# Judging a Book By Its Cover: Multimodal Distillation for Web Serial Rating Prediction

**James Gaiser**
NJIT
Newark, NJ
jcg64@njit.edu

**Nagarajan Venkata**
NJIT
Newark, NJ
nv395@njit.edu

**Ryan Li**
NJIT
Newark, NJ
rrl29@njit.edu

## Abstract

This paper examines web serials, novels serially published in installments over time through the digital medium. This subdomain of literature tends to be underrepresented in academic research, making web serials a prime target for further study. This paper uses knowledge distillation and natural language processing techniques to build a regression model to predict user-given rating scores for web serials.

## I. Introduction

Traditionally published literature tends to be sold as a physical book, an electronic format (e-books), or even in an audio format (audiobooks). However, there is a growing new domain of literature known as web fiction (web serials) that are almost exclusively published digitally online.

These stories tend to be released in installments or chapters being published serially, hence the name. They tend to be written by new or aspiring authors with a comparatively lesser quality control than traditional literature. Upon gathering more success, these authors may choose to self-publish their works online as e-books, or to work with a publishing company [1].

Web serials have garnered increasing popularity over the years. Sites like Royal Road, Scribble Hub, and Wattpad [2] gather millions of readers across the globe. In China, specifically, almost 500 million citizens are reported to read web fiction [3].

Despite this, there is a noticeable lack of research across this domain. Considering that readers engage with web serials differently than traditionally published media, there could be distinct characteristics unique to the medium.

To address this, this paper will train a multimodal model handling images, text, and tabular features on a regression task to predict a web serial's user-given rating. This complex model will be used as the teacher in a knowledge distillation task to examine the performance of a unimodal student, trained exclusively with a web serial's book cover image.

## II. Related Work

To the best of the authors' knowledge, no prior peer-reviewed studies have examined this exact combination of task and medium.

Nguyen et al. [4] come the closest, using NLP techniques to examine popularity specifically on fanfiction, mainly finding results that apply to how derivative works relate to the original, main source.

Weissburg et al. [5] studied the popularity of social media posts by analyzing how to optimize content/titles for a given audience.

Bandari et al. [6] examine the popularity of news articles based on a similar framework, however, with an emphasis on before an article even releases (so you do not need to release an article to gauge future popularity).

Pâquet [7] evaluates a case study between two different self-published web serials, exploring how popularity is gathered through community engagement and crowdsourcing.

Iwana et al. [8] train deep CNN models on book covers like our intended task, but use it to predict the probability of being in the top 1, 2, or 3 in a respective genre. They do note the difficulty of extracting useful information from purely book covers alone.

Overall, we see that these articles acknowledge the difference between offline and online medium for literature, and how different characteristics impact popularity, such as length, community engagement, timing, and appropriate framing for specific audiences.

## III. Methods

### *3.1 Data Collection*

The paper's focus is to collect web serials from various online websites where users independently post content, such as Royal Road, Scribble Hub, Wattpad, AO3, Fictionpress, Webnovel, and more.

However, these sites do not have publicly available datasets, necessitating the need for web scraping techniques to gather data. These sites have various policies on allowing this practice, leading to only some sites being ethically possible to gather from. See Notes 1 for more information.

Out of approximately 100,000 stories publicly published on Royal Road [10], the top 20,000 stories and a random selection of 5,000 stories from the remaining 80th percentile have been gathered, for a total of 24,990 stories.

The top 1000 out of approximately 40,000 stories have been gathered from the ScribbleHub [11] site.

Wattpad [12] has declined, allowing data collection.

The following data is collected from sites:

TABLE I. Raw Data Fields

| Field | Description |
|---|---|
| *Title* | The name of the story. |
| *Author* | The author of the story. |
| *Cover* | The cover image of the story. |
| *Blurb* | The short description of the story. |

| | |
|---|---|
| *Genres* | The genres of the story. |
| *Status* | Whether the story is ongoing, finished, on hiatus, or partially published as a sample. |
| *Chapter Count* | The total number of chapters published. |
| *Word Count* | The total number of words published. |
| *Ranking* | The position of the story on the website (i.e. #1 of all time). |
| *Rating* | The user-given review score (from 1 to 5). |
| *Rating Count* | The total number of user-given ratings. |
| *Views* | The total number of views on the story. |
| *Favourites* | The total number of people who favourited, or liked, the story. |
| *Chapter Dates* | The dates of publication for the first fifty chapters of the story. |
| *Chapter Names* | The names of the first fifty chapters of the story. |

### *3.2 Feature Engineering*

Converting most of the raw data into features for training models can be done easily. To ensure that different stories are treated fairly, their ranking is normalized with respect to their site's specific range.

For categorical data (genre and status), they need to be transformed into machine learning friendly formats. This can be done using a standard one-hot encoding scheme, where N possible values are converted into N different binary features. For text features (title and blurb), they use an embedding layer [9] to represent them as data-rich vectors. For our models, MiniLM-L6 was used to generate the embedding vectors.

As Nguyen et al. [4] mention in their analysis of fanfiction, metadata alone is usually not sufficient to accurately predict popularity. Feature engineering is typically done to create new, informative features to gain more predictive power. It typically involves using domain knowledge and analysis to determine what data provides the most value.

Compared to traditional publishing, web serials tend to take more advantage of the digital medium by embedding more text into their story's description or title. They may include the genre directly into the title, mention their release schedule, or what readers would be interested in their work.

Mining these features is done through the aid of binary classifiers. For our use case, we trained XGBoost models on ~1,000 manually labelled training samples to output a logit for the features.

Additionally, the initial release window for web serials could influence ratings, as inconsistency or long delays between releases could lower reader retention.

TABLE II. Engineered Features

| Field | Description |
|---|---|
| *Named Chapters* | Whether chapters are titled or just given generic numbers i.e. Chapter 1, Ch. 1, #1, etc. |
| *Title Genre* | Whether the title mentions the genre of the story. |
| *Blurb Genre* | Whether the blurb mentions the update/release frequency of the story. |
| *Blurb Audience* | Whether the blurb mentions the intended audience. |
| *Blurb Release Schedule* | Whether the blurb mentions the story's release schedule. |
| *Mean Release Frequency* | The average number of days between releases. |
| *Mean Release Standard Deviation* | The standard deviation in average number of days between releases. |
| *Title Word Length* | The length of the title in words. |
| *Blurb Word Length* | The length of the blurb in words. |
| *Title Character Length* | The length of the title in characters. |
| *Blurb Character Length* | The length of the blurb in characters. |

The correlation of some select numerical features in the dataset has been computed as seen in Fig. 1.

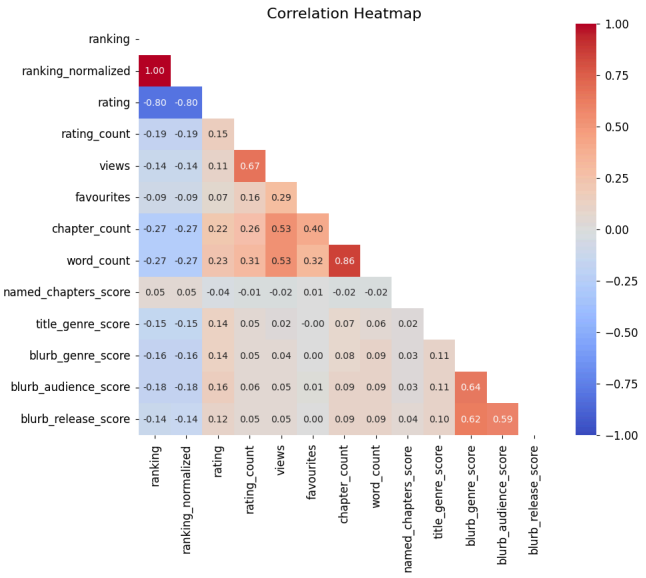



Fig. 1: Heatmap showing correlation mapping between numerical features.

When processing the data, there is also the case where several stories have identical book cover images. This is in the case when the author does not explicitly have an image to upload with their story, and the hosting website assigns a default. For our use case of having a student model predicting ratings exclusively from images, this would be unhelpful for training. As such, these examples are excluded from unimodal models' datasets.

### *3.3 Regression models and knowledge distillation.*

The overall dataset is split into a standard 80-10-10 split for training, validation and testing sets. Images are transformed during training with random resizing, cropping, and are fit to a standard 224 pixel by 224 pixel resolution.

We train two teacher models on the dataset, one trained to predict the normalized ranking of a story and the other trained to predict the user-given rating. These models use a ConvNeXt Base image encoder to process images [13] and fine-tune the last layer. This allows the model to have accurate pre-trained generic knowledge of images while being able to be trained for this specific task.

The tabular features are then concatenated with the image encoder's output to three linear perceptron layers with dropout, before passing to the final regression head. 1028, 512, and 128 were chosen for the layer sizes.

Training involved the classic Adam optimizer configured with weight decay for L1 regularization and gradient descent for the neural network with mean squared error (MSE) as the loss function. Cosine annealing is used as the scheduler.

The models are trained for a maximum of 100 epochs, though in practice, are terminated sooner due to early stopping, evaluating validation loss.

We then do a knowledge distillation where the teacher models use their added complexity to help student models train on the same dataset. For our unimodal models that have only access to image data, this makes it easier to learn the distribution of targets (rating or normalized ranking), despite not having as much information.

The unimodal models have a similar architecture to teachers, with a ConvNeXt Tiny image encoder passing to linear perceptron layers. There are only two layers of size 256 and 128, respectively.

Similar training hyperparameters are used for training students, with the exception of the loss function. For knowledge distillation, the student model is evaluated on its performance relative not only to the true target value, but also to the teacher's. The exact weighting of the teacher's influence is a hyperparameter, alpha. We set alpha to 0.5 for our task.

Besides these teacher and student models, four other models are trained to evaluate the performance of both knowledge distillation and unimodal models.

A set of smaller multimodal models was trained with the same layer sizes as the student models (256 and 128) and with the teacher distilling knowledge.

Another set of unimodal models was trained on the image dataset, but with no teacher influencing the loss function.

## IV. Results

The best performing models are those predicting ranking, with the multimodal teacher achieving the lowest root mean square error (RMSE) of 0.03. This model has a total parameter size of 90,119,041.

The next best performing model for ranking is the multimodal student with an RMSE of 0.04 with a parameter size of only 28,278,113, a 68.8% reduction.

Shown below in Table III is a comparison of all the models' performance.

TABLE III. All model performances

| Architecture | Target | Root Mean Square Error | Mean Average Error | Median Average Error | $R^2$ | Pearson | Spearman |
|---|---|---|---|---|---|---|---|
| Multimodal Teacher | Ranking | 0.032936 | 0.021269 | 0.015476 | 0.979743 | 0.989919 | 0.958966 |
| | Rating | 0.300350 | 0.225194 | 0.179510 | 0.382172 | 0.625091 | 0.630819 |
| Multimodal Student | Ranking | 0.040375 | 0.028371 | 0.021845 | 0.969559 | 0.984814 | 0.935362 |
| | Rating | 0.290053 | 0.215120 | 0.166083 | 0.423811 | 0.664280 | 0.670409 |
| Unimodal Student | Ranking | 0.222867 | 0.152478 | 0.105265 | 0.072483 | 0.272555 | 0.314195 |
| | Rating | 0.369335 | 0.285292 | 0.238775 | 0.065773 | 0.273302 | 0.277945 |
| Unimodal Standalone | Ranking | 0.219597 | 0.143837 | 0.090602 | 0.099501 | 0.324211 | 0.338457 |
| | Rating | 0.366998 | 0.284223 | 0.240062 | 0.077562 | 0.291966 | 0.293781 |

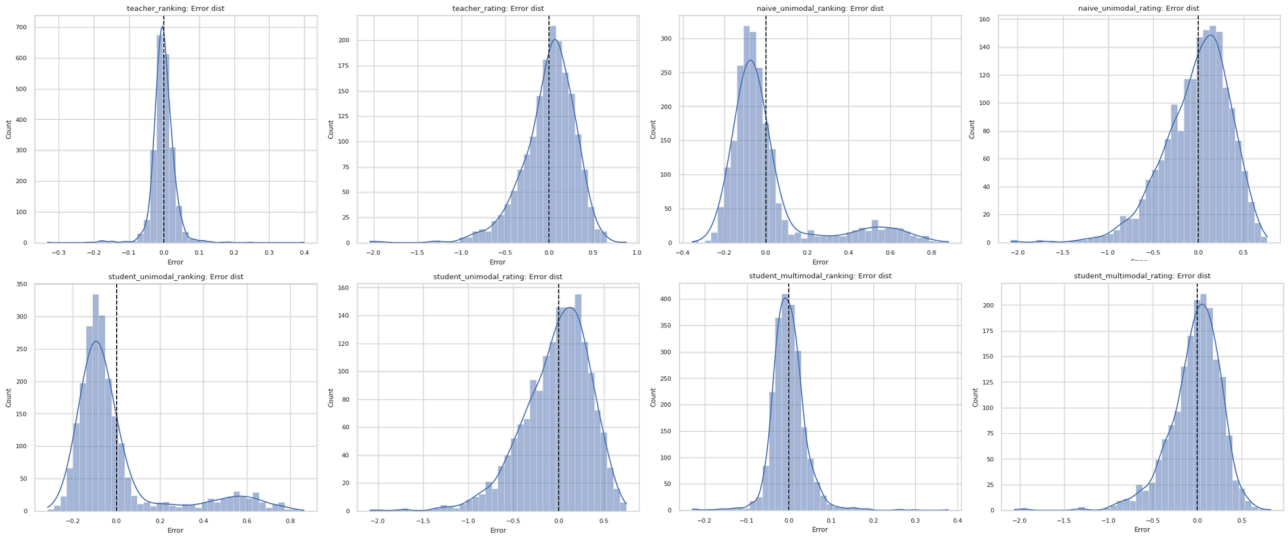



Fig. 2: The error distribution of the models' predictions on the dataset.

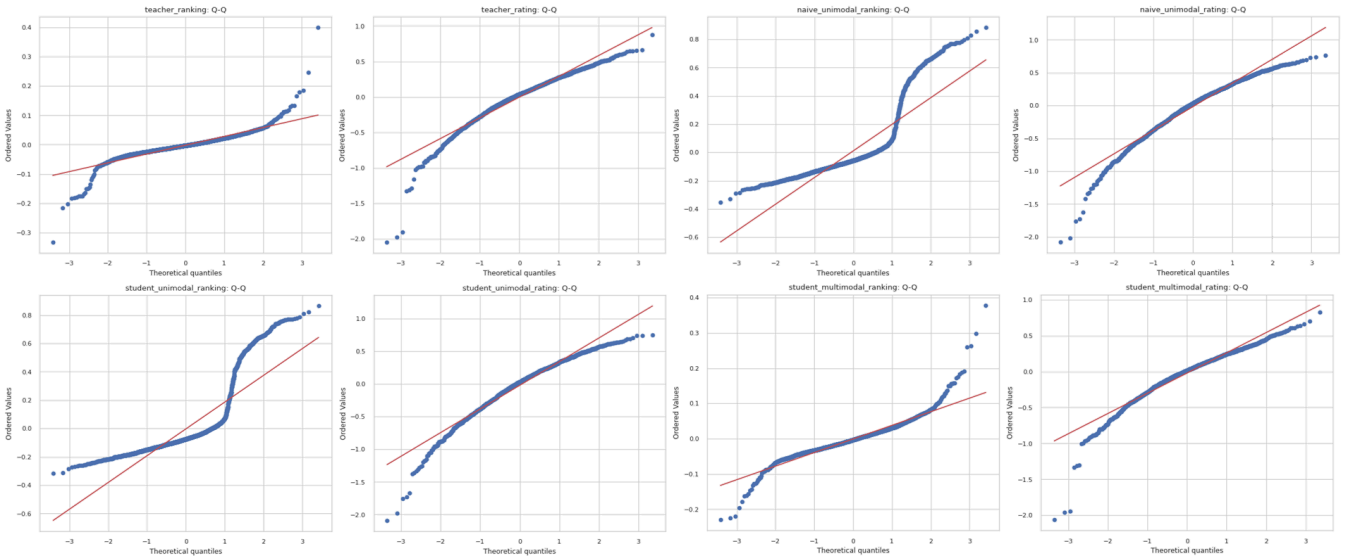



Fig. 3: The quantile-quantile plot of the models' predictions on the dataset.

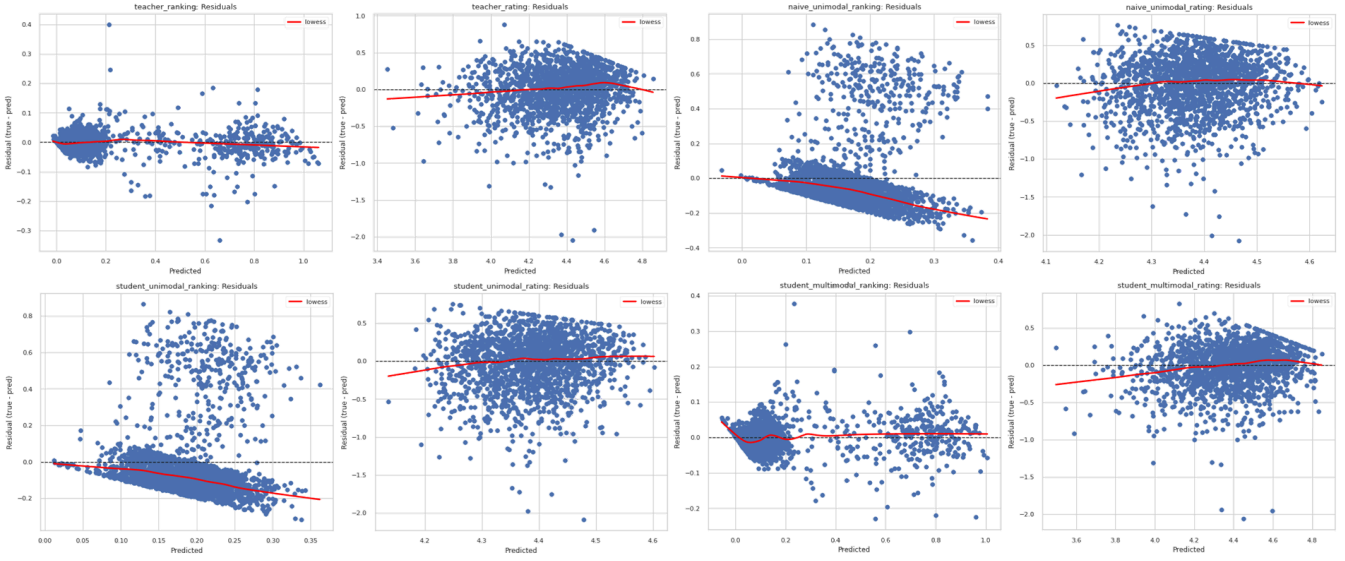



Fig. 4: The residual plot of the models' predictions on the dataset.

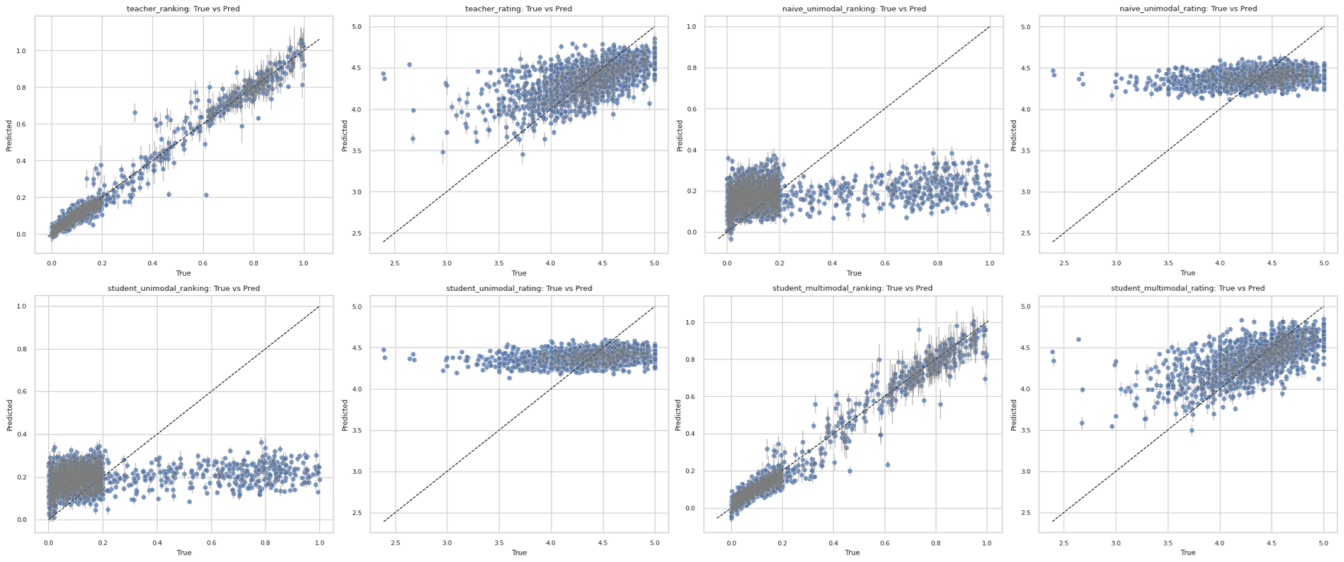



Fig. 5: The model's predicted targets on the dataset during four dropout rounds.

SHAP is a technique to help explain a model's predictions by evaluating how much a particular input contributes to the output score. When used on models' image features, it is possible to see what the model is focusing on.

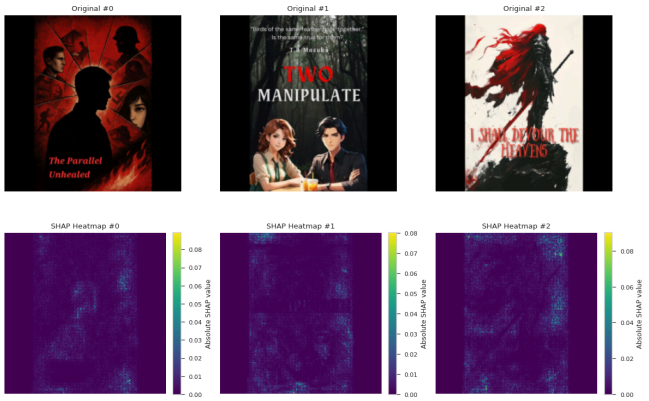



Fig. 6: The SHAP plot for the ranking teacher model.

We can also use LIME to help explain the importance of the tabular features for models, which works in a similar fashion to SHAP by evaluating how much input features influence outputs.

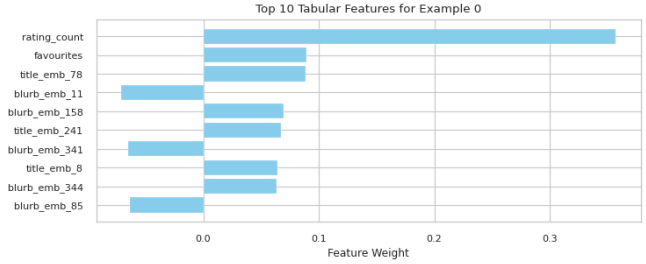



Fig. 7: The LIME plot for the ranking teacher model.

Of course, the overall importance of the text embeddings should include all the individual features of the vector. When summed up together, the text and blurb text features contribute heavily to a model's performance.

Of course, the overall importance of the text embeddings should include all the individual features of the vector. When summed up together, the text and blurb text features contribute heavily to a model's performance.

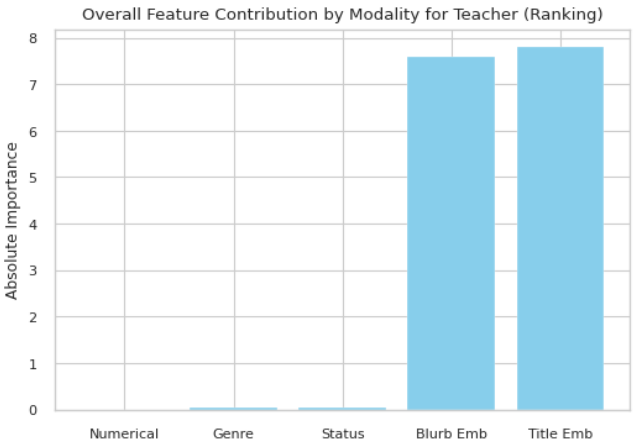



Fig. 8: The LIME plot for the ranking teacher model with the importance scores for the embeddings summed.

## V. Discussion

### 1. *Analyzing the Data*

Overall, we see that predicting user-given ratings is inherently lossy, with the best model achieving only 0.299 RMSE at that target. Given its $R^2$ value, it's only able to explain 42.4% of the dataset's variance.

We see that knowledge distillation provided the most benefit for predicting ranking, with a relatively small increase in RMSE for the multimodal student. For predicting ratings, the multinodal student actually achieved the highest performance. This is likely due to a bias-variance tradeoff where the less complex student was able to better capture the variance of the dataset at the cost of its meagre performance.

It isn't surprising for the unimodal models to perform worse than the multimodal models, given that they

inherently have less data. Additionally, Iwana et al. [8] acknowledge how image data for books does not necessarily capture that many predictive qualities, which is a potential limiter in this area.

However, it is surprising how the students of the unimodal models are the worst-performing models in their respective categories. While they aren't that much worse than their standalone unimodal counterparts (0.220 versus 0.223 for ranking and 0.367 versus 0.369 for rating), they still underperform in general. Theoretically, multimodal teachers should be able to distill higher-order knowledge to students and smooth out students' loss functions.

Considering the SHAP visualizations of images, we see how the models put relatively little emphasis on any part of the image. When viewing the predicted values for the unimodal models, it is clear that they tend to just converge to the mean of the dataset.

It can be concluded that images are largely irrelevant for predicting a web serial's overall rating. For evaluating stories, it seems easier for a model to predict a story's position (ranking) on a site than the actual rating. This is likely due to the fact that users who rate stories are inherently not representative of the overall quality of a story and are more likely to vote in extremes. Indeed, there are extremely few ratings below 3, which is the midpoint of a 1 to 5 Likert scale.

The limited dataset also makes it difficult to generalize the results to the wider population. Some genres, for instance, are more popular on Royal Road, the bulk of the dataset, which inevitably affects what patterns models are able to discover. Portal fantasy, for example, is a relatively niche genre in traditional media, but highly concentrated in this dataset, leading to it being relatively impactful.

Nevertheless, there are some trends that are likely to be widely applicable. The number of ratings, favourites, and views is highly correlated with higher ratings, which makes logical sense. We also see that the timing between releasing the initial chapters of a story impacts rating, likely because it is a measure of consistency. The number of chapters also influences ratings, again likely because it is another measure of consistency.

### 2. *Further Research*

Our work mainly involved Royal Road and partially Scribble Hub. With more time and access to other web serial sites, the performance and generalizability of the models could be improved. We can see clearly on the ranking model's performance how the variance increases after the 80th, where the data collection thinned out.

There could also be stronger predictive features that were not gathered during this study, either due to time or practical constraints. Community engagement among the online medium is particularly important and prevalent among RoyalRoad, but was not considered here. Certain metrics like subscriber count, views, etc, could be used in the future as more accurate metrics of quality.

If anyone would like to revisit this topic, we would recommend checking more mediums. We only chose a few websites to extract data from. If you use different sites there may be additional metrics you can attract e.g. some sites may have a "Time Read" function which shows viewers the average amount of time spent on a single page.

### 3. *Use in the Field*

While it may be obvious, many of these features are indirect markers of the entire web serial's quality, indicating that higher quality stories receive higher ratings. For authors, this may not seem highly actionable given the fact that a large proportion of web serial writers tend to be less experienced and independently publish.

However, the models trained have consistently shown that the title and blurb have proved to be highly influential indicators for predicting overall rating.

We can see how the combined title and blurb embeddings for multimodal models achieve higher LIME values, compared to all of the numerical, genre, and image features combined. This implies that authors should dedicate a significant amount of their time to fine-tuning and revising this portion of their work.

So while it seems that people do, in fact, not judge stories by their cover, they will, in fact, place a large emphasis on a web serial's title and blurb.

## VI. Notes

1. The Terms of Service (ToS) for Royal Road [9] explicitly prohibit web scraping without permission. Similarly, Wattpad also prohibits web scraping [10]. Scribble Hub does not have any provision against web scraping [11], but implements anti-crawling technology in the form of Cloudflare to hinder large-scale data scraping.

   To address this, permission has been requested and subsequently from Royal Road. Wattpad [12] has declined the request. A notice was sent to Scribble Hub for the data collection request.

   Data is subsequently collected using standard HTML GET requests when possible, and a simulated browser using Selenium otherwise. Collection is done over several days to never exceed the load associated with a human per period of time.

## VII. References


[1] S. Young, ‘Me Myself I: Revaluing Self-Publishing in the Electronic Age’, in *The Future of Writing*, J. Potts, Ed. London: Palgrave Macmillan UK, 2014, pp. 33–45.

[2] F. Pianzola, S. Rebora, and G. Lauer, ‘Wattpad as a resource for literary studies. Quantitative and qualitative examples of the importance of digital social reading and readers’ comments in the margins’, *PLOS ONE*, vol. 15, no. 1, pp. 1–46, 01 2020.

[3] Y. Wu. "Digital Globalization, Fan Culture and Transmedia Storytelling: The Rise of Web Fiction as a Burgeoning Literary Genre in China.," in *Critical Arts: A South-North Journal of Cultural & Media Studies*, vol. 37, no. 4, pp. 25-38, 2023.

[4] P. Nguyen. "Big data meets storytelling: using machine learning to predict popular fanfiction," in *Social Network Analysis and Mining*, vol. 14, no. 1, pp. 58, 2024.

[5] E. Weissburg, A. Kumar, and P. S. Dhillon, “Judging a Book by Its Cover: Predicting the Marginal Impact of Title on Reddit Post Popularity”, *ICWSM*, vol. 16, no. 1, pp. 1098-1108, May 2022.

[6] R. Bandari, S. Asur, and B. Huberman, “The Pulse of News in Social Media: Forecasting Popularity”, *ICWSM*, vol. 6, no. 1, pp. 26-33, Aug. 2021.

[7] L. Paquet. "The fan-networked capital of self-published web serials: A comparison of Worm and Nunslinger," in *TEXT*, vol. 23, 2019.

[8] Iwana, B., et al. "Judging a Book By its Cover," *Arxiv*, 2016.

[9] Jacob Devlin, et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[10] Royal Road, “Royal Road – terms of service,” Terms of Service | Royal Road, https://www.royalroad.com/tos (accessed Nov. 12, 2025).

[11] Wattpad, “Wattpad Policies,” Wattpad policies, https://policies.wattpad.com/terms/ (accessed Nov. 23, 2025).

[12] Scribble Hub, “Scribble Hub – terms of service,” Terms of Service | Scribble Hub, https://www.scribblehub.com/terms-of-service/ (accessed Nov. 12, 2025).

[13] Liu, Z., et al, " A ConvNet for the 2020s ," in 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , 2022, pp. 11966-11976.

## VIII. Supplementary Information

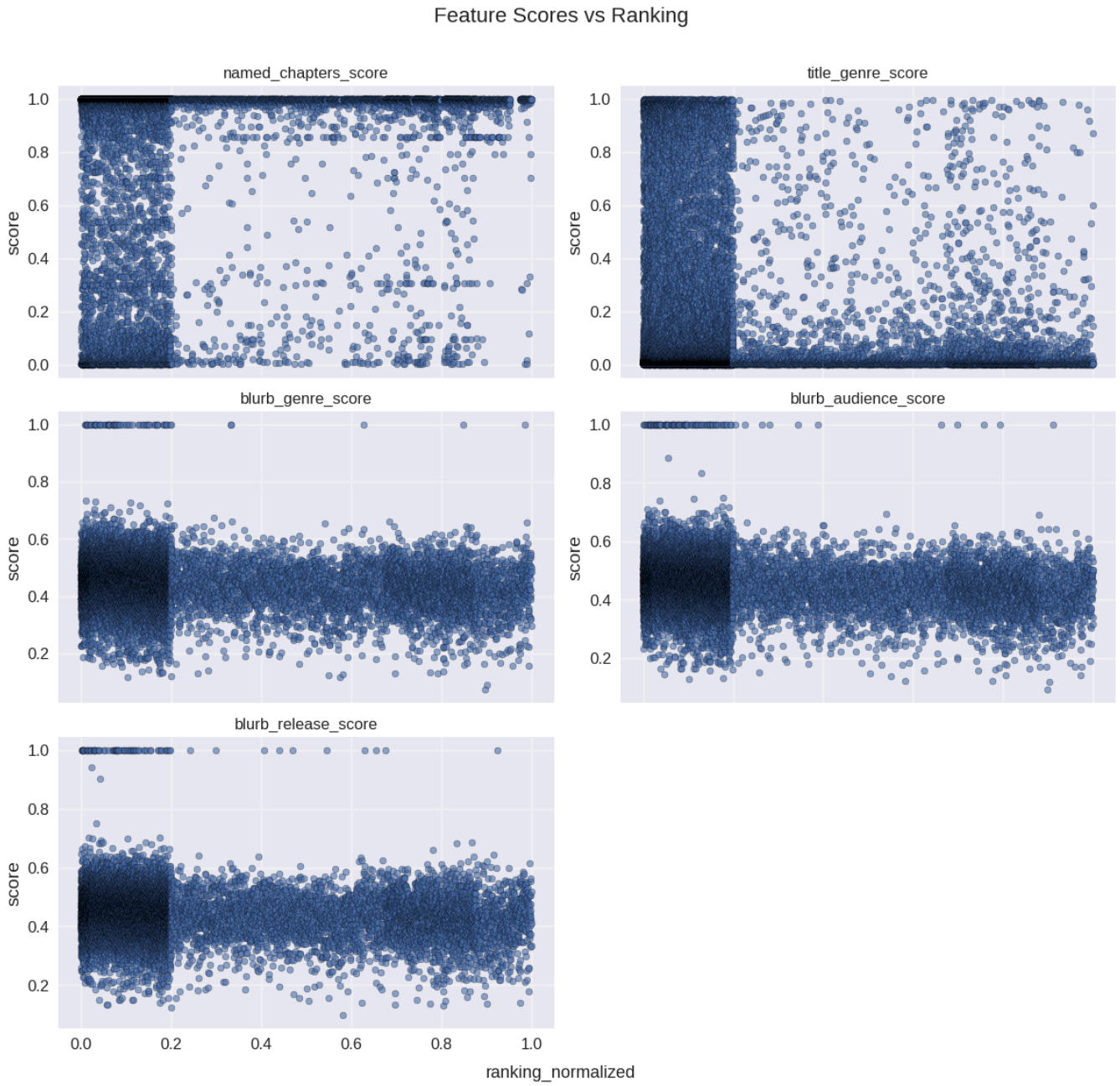
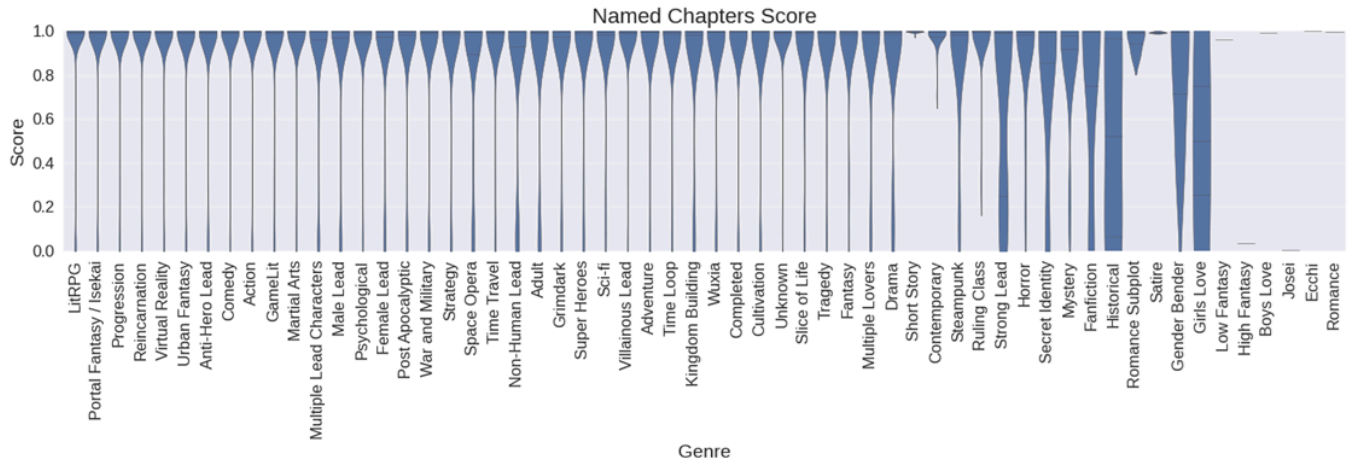
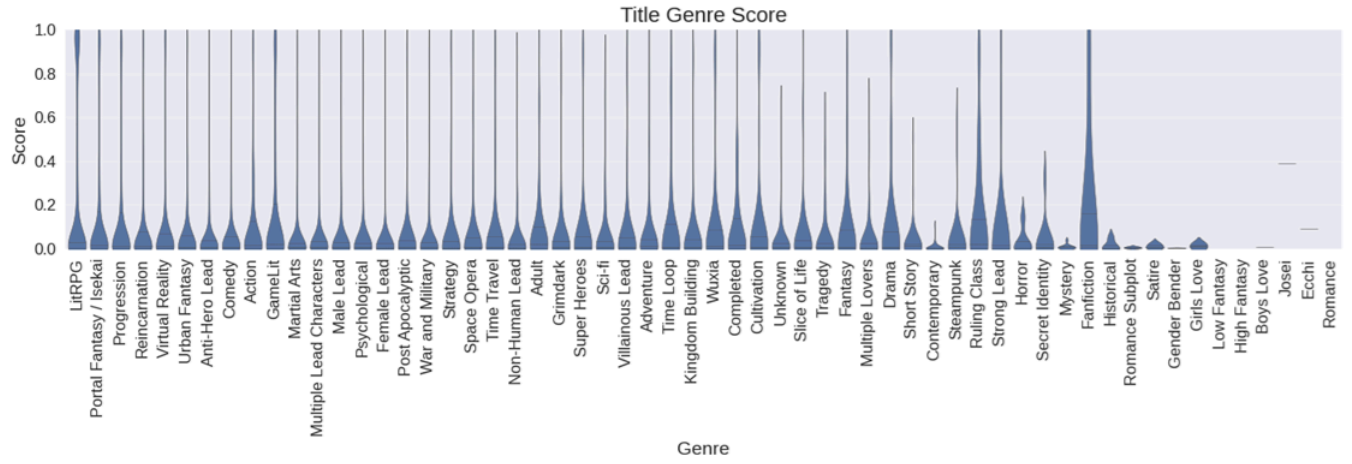
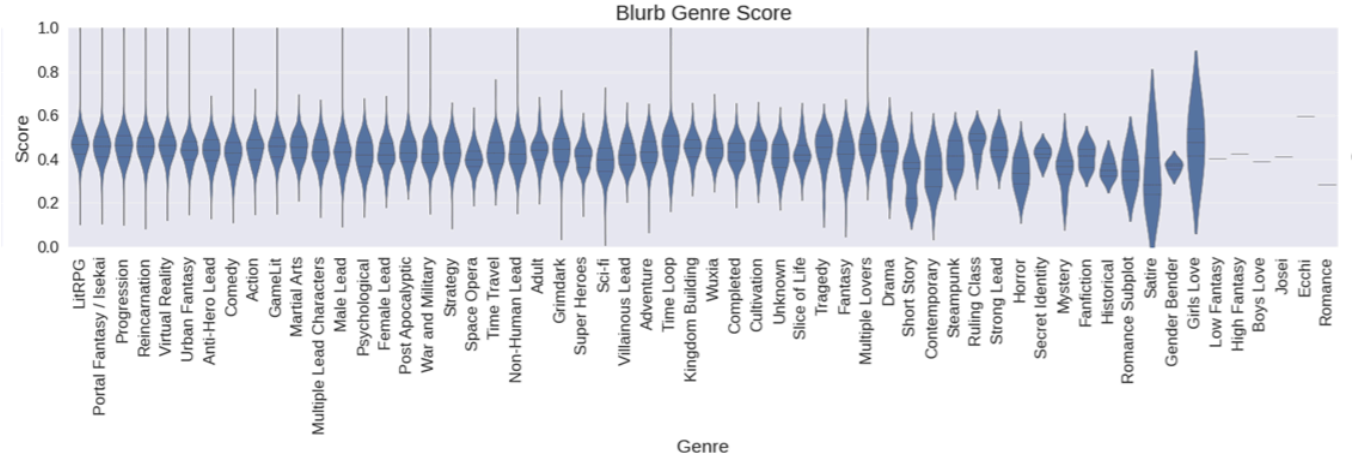



Fig. A: Visualization of the XGBoost binary classifiers. Each point refers to the score given to a specific data point.


Named Chapters Score
Score
1.0
0.8
0.6
0.4
0.2
0.0
LitRPG
Portal Fantasy / Isekai
Progression
Reincarnation
Virtual Reality
Urban Fantasy
Anti-Hero Lead
Comedy
Action
GameLit
Martial Arts
Multiple Lead Characters
Male Lead
Psychological
Female Lead
Post Apocalyptic
War and Military
Strategy
Space Opera
Time Travel
Non-Human Lead
Adult
Grimdark
Super Heroes
Sci-fi
Villainous Lead
Adventure
Time Loop
Kingdom Building
Wuxia
Completed
Cultivation
Unknown
Slice of Life
Tragedy
Fantasy
Multiple Lovers
Drama
Short Story
Contemporary
Steampunk
Ruling Class
Strong Lead
Horror
Secret Identity
Mystery
Fanfiction
Historical
Romance Subplot
Satire
Gender Bender
Girls Love
Low Fantasy
High Fantasy
Boys Love
Josei
Ecchi
Romance
Genre


Title Genre Score
Score
1.0
0.8
0.6
0.4
0.2
0.0
LitRPG
Portal Fantasy / Isekai
Progression
Reincarnation
Virtual Reality
Urban Fantasy
Anti-Hero Lead
Comedy
Action
GameLit
Martial Arts
Multiple Lead Characters
Male Lead
Psychological
Female Lead
Post Apocalyptic
War and Military
Strategy
Space Opera
Time Travel
Non-Human Lead
Adult
Grimdark
Super Heroes
Sci-fi
Villainous Lead
Adventure
Time Loop
Kingdom Building
Wuxia
Completed
Cultivation
Unknown
Slice of Life
Tragedy
Fantasy
Multiple Lovers
Drama
Short Story
Contemporary
Steampunk
Ruling Class
Strong Lead
Horror
Secret Identity
Mystery
Fanfiction
Historical
Romance Subplot
Satire
Gender Bender
Girls Love
Low Fantasy
High Fantasy
Boys Love
Josei
Ecchi
Romance
Genre


Blurb Genre Score
Score
1.0
0.8
0.6
0.4
0.2
0.0
LitRPG
Portal Fantasy / Isekai
Progression
Reincarnation
Virtual Reality
Urban Fantasy
Anti-Hero Lead
Comedy
Action
GameLit
Martial Arts
Multiple Lead Characters
Male Lead
Psychological
Female Lead
Post Apocalyptic
War and Military
Strategy
Space Opera
Time Travel
Non-Human Lead
Adult
Grimdark
Super Heroes
Sci-fi
Villainous Lead
Adventure
Time Loop
Kingdom Building
Wuxia
Completed
Cultivation
Unknown
Slice of Life
Tragedy
Fantasy
Multiple Lovers
Drama
Short Story
Contemporary
Steampunk
Ruling Class
Strong Lead
Horror
Secret Identity
Mystery
Fanfiction
Historical
Romance Subplot
Satire
Gender Bender
Girls Love
Low Fantasy
High Fantasy
Boys Love
Josei
Ecchi
Romance
Genre

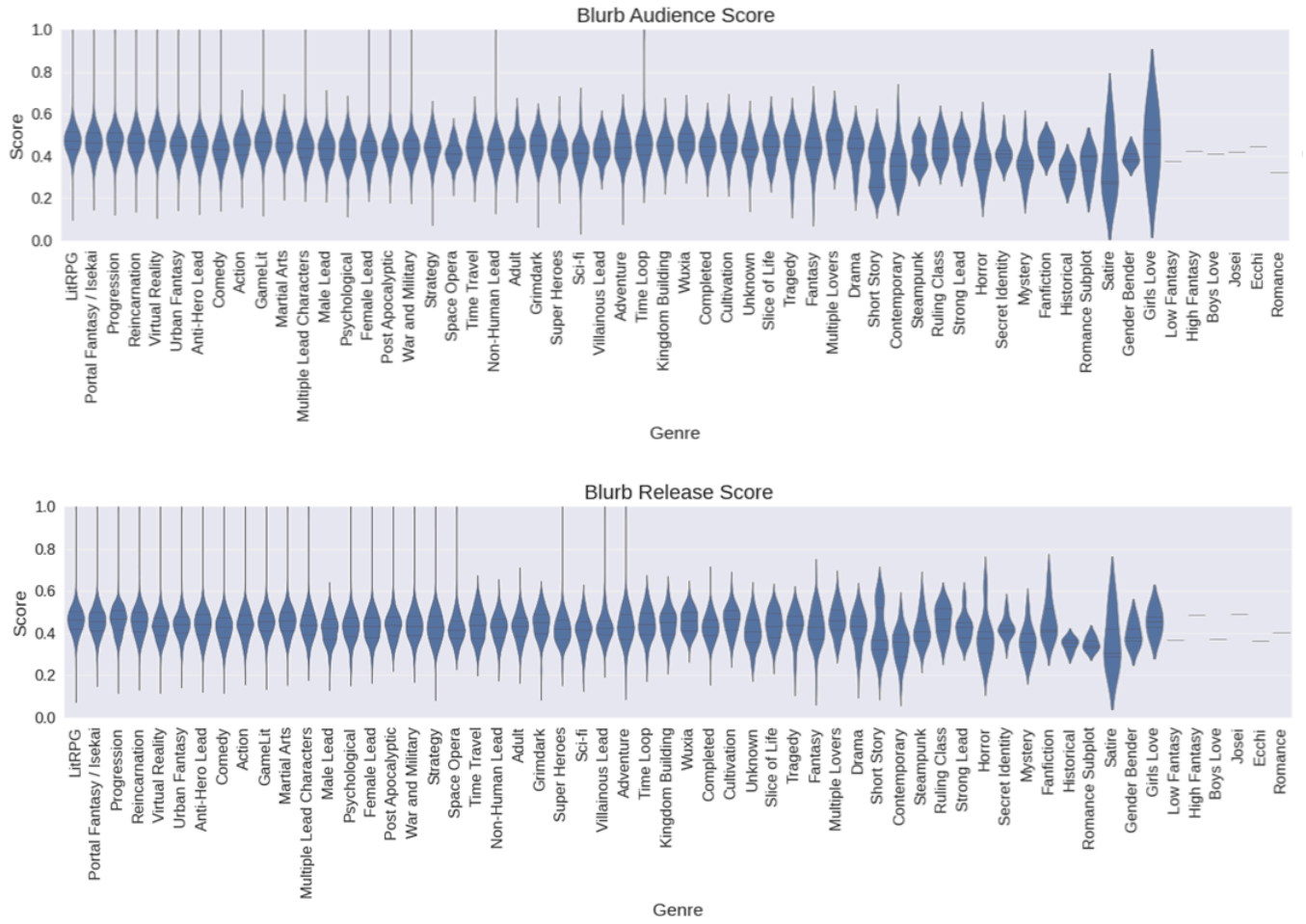



Fig. B: Violin plots of engineered feature scores by genre.

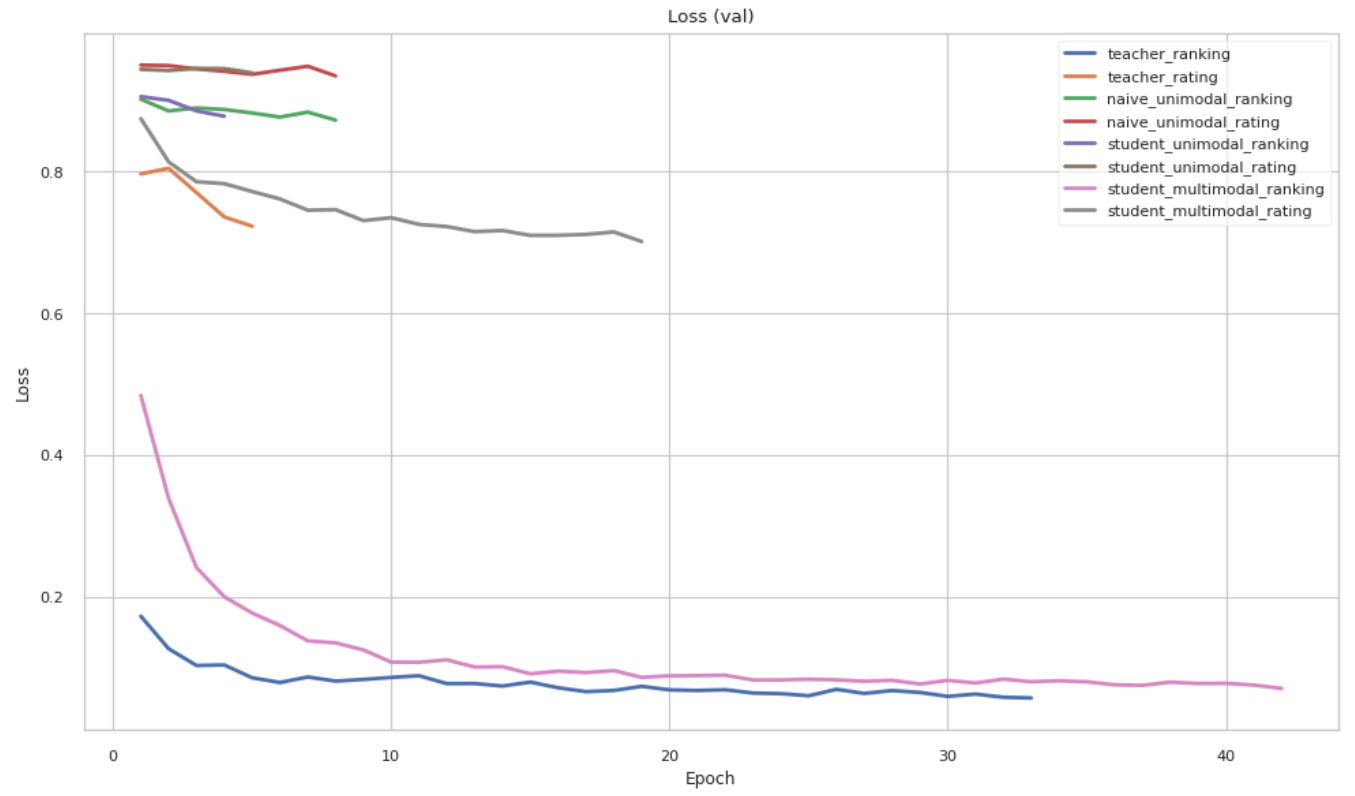



Fig. C: The validation loss of the models during training.