and *neutral* (neither entailment nor contradiction) [29]. In order to do semantic inference between the given sentence pair, the vital thing is to model the sentence semantics precisely and comprehensively. Sentence semantics depends highly on its context. Even the same sentence can have different meanings with the consideration of different contexts. As is mentioned in senction 3.1, each danmu can be precisely matched with a frame in the video, which means we can access the corresponding context of each danmu. By utilizing this feature of danmus, we are capable of evaluating the sentence semantics more precisely. To this end, Kun.et.al [29] adopted images to enhance the sentence semantic understanding and proposed an Image-Enhanced Multi-Level Sentence Representation Net (*IEMLRN*). Since the video frame and corresponding comments can be matched exactly in *Danmu* dataset, the video frame can be treated as the context of corresponding comments and be helpful for sentence semantic understanding. Therefore, the authors designed a multi-level architecture to understand sentences from different granularity, which is in favor of the integration of images and text. They extracted part of *Danmu* data as NLI alike data and did extensive experiments on the data. The results demonstrate the effectiveness of *IEMLRN* and also prove that *Danmu* dataset is very useful in the field of sentence semantic understanding, representations and matching.

The image-enhanced NLI is a simple attempt to apply danmus to other domains. While we hope the application of danmus is not limited to pure NLP tasks. First, since danmu is in fact a communication method among users,

we naturally consider whether the data can be also applied to human-computer interaction [49], [50], chat robots [51], [52] and VQA [53], [54] systems. Second, as danmus can reflect user behavior to a large extent, can it be used for abnormal user (information) detection [55]? Moreover, studies around informal/special expression [56] may also be helpful in detecting sensitive information. At last, some of the behaviors in danmus also reflect characteristics of the audiences, which means that we can mine different preferences of audiences for videos. Thus, we can better utilize the danmus to model the users' interests and make some improvements in plenty of research areas, such as recommendation system [57], [58], user profile [59], [60] and personalized service [61].

7 Conclusion

In this paper, we made a deep understanding to users and videos based on a *Danmu* dataset (we have made public available) which was 1.7 TB in size and collected from a real-world danmu-enabled video platform across 8 various video categories. Our analysis on both basic and semantic levels showed how danmus establish relationships among language, multimedia and user behaviors. In order to further demonstrate the potential value of *Danmu* dataset, we also introduced our four applications, in which *Video Popularity Prediction* models the connection between danmus and user behaviors, *Video Temporal Labeling* and *Synchronized Video Plot Description* show how danmus help to model videos, and *Image-Enhanced Sentence Representation*, a NLP task, benefits from *Danmu* data. At last, we proposed the possible future directions for each application. We will keep improving this dataset and hope it would inspire new ideas in various research domains.

References

[1] B. Li, Z. Wang, J. Liu, and W. Zhu, "Two decades of internet video streaming: A retrospective view," *TOMM*, vol. 9, no. 1s, p. 33, 2013.

[2] V. N. Index, "White paper: Cisco vni forecast and methodology, 2015-2020," 2016.

[3] M. He, Y. Ge, L. Wu, E. Chen, and C. Tan, "Predicting the popularity of danmu-enabled videos: A multi-factor view," in *DASFAA*. Springer, 2016, pp. 351–366.

[4] G. Lv, T. Xu, E. Chen, Q. Liu, and Y. Zheng, "Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding." in *AAAI*, 2016, pp. 3000–3006.

[5] X. Lin, E. Ito, and S. Hirokawa, "Chinese tag analysis for foreign movie contents," in *ICIS*. IEEE, 2014, pp. 163–166.

[6] Z. Wu and E. Ito, "Correlation analysis between user's emotional comments and popularity measures," in *IIAIAAI*. IEEE, 2014.

[7] B. Wu, E. Zhong, B. Tan, A. Horner, and Q. Yang, "Crowdsourced time-sync video tagging using temporal and personalized topic modeling," in *SIGKDD*. ACM, 2014, pp. 721–730.

[8] E. Chen, G. Zeng, P. Luo, H. Zhu, J. Tian, and H. Xiong, "Discerning individual interests and shared interests for social user profiling," *WWW*, vol. 20, pp. 417–435, 2016.

[9] X. Li, G. Xu, E. Chen, and L. Li, "Learning user preferences across multiple aspects for merchant recommendation," *ICDM*, 2015.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[11] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[13] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[14] C. L. Zitnick, R. Vedantam, and D. Parikh, "Adopting abstract images for semantic scene understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, 2016.

[15] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *CVPR*, 2018.

[16] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, 2017.

[17] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *ICCV*, 2017.

[18] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: Understanding Stories in Movies through Question-Answering," in *CVPR*, 2016.

[19] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *ICCV*, 2017.

[20] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*, 2009.

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *ICCV*, 2011.

[22] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[23] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[24] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[25] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv:1705.06950*, 2017.

[26] M. Cheung, J. She, and N. Wang, "Characterizing user connections in social media through user-shared images," *IEEE Trans. on Big Data*, vol. 4, no. 4, pp. 447–458, 2018.

[27] K. Kuang, M. Jiang, P. Cui, H. Luo, and S. Yang, "Effective promotional strategies selection in social media: A data-driven approach," *IEEE Trans. on Big Data*, vol. 4, no. 4, pp. 487–501, 2018.

[28] B. Lwowski, P. Rad, and K.-K. R. Choo, "Geospatial event detection by grouping emotion contagion in social media," *IEEE Trans. on Big Data*, 2018.

[29] K. Zhang, G. Lv, L. Wu, E. Chen, Q. Liu, and H. Wu, "Image-enhanced multi-level sentence representation net for natural language inference," in *ICDM*, 2018.

[30] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE TCSVT*, vol. 11, no. 6, pp. 703–715, 2001.

[31] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[33] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[34] G. G. Chowdhury, *Introduction to modern information retrieval*. Facet publishing, 2010.

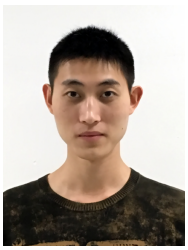
[35] Z. Zhang, R. Sun, X. Wang, and C. Zhao, "A situational analytic method for user behavior pattern in multimedia social networks," *IEEE Trans. on Big Data*, no. 1, pp. 1–1, 2017.

[36] Z. Li, H. Zhao, Q. F. Liu, Z. Huang, T. Mei, and E. Chen, "Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors," in *KDD*, 2018.

[37] L. Wu, Q. F. Liu, E. Chen, N. J. Yuan, G. Guo, and X. Xie, "Relevance meets coverage: A unified framework to generate diversified recommendations," *TIST*, vol. 7, pp. 39:1–39:30, 2016.

[38] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo, "Tell-and-answer: Towards explainable visual question answering using attributes and captions," *arXiv preprint arXiv:1801.09041*, 2018.

[39] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *CVPR*, vol. 2, 2017.
[40] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models." in *AAAI*, vol. 16, 2016.
[41] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu, "Analyzing the training processes of deep generative models," *IEEE Trans. on visualization and computer graphics*, vol. 24, no. 1, pp. 77–87, 2018.
[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *CVPR*, pp. 3156–3164, 2015.
[43] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," *CVPR*, 2016.
[44] Q. Li, Q. Tao, S. Joty, J. Cai, and J. Luo, "Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions," *arXiv preprint arXiv:1803.07464*, 2018.
[45] Y. Tu, X. Zhang, B. Liu, and C. Yan, "Video description with spatial-temporal attention," in *MM*. ACM, 2017, pp. 1014–1022.
[46] L. Xu and C. Zhang, "Bridging video content and comments: Synchronized video description with temporal summarization of crowdsourced time-sync comments." in *AAAI*, 2017.
[47] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," *CoRR*, vol. abs/1709.04696, 2017.
[48] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *TCSVT*, 2018.
[49] A. Dix, "Human-computer interaction," in *Encyclopedia of database systems*. Springer, 2009, pp. 1327–1331.
[50] G. Sinha, R. Shahi, and M. Shankar, "Human computer interaction," in *ICETET*. IEEE, 2010, pp. 1–4.
[51] X. Sun and J. Li, "Emotional conversation generation orientated syntactically constrained bidirectional-asynchronous framework," *arXiv preprint arXiv:1806.07000*, 2018.
[52] J. Huang, M. Zhou, and D. Yang, "Extracting chatbot knowledge from online discussion forums." in *IJCAI*, vol. 7, 2007, pp. 423–428.
[53] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Fvqa: Fact-based visual question answering," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
[54] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang, "R-vqa: Learning visual relation facts with semantic attention for visual question answering," *arXiv preprint arXiv:1805.09701*, 2018.
[55] D. W. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE signal processing magazine*, vol. 19, no. 1, 2002.
[56] A. Bessi, "On the statistical properties of viral misinformation in online social media," *Physica A: Statistical Mechanics and its Applications*, vol. 469, pp. 459–470, 2017.
[57] C. Guan, Y. Fu, X. Lu, E. Chen, X. Li, and H. Xiong, "Efficient karaoke song recommendation via multiple kernel learning approximation," *Neurocomputing*, vol. 254, pp. 22–32, 2017.
[58] L. Wu, Q. Liu, E. Chen, N. J. Yuan, G. Guo, and X. Xie, "Relevance meets coverage: A unified framework to generate diversified recommendations," *TIST*, vol. 7, no. 3, p. 39, 2016.
[59] L. Wu, Y. Ge, Q. Liu, E. Chen, R. Hong, J. Du, and M. Wang, "Modeling the evolution of users' preferences and social links in social networking services," *TKDE*, 2017.
[60] E. Chen, G. Zeng, P. Luo, H. Zhu, J. Tian, and H. Xiong, "Discerning individual interests and shared interests for social user profiling," *WWW*, vol. 20, no. 2, pp. 417–435, 2017.
[61] Z. Wu, G. Li, Q. Liu, G. Xu, and E. Chen, "Covering the sensitive subjects to protect personal privacy in personalized recommendation," *IEEE Transactions on Services Computing*, vol. 11, no. 3, 2018.



Guangyi Lv received the B.E. degree in Computer Science and Technology in 2013 from Sichuan University, Chengdu, P. R. China. He is currently a PhD student in the School of Computer Science and Technology at University of Science and Technology of China (USTC), P. R. China. His major research interests include deep learning, natural language processing and recommendation system. He has published several papers in refereed conference proceddings, such as AAAI, ICDM, PAKDD.

Kun Zhang received the B.E. degree in computer science and technology from University of Science and Technology of China, Hefei, China, in 2014. He is currently pursuing the PhD degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include natural language processing, and text mining. He has published several papers in refereed conference proceedings such AAAI, KDD, ICDM.

Le Wu received the PhD degree in computer science from the University of Science and Technology of China (USTC). She is currently a faculty member with the Hefei University of Technology, China. Her general area of research is data mining, recommender system, and social network analysis. She has published several papers in referred journals and conferences, such as the IEEE Transactions on Knowledge and Data Engineering, the ACM Transactions on Intelligent Systems and Technology, AAAI, IJCAI, KDD, SDM, and ICDM. She is the recipient of the Best of SDM 2015 Award.

Enhong Chen (SM'07) received the PhD degree from USTC. He is a professor and vice dean of the School of Computer Science, USTC. His general area of research includes data mining and machine learning, social network analysis, and recommender systems. He has published more than 100 papers in refereed conferences and journals, including the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Mobile Computing, KDD, ICDM, NIPS, and CIKM. He was on program committees of numerous conferences including KDD, ICDM, and SDM. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is a senior member of the IEEE.

Tong Xu currently working as a Associate Researcher of the Anhui Province Key Labora- tory of Big Data Analysis and Application, USTC. He has authored 20+ journal and conference papers in the fields of data mining, including KDD, AAAI, ICDM, SDM, etc.

Qi Liu received the PhD degree in computer science from USTC. He is an associate professor with USTC. His general area of research is data mining and knowledge discovery. He has published prolifically in refereed journals and conference proceedings, e.g., the IEEE Transactions on Knowledge and Data Engineering, the ACM Transactions on Information Systems, the ACM Transactions on Knowledge Discovery from Data, the ACM Transactions on Intelligent Systems and Technology, KDD, IJCAI, AAAI, ICDM, SDM, and CIKM. He is a member of the ACM and the IEEE. He received the ICDM 2011 Best Research Paper Award and the Best of SDM 2015 Award.

Weidong He received the B.E. degree in computer science and technology from University of Science and Technology of China, Hefei, China, in 2016. He is currently a PhD student in the School of Computer Science and Technology at University of Science and Technology of China (USTC), P. R. China. His research interests include deep learning and natural language processing.