# Content Rating Classification in Fan Fiction Using Active Learning and Explainable Artificial Intelligence

Heng Yi Sheng and James Pope[a]
*Department of Engineering Maths, University of Bristol, Bristol, U.K.*

Keywords: Natural Language Processing, Text Classification, Fan Fiction, Content Rating Classification, Explainable Artificial Intelligence (XAI), Active Learning.

Abstract: The emergence of fan fiction websites, where fans write their own storied about a topic/genre, has resulted in serious content rating issues. The websites are accessible to general audiences but often includes explicit content. The authors can rate their own fan fiction stories but this is not required and many stories are unrated. This motivates automatically predicting the content rating using recent natural languages processing techniques. The length of the fan fiction text, ambiguity in ratings schemes, self-annotated (weak) labels, and style of writing all make automatic content rating prediction very difficult. In this paper, we propose several embedding techniques and classification models to address these problem. Based on a dataset from a popular fan fiction website, we show that binary classification is better than multiclass classification and can achieve nearly 70% accuracy using a transformer-based model. When computation is considered, we show that a traditional word embedding technique and Logistic Regression produce the best results with 66% accuracy and 0.1 seconds computation (approximately 15,000 times faster than DistilBERT). We further show that many of the labels are not correct and require subsequent preprocessing techniques to correct the labels. We propose an Active Learning approach, that while the results are not conclusive, suggest further work to address.

## 1 INTRODUCTION

The rapid evolution of fan fiction works throughout the years was driven by the endless creativity of fan fiction writers, whereby fan fiction culture has now been recognised globally, and it is also a form of social interaction among fandom communities. Fan fiction refers to fictional stories about fictional characters created by the fans of the original story or work, which can often be found published online on the internet. The original story could be from a famous TV series, movie, anime, video game, or book. Not to mention, the fanmade stories could also involve non-fictional settings such as celebrities or historical figures. In this digital era, fan fiction works have slowly become a digital practice on a global scale, whereby the fan fiction works that were shared online will be read by other fans (Vazquez-Calvo et al., 2019). Fan fiction stories can take various forms ranging from novels and short stories to poetry, whereby the content and length of the fan fiction stories can vary depending on the writers. Most importantly, fan fiction stories are created by the fans, contributing to the fictional worlds that the fans adore, which are not affiliated with the original author's works. The freedom of fan fiction allows writers to explore alternate storylines, different genres and sometimes to fill the gaps in the original work.

Fan fiction writers often publish their work on fan fiction platforms, such as Archive of Our Own (AO3) or fanfiction.net. According to AO3 (Archive of Our Own, 2023), more than 11 million works from more than 59 thousand fandoms are published on the website, as of August 2023. Moreover, a content rating should be given for each published fan fiction work, guiding the readers to understand the nature of the content the readers are about to read. However, problems arise when fan fiction writers are responsible for providing content ratings for the works. In this sense, either the content ratings are not provided for the fan fiction work, or the fan fiction works might be annotated incorrectly, which could be problematic. According to an article from CNN Health, fan fiction has many unhealthy themes and genres that are inappropriate for certain age groups (Knorr, 2017). For instance, sexual assault, domestic violence, toxic re-

[a] https://orcid.org/0000-0003-2656-363X



Heng, Y. and Pope, J.
Content Rating Classification in Fan Fiction Using Active Learning and Explainable Artificial Intelligence.
DOI: 10.5220/0012313400003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 224-231
ISBN: 978-989-758-684-2; ISSN: 2184-4313

lationships, and other explicit content. Furthermore, a recent UK study has reported that more than 63% of children aged 3 to 17 own a mobile phone and have regular access to the internet (Ofcom, 2022). Regularly exposing underage children to potentially unhealthy fan fiction content can seriously harm children's development and well-being. Hence, having a content rating for each fan fiction work will help protect vulnerable audiences, especially underage children, from exposure to content that might be harmful whilst preventing any psychological or emotional damage at a young age.

Natural language processing (NLP) is a growing interdisciplinary field, especially due to the ever-expanding text data in different industries (Minaee et al., 2021). As such, automated text classification has become widely popular and important. Hence, to solve the issue mentioned earlier, this paper proposes different classification techniques using machine learning and deep learning approaches to tackle the problem of content rating annotation whilst reducing the burden of the fan fiction writers of having to decide on a rating for the writer's own work. The classes for the content rating of fan fiction, which is the target variable, are as follows:

- **G-Rated.** General audiences
- **T-Rated.** Teen audiences, suitable for ages 13 and above
- **M-Rated.** Mature audiences
- **E-Rated.** Explicit content and only suitable for adults

The motivation of this project is to help automate the annotation of the content rating for fan fiction works for the convenience of the writers needing to annotate the content rating manually. Most importantly, it will help protect underage children from harmful, sensitive, or explicit content whilst complying with industry standards and compliance.

The contributions of the paper are as follows:

- Comparative analysis between different embedding techniques and classification models for fan fiction content rating.
- Used Explainable AI to discover weak labels in the dataset whilst potentially establishing trust in the model.
- Applied Active Learning (AL) approach to address the weak labels.

To our knowledge, this is the first fan fiction content rating analysis and use of Explainable Artificial Intelligence (XAI).

# 2 LITERATURE REVIEW

Numerous content rating classifications research of different domains has been conducted in the past. For instance, content rating for digital fiction books (Glazkova, 2020), social media (Barfian et al., 2017) and movies (Shafaei et al., 2019; Mohamed and Ha, 2020; Murat, 2023).

After conducting thorough research, most of the content rating classification research centres around movies or social media, but the fan fiction domain is given less attention. Since (Qiao and Pope, 2022) has highlighted the annotation issues that persist in the study, AL strategies will be adopted to address the weak labels. Moreover, it is evident that most of the content rating classification research only focuses on improving the performance of the classifiers. However, none of the studies have attempted to use XAI to understand how ML and AI model predicted the outcome and discover weak labels in the dataset. This is a huge gap presented in past studies. Hence, XAI techniques will be deployed that will act as a surrogate model in this study.

# 3 DATA PRE-PROCESSING PIPELINE

## 3.1 Data Collection Process

The fan fiction corpus will be scraped from the fan fiction online platform, Archive of Our Own (AO3) with the help of the prebuilt web crawler written in Java by (Donaldson and Pope, 2022). The working web crawler aims to scrape English fan fiction webpages from AO3. This includes the content rating, language, word counts, and other useful additional features (kudos, bookmark, and hits). In this project, almost 18,000 fan fiction (i.e., approximately 7 GB of data) was scraped from the AO3 website. However, a subset of these data will be randomly selected for experimentation, which will be discussed in the next section. Figure 1 shows the overview of the data collection process from the AO3 website.

## 3.2 Data Pre-Processing

Many researchers have outlined the importance of data cleaning, which helps improve data quality, enabling more robust ML and AI models to be created, even if the data cleaning process is costly and time-consuming (Ridzuan and Wan Zainon, 2019; Tae et al., 2019). As such, a data pre-processing pipeline

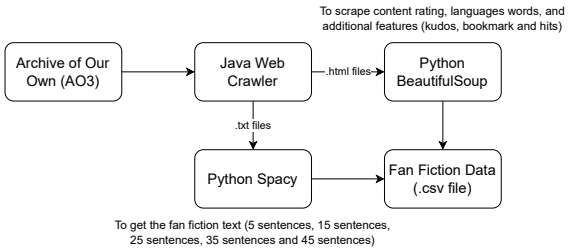



Figure 1: Data Collection Process.

will be designed to clean the fan fiction text corpus thoroughly before fitting it into the ML and AI models. Before applying the data pre-processing pipeline, a quick inspection of the fan fiction dataset was conducted, whereby it was evident that there were a few blank fan fiction contents. Also, some fan fiction text scraped by the web crawler was not English language. Hence, these abnormalities were filtered out accordingly. Furthermore, **to address the data imbalanced issue, a subset of 3000 instances from each class will be collected randomly** from the entire fan fiction data scraped from the AO3 website. After that, both rated and unrated fan fiction will be cleaned thoroughly:

1. Remove any URLs from the text
2. Replace the contractions
3. Remove the entity names (i.e., organisation, person, and location names)
4. Remove non-ASCII characters
5. Remove any hashtags, punctuations, and non-alphanumeric characters
6. Convert all the text to lowercase
7. Remove stopwords
8. Perform lemmatization while taking into account the verbs and adjectives part-of-speech (POS)

Finally, the cleaned data, excluding the "Not Rated" fan fiction, will be divided into training (64%), validation (16%) and testing (20%) datasets for further analysis and model building.

# 4 MODEL BUILDING AND OPTIMISATION

## 4.1 Evaluating Indicator

Model evaluation is an important part of the content rating classification task. This paper used accuracy and F1 score as part of the content rating classification evaluation process. Accuracy refers to the percentage of correct classification that a fully trained model achieved, whereas the F1 score is the harmonic mean of precision and recall. Since the class imbalance issue has been addressed as outlined in section 3.2, there will be not much discrepancy in the accuracy and F1 score. Furthermore, to ensure a fair comparison between models whilst mitigating overfitting issues, **the training process will be executed three times, whereby the mean of the accuracy and Macro-averaged F1 score will be recorded**.

## 4.2 Accuracy Against Different Number of Sentences per Instance

Figure 2 shows the experimentation conducted with the TF-IDF + Logistic Regression (LR) model to study the changes in accuracy on varying numbers of sentences per instance. The experiment reveals that the accuracy increases as the number of sentences per instance increases. Although it has a steep increase in accuracy from 5 to 25 sentences per instance, any further increase in the number of sentences will not lead to apparent accuracy increases. One assumption that can be made is that not all fan fictions are lengthy that consists of many sentences, as some fan fiction are just short stories. Hence, picking more sentences doesn't mean that it could result in better predictive accuracy. Not to mention, training an ML or AI model with more sentences could take a longer time and requires more computational resources. Therefore, after careful deliberation, **25 sentences per instance would be an ideal number for further experimentation**.

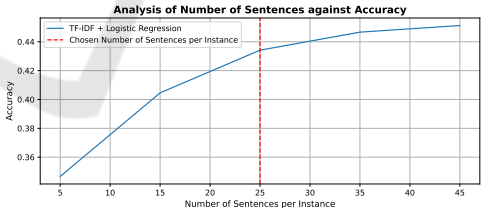



Figure 2: Analysis of Number of Sentences against Accuracy.

## 4.3 Experiment with Different Word Embedding Techniques

Table 1 shows the LR experimentation conducted with different types of word embedding techniques such as TF-IDF, GloVe, and Word2Vec to identify how different types of word embedding techniques affect the performance of the models. The research primarily relied on unigrams for experimentation and analysis. The result shows that TF-IDF outperforms GloVe and Word2Vec embeddings, even though both

GloVe and Word2Vec are part of the more recent developments in NLP.

One assumption that could be made is that the complex semantic relationships that GloVe or Word2Vec captured or learned are ineffective in the fan fiction task. Instead, a simpler representation such as TF-IDF is sufficient for the given task, and hence it outperforms much more complex word embedding techniques. Thus, **TF-IDF will be employed in this study**.

Table 1: Performance of Various Word Embedding Techniques.

| *Model* | *Word Embedding Techniques* | *Mean Accuracy* | *Mean F1 Score* |
|---|---|---|---|
| LR | **TF-IDF** | **41.54** % | **41.69** % |
| | GloVe | 38.75 % | 38.42 % |
| | Word2Vec | 28.63 % | 26.82 % |

## 4.4 Multiclass Classification

Table 2 shows the multiclass classification performance of different models with and without additional features (Kudos, Bookmarks and Hits). LR, XGBoost, BiLSTM and transformer-based models such as DistilBERT and TinyBERT slightly improved when it was trained with the additional features compared to the models trained without additional features. Nonetheless, some models like RF and LightGBM show no discernable improvements in classification performance, when additional features are added. It is suspected that these additional features, when included in the fan fiction text, might confuse these models during the training process. Among all the models trained with additional features, DistilBERT clearly outperforms all other models with an accuracy and F1 score of 44.18% and 44.91%, respectively. However, suppose the model's complexity, computation resources required, and training time are taken into account, a simple model such as LR model, which requires lesser computational power, is generally a better pick since the accuracy difference is only around 2% compared to DistilBERT that is complex in nature. Table 3 shows the computation time for all the models. Thus, further experimentation will be conducted with the binary classification approach before making a conclusive decision.

## 4.5 Binary Classification

Due to the poor multiclass classification performance, binary classification approach was adopted and evaluated accordingly. In this sense, the binary approach will be to classify fan fiction into either the "General Audiences" classes or the "Explicit" classes. Moreover, to avoid having to drop the "Teens And Up Audiences" classes and "Mature" classes and having to lose half of the training data, these labels will be converted to the respective labels shown as follows:

- Convert all "Teens And Up Audiences" labels to "General Audiences" labels
- Convert all "Mature" labels to "Explicit" labels

Table 4 shows the performance results of the binary classification. Surprisingly, transformer-based models such as DistilBERT and TinyBERT demonstrated a noteworthy boost in classification performance, showing an approximate increase of 3% when additional features are considered. For instance, DistilBERT emerged as one of the best-performing models achieving an accuracy of 69.74% and an F1 score of 69.72%, nearly crossing the 70% threshold.

However, given the complex architecture, substantial computation time (proven in table 3), and hardware prerequisites of the transformer-based models, it might not always be optimal to employ the transformer-based model. Instead, a much simpler alternative like the LR proves to be effective, delivering respectable performance. For instance, LR emerges as one of the top-performing models among the statistical and ensemble models, attaining an accuracy of 66.50% and an F1 score of 66.06%. Notably, this achievement is merely 3% lower than that of DistilBERT.

## 4.6 Summary and next Step

In summary, 25 sentences per instance is an ideal number to be fed into the ML and AI models during the training process, whereby employing more sentences per instance will not lead to an apparent increase in accuracy and might lengthen training time.

After performing various experimentations, it is evident that binary classification methods generally exhibit at least a 20% accuracy increment compared to multiclass classification methods, whereby this difference arises due to the inherently simpler nature of distinguishing two classes, as opposed to the fairly more complicated task of classifying four classes. In addition, DistilBERT, a state-of-the-art method clearly outperforms all models with the highest accuracy and F1 score. However, considering the limitation of employing the state-of-the-art methods explained previously, **adopting LR model might be a more feasible option** without having to worry about substantial computational time or hardware prerequisites. Hence, the LR model, which is considered a

Table 2: Performance of Different ML and AI Models - Multiclass Classification.

| **Model** | **Word Embedding Techniques** | *Without AF* **Accuracy** | *Without AF* **F1 Score** | *With AF* **Accuracy** | *With AF* **F1 Score** |
|---|---|---|---|---|---|
| **Logistic Regression** | **TF-IDF** | **42.01%** | **41.86%** | **42.11%** | **41.89%** |
| Random Forest | TF-IDF | 40.04% | 40.15% | 39.72% | 39.76% |
| XGBoost | TF-IDF | 39.25% | 39.44% | 39.75% | 39.95% |
| LightGBM | TF-IDF | 41.25% | 41.55% | 40.75% | 41.13% |
| BiLSTM | Tensorflow Embedding | 37.79% | 37.77% | 38.56% | 38.46% |
| **DistilBERT** | **DistilBERT** | **42.46%** | **42.98%** | **44.18%** | **44.91%** |
| TinyBERT | TinyBERT | 41.28% | 41.81% | 43.14% | 43.93% |

*Note: AF = additional features*

Table 3: Computation Time of Different ML and AI Models.

| ***Model*** | ***Computation Time (s)*** |
|---|---|
| Logistic Regression | 0.1147 |
| Random Forest | 9.5757 |
| XGBoost | 53.9561 |
| LightGBM | 58.2580 |
| BiLSTM | 157.0404 |
| DistilBERT | 1922.4387 |
| TinyBERT | 420.6284 |

relatively simple and computationally efficient model, will be employed to perform further experimentation.

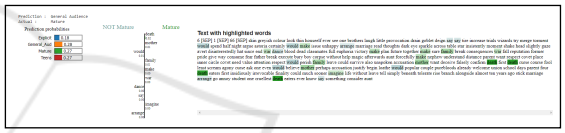
# 5 EXPLAINABLE ARTIFICIAL INTELLIGENCE

## 5.1 LIME with Multiclass Classification

According to the experimentation, all the prediction probabilities for most of the fan fiction content are nearly identical except for the explicit content, suggesting that the LR model is very uncertain about which class to choose. To be precise, the LR model lacks clear confidence in distinguishing between the four content rating classes, probably due to factors like inherent similarities between the classes or poor quality of labelled data.

The LIME local explainer helps improve the LR model's interpretability by uncovering important features that highly contribute to the model's decision, as presented in those highlighted words. Besides analysing the explanation for the correct prediction, the LIME explainer can also shed light by providing explanations for wrong predictions made by the LR model. Figure 3 shows the LIME explanation given the scenario when the LR model made the wrong predictions.

(a) Predicted Teens Class but Actual is General Audience Class.



(b) Predicted General Audience Class but Actual is Mature Class.

Figure 3: Comparison of the LIME Explanation Output in Multiclass Classification (Wrong Prediction).

Figure 3a reveals that the LR model predicted the fan fiction as a teen's content rating, but the actual is a general audience content rating. However, upon analysing the text, it is evident that the word "kill" is present, suggesting that this fan fiction might be more appropriate for teens or mature audiences. In contrast, figure 3b reveals that the LR model predicted the fan fiction suitable for the general audience, but the actual is a mature content rating. However, the prediction probabilities are almost similar across the general audience, teens, and mature content rating, with the general audience being 1% higher than the others. Besides, the text contains many "death" and "war" words, suggesting it is much more suitable for a mature audience. After careful deliberation and deduction, it is clear that there are some noisy fan fiction labels, as evident in figure 3a, causing the poor model's predictive performance whilst confusing the LR model during the training process. The following section will examine how the LIME explanation behaves in the context of the binary classification.

Table 4: Performance of Different ML and AI Models - Binary Classification.

| **Model** | **Word Embedding Techniques** | *Without AF* **Accuracy** | *Without AF* **F1 Score** | *With AF* **Accuracy** | *With AF* **F1 Score** |
|---|---|---|---|---|---|
| **Logistic Regression** | **TF-IDF** | **66.46%** | **65.93%** | **66.50%** | **66.06%** |
| Random Forest | TF-IDF | 64.64% | 64.54% | 65.00% | 64.90% |
| XGBoost | TF-IDF | 65.29% | 63.99% | 65.00% | 63.55% |
| LightGBM | TF-IDF | 66.08% | 65.28% | 65.42% | 64.53% |
| BiLSTM | Tensorflow Embedding | 64.10% | 63.94% | 63.68% | 63.62% |
| **DistilBERT** | **DistilBERT** | **67.32%** | **67.31%** | **69.74%** | **69.72%** |
| TinyBERT | TinyBERT | 65.35% | 65.34% | 68.29% | 68.29% |

*Note: AF = additional features*

## 5.2 LIME with Binary Classification

Similar experimentation was also being conducted in the context of binary classification, where XAI shows promising explainability outputs. However, some discrepancies still persist as shown in figure 4. For instance, the LR model predicted the fan fiction text as a general audience content rating, but the actual label is explicit. Nonetheless, the text itself does not seem to contain any explicit elements, which is evidence of poor quality labels presented in the fan fiction dataset.


Figure 4: Comparison of the LIME Explanation Output in Binary Classification (Wrong Prediction).

## 5.3 Critical Evaluation of XAI

In summary, the XAI technique, specifically LIME, helps provide insight into how the LR model arrives at a specific decision. Besides improving the robustness of the LR model, employing LIME helps identify weak labels and facilitates transparency in the LR model.

Upon analysing the LIME results, it is evident that there are lots of inconsistencies among the labels of the fan fiction datasets, resulting in poor predictive accuracy. In most cases, the model is very uncertain about which classes to choose, especially for multi-class classification, whereby even humans face difficulty categorising the content rating for certain fan fiction. The label inconsistencies are believed to stem from the stochastic nature of data collection when 25 sentences are gathered for each fan fiction. For instance, explicit labels do not contain explicit texts within the 25 sentences collected. Thus, future research should probably take note of such issues and potentially revise the data collection process before model building.

# 6 ACTIVE LEARNING

## 6.1 Pool-Based Active Learning Framework

Active learning (AL) is an learning algorithm that interacts with a human annotator (i.e., oracle) using a querying strategy to select and annotate the most informative instances from the unlabeled fan fiction dataset. The primary goal of AL is to potentially elevate model performance whilst helping the model to better generalise on unseen data. In addition, AL is an emerging approach that has been adopted in various text classification scenarios (Ul Haque et al., 2021; Zang, 2021; Al-Tamimi et al., 2021). As the performance of the fan fiction content rating classification is still unsatisfactory, the AL approach, specifically the pool-based AL framework, will be implemented in this study while evaluating its effectiveness towards the performance of the content rating classification task.

The pool of unlabelled fan fiction datasets (e.g., "Not Rated" content rating) will be gathered, whereby uncertainties for each unlabelled data point are computed. The strategy chosen to estimate the uncertainty is entropy, and the mathematical equation for entropy is given by:

$$\hat{x} = \underset{x}{argmax} - \sum_{i=1}^{C} P(\hat{y}_i|x) log P(\hat{y}_i|x) \qquad (1)$$

where $P(\hat{y}_i|x)$ is the conditional probability of the ith class $y$ for the given unlabelled instance x whereas $C$ refers to the number of classes.

Furthermore, for simplicity, the calculated entropy will be normalised to values between 0 and 1, such that the entropy value will be divided by the maximum uncertainty. In this sense, 0 signifies that there is no uncertainty, and 1 signifies that the model is very uncertain and will be selected for annotation. The selected data points will be passed over to the human annotator, and a new label will be given based on the fan fiction text. After the labelling process, the newly labelled data points will be added to the training set and retrain the LR model, in which the performance of the LR model will be measured using accuracy and F1 score. Figure 5 illustrates the workflow of the pool-based AL approach implemented in this study:

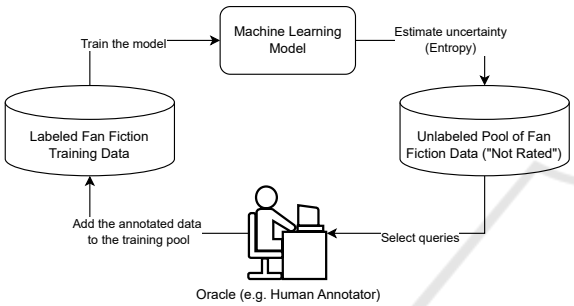



Figure 5: Pool-based Active Learning Framework.

## 6.2 Active Learning with Multiclass Classification

Figure 6 shows the performance results of AL in the context of multiclass classification comparing LR trained with only fan fiction text corpus and LR trained with additional features in addition to fan fiction text. The results reveal that adopting AL did not help improve the LR model's performance in classifying the fan fiction content rating, whereby the graph fluctuates around 41% and 42% accuracy. Moreover, it is also suspected that performing AL led to further confusion in the LR model, resulting in decreased accuracy as the number of newly annotated fan fiction increases. According to figure 6 , LR trained with plain text corpus achieves the highest accuracy of 41.80% with 250 newly annotated fan fiction, whereas LR trained with additional features in addition to fan fiction text achieves the highest accuracy of 42.24% with only 50 newly annotated fan fiction. Even though the accuracy discrepancy is only marginal, the model with additional features still performs better than the model trained with only plain text corpus. The next section will proceed with the AL experimentation using the binary classification approach before making a conclusive decision.

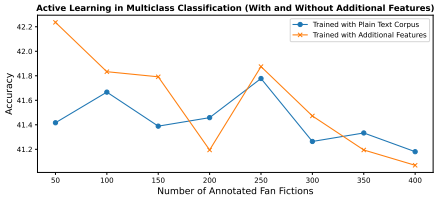



Figure 6: Active Learning in the context of Multiclass Classification (With and Without Additional Features).

## 6.3 Active Learning with Binary Classification

Figure 7 shows the performance results of AL in the context of binary classification comparing LR trained with only fan fiction text corpus against LR trained with additional features in addition to fan fiction text. Similar to the previous experimentation, the LR did not show obvious performance improvement even if the AL approach is adopted. Instead, the line graph shows a downward trend indicating that the model's accuracy worsens as the number of newly annotated fan fiction increases for both scenarios. Nonetheless, according to figure 7, LR trained with plain text corpus achieves the highest accuracy of 66.28% with 200 newly annotated fan fiction, whereas LR trained with additional features in addition to fan fiction text achieves the highest accuracy of 66.58% with only 50 newly annotated fan fiction. The following section will critically evaluate the AL approach in content rating classification while considering both multiclass and binary classification performances.

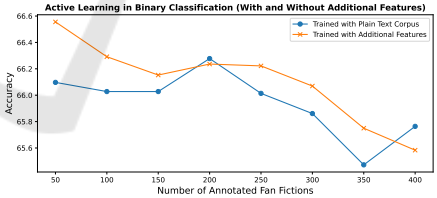



Figure 7: Active Learning in the context of Binary Classification (With and Without Additional Features).

## 6.4 Critical Evaluation of Active Learning

The pool-based AL approach has been applied and experimented on both multiclass and binary classification problems. However, upon meticulous evaluation of the outcomes, it was observed that the AL approach, in general, did not yield notable improvements in the predictive performance of the LR model for enhancing the predictive capabilities of the fan fiction content rating classification problem.

Several assumptions could be made about why AL

did not help improve the classification performance of the LR model. These include the existence of noisy labels in the initial pool of labelled fan fiction data, the selection of the uncertainty measure for the AL strategy, and the quality of the new annotations.

# 7 CONCLUSION AND FUTURE WORKS

Despite the results of the content rating classification carried out in this research, several improvements can be made to this project for future enhancement. For instance, future works could consider **including the summary part of the fan fiction apart from the main stories**, whereby this approach may help in better generalising the model.

In addition, future studies should also consider **adopting different AL strategies**, which could involve **experimenting with uncertainty measures other than entropy**. For example, least confidence, margin sampling, ratio sampling, or other uncertainty measurement techniques. Moreover, while **employing a professional annotator** might be costly, it ensures consistent annotations throughout the fan fiction dataset, maintaining controlled quality for the labels.

Last but not least, ML and AI models with **global explanations** could also be explored, which provide a high-level overview of how these models make certain decisions. An example will be employing the **SHAP** technique, whereby the impact of the features on the model output was computed with the Shapley value (Lundberg and Lee, 2017). Other XAI techniques, such as **Dalex** or **Shapash**, that support local and global explanations could also be taken into consideration, which might bring additional value to the content rating classification task.

# REFERENCES


Al-Tamimi, A.-K., Bani-Isaa, E., and Al-Alami, A. (2021). Active learning for arabic text classification. In *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pages 123–126.

Archive of Our Own (2023). A fan-created, fan-run, nonprofit, noncommercial archive for transformative fanworks, like fanfiction, fanart, fan videos, and podfic. https://archiveofourown.org/.

Barfian, E., Iswanto, B. H., and Isa, S. M. (2017). Twitter pornography multilingual content identification based on machine learning. *Procedia Computer Science*, 116:129–136. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

Donaldson, C. and Pope, J. (2022). Data collection and analysis of print and fan fiction classification. In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM,*, pages 511–517. INSTICC, SciTePress.

Glazkova, A. (2020). Text age rating methods for digital libraries. In *Proceedings of the International Scientific Conference "Digitalization of Education: History, Trends and Prospects" (DETP 2020)*, pages 364–368. Atlantis Press.

Knorr, C. (2017). Inside the racy, nerdy world of fanfiction. https://edition.cnn.com/2017/07/05/health/kids-teens-fanfiction-partner/index.html.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning based text classification: A comprehensive review.

Mohamed, E. and Ha, L. A. (2020). A first dataset for film age appropriateness investigation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1311–1317, Marseille, France. European Language Resources Association.

Murat, I. (2023). A smart movie suitability rating system based on subtitle. *Gazi University Journal of Science Part C: Design and Technology*, 11(1):252–262.

Ofcom (2022). Children and parents: media use and attitudes report 2022. Annual report, Ofcom.

Qiao, Y. and Pope, J. (2022). Content rating classification for fan fiction.

Ridzuan, F. and Wan Zainon, W. M. N. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.

Shafaei, M., Samghabadi, N. S., Kar, S., and Solorio, T. (2019). Rating for parents: Predicting children suitability rating for movies based on language of the movies.

Tae, K. H., Roh, Y., Oh, Y. H., Kim, H., and Whang, S. E. (2019). Data cleaning for accurate, fair, and robust models: A big data - ai integration approach.

Ul Haque, M. A., Rahman, A., and Hashem, M. M. A. (2021). Sentiment analysis in low-resource bangla text using active learning. In *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6.

Vazquez-Calvo, B., Zhang, L.-T., Pascual, M., and Cassany, D. (2019). Fan translation of games, anime, and fanfiction. *Language, Learning and Technology*, 23(1):49–71.

Zang, T. (2021). Active learning approach for spam filtering. In *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM)*, pages 366–370.