Happily Ever After: Comparing Sentiment Arcs in Emotionally-Inflected Fanfiction Genres Across Fandoms

Julia Neugarten[1], Pascale Feldkamp[2], Mia Jacobsen[2], and Yuri Bizzoni[2]

[1] Department of Arts and Culture Studies, Radboud University, Nijmegen, The Netherlands
[2] Center for Humanities Computing, Aarhus University, Aarhus, Denmark

Abstract

This paper uses sentiment arcs to compare three different genres of fanfiction – *Angst, Fluff* and *Hurt/comfort* – each characterized by particular emotionally-inflected content. We examine whether these arcs and arc development throughout stories reveal differences between the three genres. We also compare sentiment arcs across four fandoms: *Ancient Greek Religion and Lore, Harry Potter, Lord of the Rings*, and *Percy Jackson*. When using two different Sentiment Analysis methods – a BERT-model and the Syuzhet package [27] – mean sentiment differs significantly between two of the three genres. Additionally, four detectable clusters of sentiment arcs are dominated by particular genres in each case, conforming to expected patterns. Additionally, we find an ending effect – a tendency for stories' endings to be more positive than their beginning – in most stories regardless of genre. This suggests the therapeutic potential of fanfiction, as even the gloomiest stories tend to progress towards happiness or positivity in their sentiment. Finally, we also find that each fandom has its own emotional "bandwidth" with stories in the *Lord of the Rings* fandom consistently displaying the most positive sentiments while stories in the *Percy Jackson* fandom consistently display the most negative sentiments, regardless of genre.

Keywords: fanfiction, genre, sentiment arcs, happy endings, fandom

1 Introduction

Fanfiction – stories written by and for fans of existing stories, inspired by these stories and often published for free online – has been evocatively described as "emotional landscapes of reading" [43]. This description highlights the central role of emotions in fanfiction, both intradiegetically, as fictional characters experience them and as they often function to propel the plot forward, and extradiegetically, as readers seek emotional experiences in their reading choices. To navigate this immense emotional landscape of reading that fanfiction offers,[1] fans use various paratextual clues, including recommendations from fellow fans, metadata provided by platforms, and information about the reader reception of fanfiction through comments and hits.

The central role of emotion in fanfiction, which in early scholarly work on fan culture was termed the "affective sensibility of fandom" [18] even extends to the existence of particular genres defined by stories' emotional content and their intended emotional impact on readers. This paper examines three of these genres: *Angst*, *Fluff*, and *Hurt/comfort*. The *Angst*-genre denotes "writing which dwells on emotional problems, tortured minds and unconsummated relationships, and enjoys them immensely" [39, p. 242]. Conversely, *Fluff*-stories "lack emotional difficulties"

Julia Neugarten, Pascale Feldkamp, Mia Jacobsen, and Yuri Bizzoni. "Happily Ever After: Comparing Sentiment Arcs in Emotionally-Inflected Fanfiction Genres Across Fandoms." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 736–758. https://doi.org/10.63744/vimmV7M89Fq9.

[1] At the time of writing, Archive of Our Own (AO3) hosts over 16 million works. It is one of the largest English-language fanfiction platforms and the source of the datasets used in this paper.

[34, p. 207], focusing instead on (romantic) happiness and cuteness. Finally, *Hurt/comfort* encompasses stories where one character is hurt mentally or physically, and another character comforts or heals them. *Hurt/comfort* thus "combines elements from *Angst* and *Fluff*: the genre uses both emotional pain and its amelioration or absence to create narrative and emotional tension" [34, p. 207]. *Hurt/comfort* therefore often progresses from *Angst* to *Fluff*.

These genre-categories can be understood as emotionally-inflected and communally negotiated clues that help readers navigate the vast array of fanfiction available. Thus, they may function similarly to recommendations and metadata provided by other reading platforms, especially if stories exhibit within-genre similarity in sentiment dynamics – also across fandoms. If this is the case, based on these definitions, we expect the three genres to display differing patterns when it comes to valence, the "positiveness–negativeness/pleasure–displeasure" [32, p. 174] associated with a word or sentence. In this paper, we use sentiment arcs, a method of mapping the emotional trajectories of stories through their valence (positive/negative) calculated at the sentence level, to investigate three research questions about these three emotionally-inflected fanfiction genres:

- **RQ 1:** Do sentiment arcs differ significantly between the genres *Angst*, *Fluff*, and *Hurt/comfort*?
  - **Hypothesis:** We expect to find more negative sentiment in *Angst* and more positive sentiment in *Fluff*, with *Hurt/comfort* showing a development from negative to positive sentiment on the scale of individual stories.
- **RQ 2:** How do sentiment arcs develop over the course of the stories within genres?
  - **Hypothesis:** We expect sentiments to be relatively stable in *Angst* and *Fluff*, so that an *Angst*-story is sad throughout while a *Fluff*-story is happy throughout, accounting for consistent differences between genres. We also expect the order of the sentiments to influence genre categorization. Research shows that the valence of a story's ending impacts its reception [33], so we expect *Fluff* to end particularly positively while *Angst* ends especially negatively. We also expect *Hurt/comfort* stories to develop from negative to positive sentiment, as their protagonists progress from being hurt to being comforted.
- **RQ 3:** How do the sentiment arcs associated with the different genres compare across different fandoms?
  - **Hypothesis:** Previous research has shown similarity in sentiment arcs between three fandoms [22], but this analysis did not take within-fandom genre differences into account. The same paper identified stylistic differences between fandoms, so we have no strong hypothesis for this question.

2 Genres and Sentiment in Fanfiction

Fanfiction is often praised for its transformative capacities; since fanfiction is produced "outside of the literary marketplace" [11, p. 2], it can interrogate the cultural norms that structure mainstream literary production, for example by queering heteronormative stories [15] or through "racebending": rewriting white fictional characters as people of color [16]. Yet fanfiction is also "written within and to the standards of a particular fannish community" [11, p. 7], so its transformative potential is sometimes curtailed by community norms.

In this context, with fanfiction oscillating between the seemingly unlimited freedom of rewriting and the constraints of community expectations, emotionally inflected fanfiction genres are important objects of study. The genre-labels of *Angst*, *Fluff* and *Hurt/comfort* are unique to fanfiction

communities. In this sense, they illustrate the transformative potential of fanfiction, particularly in terms of its capacity to foreground, explore, and celebrate emotion. On the other hand, genre-labels also codify community norms that may impact textual production and constrain creativity. In that regard, they illustrate that fanfiction production is perhaps equally indebted to its marketplace dynamics as published fiction. Because of this tension, genre in fanfiction communities has been described as both 'stabiliz(ing) interpretation' and 'open-ended', like ongoing conversation [29, p. 6].

Furthermore, studying how fanfiction communities create, combine, and question genre categories – some of which are innovative from the perspective of more traditional literary scholarship – may generate new insights into how genres function in texts and reading more generally. Specifically, analyzing fanfiction's emotionally-inflected genres emphasizes that all categorizations of narrative – including but not limited to genres – and the choices prospective readers make based on them, are potentially emotionally inflected, as well as structured by value systems related to taste and prestige.

This research also connects to existing scholarship indicating that negative emotions (fear, sadness, disgust, anger) are more closely associated with fanfiction in the *Angst*-genre than in the *Fluff*-genre, when cosine similarity in a vector space model is used to operationalize genre-emotion similarity [37]. However, contrary to expectations, no statistically significant correlations were found between either *Angst* or *Fluff* and happiness. Because sentiment arcs can be measured at the level of individual stories rather than corpus-wide patterns, this paper can contextualize and perhaps explain these findings.

Despite the prevalence of emotionally-inflected genres and their integral role in fanfiction culture, they remain understudied within computational and digital humanities. Most studies using computational methods focus on the style of successful or well-liked fanfiction [23; 30; 38; 44] or on gendered power difference in fanfiction [35; 36; 45]. One study examined the reception of genres through a distant reading of fanfiction comments in *Harry Potter* fandom [34], but without considering the role of story content. By measuring textual features – sentiment arcs – from the stories themselves, the current paper addresses that research gap.

Specifically for sentiment analysis, previous scholarship has examined the persistence of sentiment arcs in fanfiction. Jacobsen et al. [22] measured the Hurst exponent – a measure of whether a particular arc is "trending, mean-reverting, or exhibiting a random walk behaviour" [22, p. 724] – of sentiment arcs in a corpus of fanfiction from three fandoms, along with two other textual measures: nominal style and readability. They found that each group of fanfiction, specifically fanfiction about *Harry Potter, Percy Jackson and the Olympians*, and *Lord of the Rings*, had distinct textual styles that reflected the complexity of their respective source materials. However, regarding narrative structure, the Hurst exponent was similar across fandoms. If fandoms differ in other ways, why not in terms of sentiment arcs dynamics – such as the ones measured by the Hurst-exponent? This might have to do with Hurst being (just) one possible measure of sentiment structure. This also raises the question of whether the lack of difference is related to internal diversity in terms of sentiment structure in each fandom. We turn attention to that matter in this paper.

3 Data

We use two datasets of fanfiction from four different fandoms. All fanfiction was collected from popular fanfiction-platform *Archive of Our Own* (AO3) in accordance with their terms of service.[2] All subsets contain only fanfiction written in English. Crossover stories tagged with more than one fandom were excluded. The fandoms are *Ancient Greek Religion and Lore* (AGRL), *Percy*

[2] `https://archiveofourown.org/tos`

Fandom	**Fics**	**Sentences**	**Words**	Other	Fluff	Angst	H/C
Ancient Greek Religion and Lore (AGRL)	3,584	1,932,084	25,168,026	2,940	268	290	86
Percy Jackson and the Olympians (PJ)	2,877	2,155,399	25,793,730	1,984	531	263	99
Harry Potter (HP)	2,886	2,886,189	38,269,933	2,194	347	265	80
Lord of the Rings (LOTR)	2,772	2,466,164	34,657,528	2,186	218	237	131
Total	12,119	9,439,836	123,889,217	9,304	1,364	1,055	396

Table 1: Dataset description: All fanfiction of over 50 sentences, divided by fandom. From left to right: number of stories in each fandom, number of sentences, and number of words. On the righthand side: number of stories in the category "Other" and in the three relevant genres per fandom. Note that as we applied sentiment models on the sentence-level, the number of sentences also corresponds to the final number of datapoints in our analyses (i.e., $n = 9,439,836$).

Jackson and the Olympians (PJ), *Harry Potter* (HP) and *Lord of the Rings* (LOTR). The AGRL dataset contains around 5,000 stories published in that fandom on AO3 at the time of data collection (2022). This dataset was previously described by Neugarten [35]. The fanfiction from PJ, HP, and LOTR in this study constitutes a subset of 3,000 texts from each group available on AO3. This corpus was first described in Jacobsen et al. [22]. Its texts were published between January 2002 and December 2023, and collected in the beginning of 2024.

Stylistic features, including the aforementioned Hurst exponent of sentiment arcs, have been compared before for the datasets taken from HP, PJ, and LOTR fandoms [22]. In this paper, we build on that analysis by examining the effect of genre-labels on sentiment arcs. Additionally, by expanding our analysis to fanfiction about Greek mythology, a fan community without clearly circumscribed source material, we increase the generalizability of our findings to other fandoms. This combination of datasets also potentially helps us understand the influence of a source material's cultural prestige in the fanfiction it inspires, since Greek mythology is traditionally more associated with prestige than young adult or fantasy fiction. In her analysis of the relationship between social class and Classics, Edith Hall observes that "education in the ancient Latin and Greek languages has always been an exclusive practice, used to define membership in an elite" [19, p. 386]. Consequently, Classics – and the associated domain of Greek myth – have traditionally been associated with academic prestige. Conversely, the "tradition of fantasy literature (...) has generally defined itself by its distance from the high-cultural (...) canon" [12, p. 60]. The differing levels of literary prestige associated with these fandoms may impact the fanfictions' sentiment arcs. Additionally, including fanfiction not based on contemporary fantasy literature may also reveal characteristics of fanfiction that exist independently from the genre or era of publication of the material being rewritten.

From each fandom, we selected only stories tagged in AO3's user-driven tagging system with 'Angst' or 'Fluff' or 'Hurt/Comfort'. To avoid interpretive confusion and ensure we were measuring textual characteristics specific to each genre, stories with more than one of these tags, or with tags indicating a blend or combination, such as "Angst and Fluff" or "Flangst", were discarded. This led to the exclusion of between 35% and 55% of stories per genre. Descriptive statistics for the resulting datasets are listed in Table 1. Additionally, as a control group, we used the remaining fanfiction not included in any of the three genre groups – listed in Table 1 under 'Other'.

4 Method

To measure differences in the ”sentimental palette” of different genres and across several fandoms, we use Sentiment Analysis to retrieve both the mean and the standard deviation of sentiment scores for each story's sentiment arc calculated at the sentence level (subsection 4.1), as well as to extract the raw and detrended versions of the stories' sentiment arcs. To observe whether there are significant differences in sentiment's mean and std between fandoms and genres, we use linear mixed-effects models (subsection 4.2). To check whether some sentimental trajectories are prevalent, overall or in specific subgroups, we use techniques of interpolation and clustering (subsection 4.3).

4.1 Measuring Sentiment Arcs

In computational literary studies, Sentiment Analysis has become a popular method for exploring the affective dimensions of literature [41]. It is often used to visualize the emotional arc of a narrative [26; 40], and to model affective dynamics over time using time-series techniques [4]. Since the introduction of the `Syuzhet` package in 2014 – the first sentiment analysis tool tailored to literary scholarship – more sophisticated approaches have emerged, including fine-tuned transformer models like BERT for classifying sentiment at the sentence or passage level [13; 28]. Typically, these models assign a single compound sentiment score (or valence) per sentence.

To retrieve the raw sentiment scores, we tested 4 different measures scoring sentences for sentiment (see Table 8 in Appendix A). We chose the dictionary method performing best on our test datasets – `Syzhet` [27] – and the best-performing transformer-based method – `xlm-roberta-base-sentiment-multilingual`[3] – here. We chose both methods, considering that dictionary-based tools have the advantage of staying closer to a human-like distribution overall, while transformer-based scores tend to “overestimate” scores to form distributions underlining extremes [13].

4.1.1 Detrending

As arcs based on valences are inherently noisy and nonlinear, studies typically apply some technique for detrending the arcs to reduce noise and extract global narrative trends – from a simple moving average window to more complex noise reduction techniques [6; 10; 17; 27]. As wavelet approaches typically used for noise reduction are not ideal for nonlinear series, [25] proposed an adaptive filtering technique for nonlinear series. The usefulness of adaptive filtering applied to sentiment arcs has been demonstrated, especially in the context of estimating the dynamics of sentiment arcs [5; 20], which is why we use it here (for a visual example of denoised arcs, see Figure 1).[4]

4.2 Statistical Analysis of Sentiment Arcs

To test for differences between fandoms and genres in their sentiment arcs, we used linear mixed effects models. These models are useful in this context, as they produce robust results even when dealing with imbalanced data. Additionally, they let us explicitly model the fact that one author may be in the data multiple times, if they have written multiple stories that fit the search criteria.

We created two linear mixed effects models to investigate the difference in sentiment arcs across fandoms and genres. The first model seeks to predict the mean of the sentiment arc for each work of fanfiction, based on an interaction between fandom and genre. This analysis shows

[3] `https://huggingface.co/cardiffnlp/xlm-roberta-base-sentiment-multilingual` We use the shorthand `xlm-roberta`. For full model names see Table 9 in Appendix A.

[4] All code for sentiment score retrieval and detrending is available at: `https://github.com/centre-for-humanities-computing/fanfic_sentiment`.

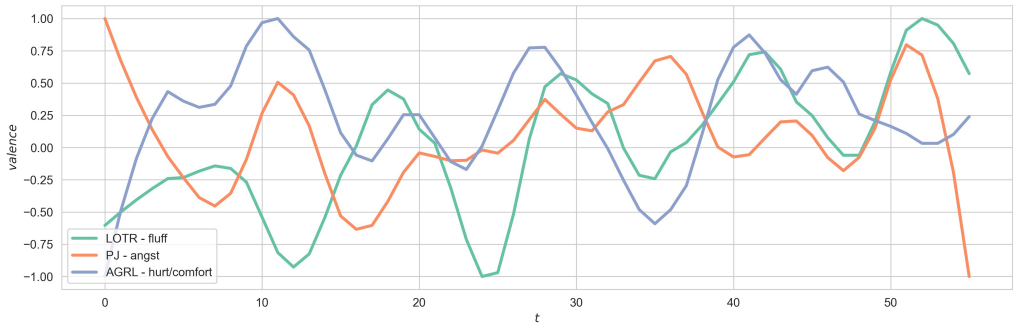


Figure 1: Example of 3 detrended `Syuzhet` arcs from *Lord of the Rings (LOTR)*, *Percy Jackson (PJ)*, and *Ancient Greek Religion and Lore (AGRL)*. Arcs are detrended using adaptive filtering.

whether some genres or fandoms are more positive or negative than others (**RQ 1**), and whether specific genres are written with different levels of sentiment in different fandoms (**RQ 3**).

The second model seeks to predict the standard deviation of the sentiment arc for each work of fanfiction, based on an interaction between fandom and genre. This analysis indicates how stable these arcs in specific groups are, based on which genres or fandoms have the greatest variation in sentiment over the course of the texts (**RQ 2** and **3**). Again, we are interested in both genre and fandom main effects, as well as the possibility of fandom-specific effects on the sentiment in different genres.

We added publishing date and word count to control for any confounding effect they might have. Since fanfiction is dynamic and published on an ongoing basis, it is appropriate to assume both time and length will have an effect on sentiment that we wish to control for. Additionally, we added a random intercept for author to control for repeating authors. To aid model convergence, word count was scaled. The specific model formulations can be seen below.

$$Mean\ Sentiment \sim genre * fandom + published\ date + word\ count + (1|author) \quad (1)$$

$$SD\ Sentiment \sim genre * fandom + published\ date + word\ count + (1|author) \quad (2)$$

4.3 Arc Shapes' Analysis

To analyze the shapes of sentiment arcs, we represented each narrative by a univariate time-series capturing the detrended sentiment score at successive narrative positions. Because the underlying fanfiction narratives differ in length, direct comparison required resampling every trajectory to a common temporal grid. Let $\mathbf{s} = (s\_1, \ldots, s\_n)$ denote the original sentiment arc of length n. A piece-wise linear interpolant $f : [0, n-1] \to \mathbb{R}$ was constructed.[5] The interpolant was then evaluated at $L = 50$ equidistant points,

$$t_j = (j-1)\,\frac{n-1}{L-1}, \qquad j = 1, \ldots, L,$$

producing the length–normalized trajectory $\mathbf{s}' \in \mathbb{R}^{50}$. We selected linear interpolation for its robustness to outliers and its preservation of monotonic segments; empirical inspection showed negligible benefit from higher-order kernels. We applied the procedure to both sets of sentiment estimates (`Syuzhet` and `xlm-roberta`), yielding two matrices $\mathbf{X} *$ Syuzhet, $\mathbf{X} *$ XLM $\in \mathbb{R}^{N \times 50}$, where N is the number of texts.

[5] Using `SciPy`'s `interp1d` (method `kind="linear"`)

We then performed unsupervised clustering with the K-means algorithm as implemented in `scikit-learn v.1.4.2`. The algorithm minimizes the within-cluster sum of squared Euclidean distances:

$$\arg\min_{\{C_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_2^2,$$

where $\boldsymbol{\mu}\_k$ is the centroid of cluster C_k.[6]

We identified the appropriate number of clusters, K, through the mean silhouette coefficient [42]. We chose silhouette analysis to determine both the optimal number and the coherence of clusters, as it quantifies both intra-cluster cohesion and inter-cluster separation, which allows to assess how distinct and consistent each cluster is.

5 Results

5.1 Statistical Models

With AGRL and *Other* as baseline conditions, results are in Tables 2 and 3.[7] For mean of the sentiment arc, we find – across sentiment methods – that *Angst* has lower mean sentiment and *Fluff* has higher mean sentiment than the set of non-genre works, *Other*. We find no effect for *Hurt/comfort*, meaning it displays no significant difference with regards to mean sentiment from *Other*. We find that LOTR fanfiction has higher mean sentiment than AGRL, but there is no effect for HP. For PJ, when using `xlm-roberta`, we find lower mean sentiment compared to AGRL, and both methods (`Syuzhet/xlm-roberta` show an interaction effect for PJ *Angst* fanfiction. This means that works of PJ fanfiction in general have lower sentiment, and that *Angst* stories in that fandom also have lower mean sentiment compared to *Angst* stories from other groups.

Mean `Syuzhet`	β	SE	t-value	p-value
Angst	-0.062	0.01	-5.97	<0.001*
Fluff	0.11	0.011	10.03	<0.001*
Hurt/comfort	-0.022	0.0185	-1.17	0.24
HP	0.00094	0.0053	.177	0.86
LOTR	0.044	0.0056	7.91	<0.001*
PJ	-0.0074	0.00559	-1.32	.19
PJ:Angst	-0.032	0.015	-2.06	<0.05*
Mean `xlm-roberta`	β	SE	t-value	p-value
Angst	-0.041	0.0066	-6.28	<0.001*
Fluff	0.0688	0.0068	10.12	<0.001*
Hurt/comfort	-0.012	0.012	-1.08	0.28
HP	-0.0053	0.0034	-1.56	0.11
LOTR	0.0097	0.0035	2.73	<0.01*
PJ	-0.0075	0.0035	-2.13	<0.05*
PJ:Angst	-0.029	0.0096	-3.02	<0.01*

Table 2: Estimates for model (1) for each sentiment analysis method (`Syuzhet/xlm-roberta`).

[6] We used fixed parameters to `init="k-means++"`, `n_init=10`, `max_iter=300`, and `random_state=42` to ensure replicability while mitigating sensitivity to initialization.

[7] For conciseness, we included only fandom and genre main effects in the table plus any significant interaction effects. See Appendix C for full model outputs.

SD `Syuzhet`		β	SE	t-value	p-value
	Angst	0.025	0.0090	2.77	<0.01*
	Fluff	-0.016	0.0092	-1.75	0.08
	Hurt/comfort	-0.0019	0.015	-0.12	0.90
	HP	-0.041	0.0051	-8.0	<0.001*
	LOTR	0.024	0.0055	4.38	<0.001*
	PJ	-0.074	0.0054	-13.68	<0.001*
SD `xlm-roberta`		β	SE	t-value	p-value
	Angst	0.0054	0.0027	2.0	<0.05*
	Fluff	0.0053	0.0028	1.89	0.059
	Hurt/comfort	0.0019	0.0048	0.41	0.68
	HP	-0.019	0.0014	-12.98	<0.001*
	LOTR	-0.010	0.0015	-6.77	<0.001*
	PJ	-0.018	0.0015	-11.65	<0.001*

Table 3: Estimates for model (2) for each sentiment analysis method (`Syuzhet`/`xlm-roberta`).

The lack of other interaction effects indicates that for *Fluff* and *Hurt/comfort*, there are no fandom effects, meaning the size of the difference in sentiment between *Other* and *Fluff* and between *Other* and *Hurt/comfort* is similar across fandoms. There is also no difference between HP fanfiction and AGRL fanfiction when it comes to mean sentiment. This suggests that rather than differing from fandoms with a more recent and circumscribed source material, fanfiction about Greek myth is similar to other, prominent contemporary fandoms.

The lack of interaction effects means that these genres are similar across groups. In other words, the intercepts (fandom effect) are different for each group but the slopes (genre effect) are similar. We show these findings in Figure 2. These visualizations show that each fandom has its own emotional "bandwidth"; in each fandom, *Fluff* has a more positive mean sentiment than *Angst* – the genre effect is the same in each group. This conforms to our expectations for **RQ 1**. However, the fandoms show up on these plots in the same general order (from most positive to most negative): LOTR, AGRL, HP, PJ. It thus seems that each fandom has its own baseline for sentiments – the fandom effects differ between groups. This difference between fandoms provides some insight into **RQ 3**. We hypothesize that these differences may be due to differing fandom demographics, with the *Percy Jackson* fandom skewing younger than the participants in the other fandoms.

For standard deviation (SD) of the sentiment arcs, the results are more method-dependent. Both methods find that only *Angst* has a significantly different sentiment SD compared to *Other* stories. The positive effect indicates that sentiment arcs for *Angst* display a greater variation than other fanfiction genres. The sentiment arcs for *Angst*-fanfiction are more varied than for the other groups – a partial answer to **RQ 2**. For HP and PJ, both methods find a negative effect, meaning compared to AGRL, HP and PJ fanfics have less variation in their sentiment arcs overall. For LOTR, `Syuzhet` finds a positive effect, meaning more variation, whereas `Roberta` finds a negative effect. This discrepancy might arise because `Syuzhet` is a continuous measure, while `Roberta` is categorical, meaning some confounding effects are potentially influencing the results. Because the specific effects as seen in Table 3 are rather small, further research is needed to cement these findings.

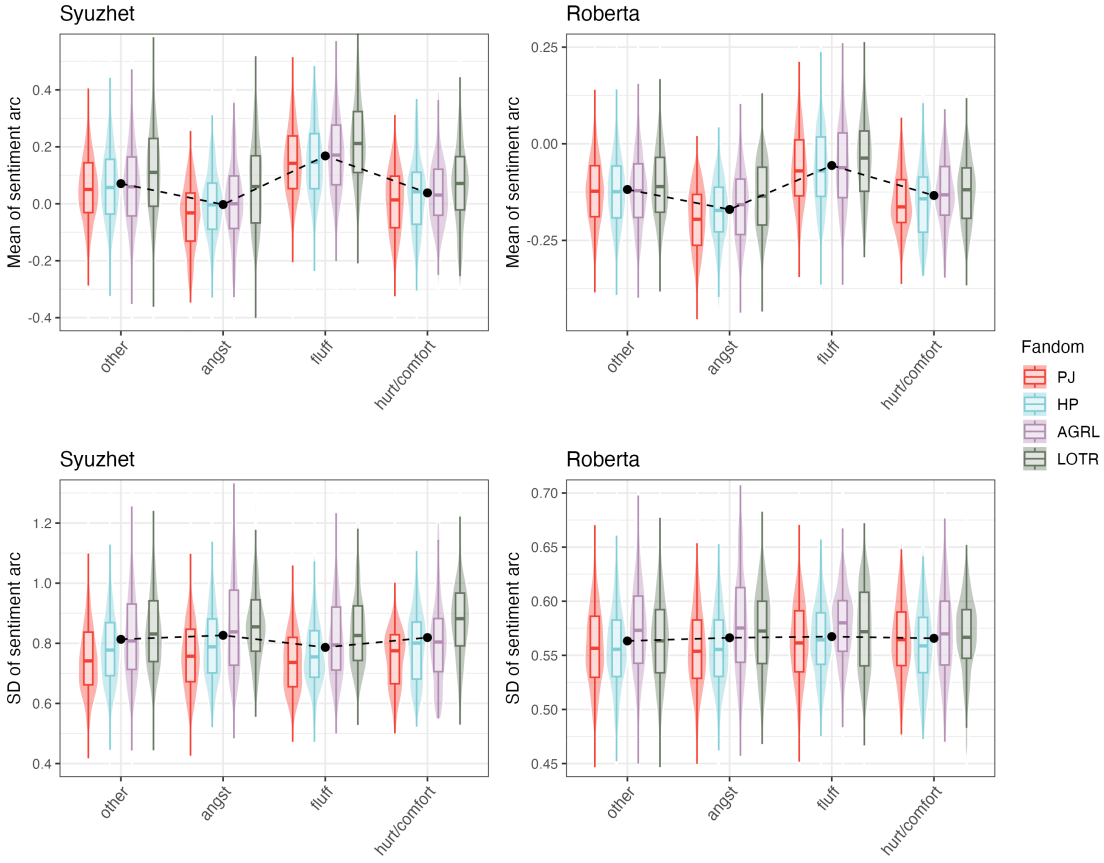


Figure 2: The mean and standard deviation for the sentiment arcs, across genres, fandoms, and sentiment methods. The boxplot indicates the 25th quantile, median, and 75th quantile for the fandom. The black dot shows the mean for the whole genre group.

5.2 Arcs' Shapes

Using cluster analysis, we can observe whether stories belonging to particular genres tend to replicate specific shapes in their sentiment arcs (**RQ 2**). Through silhouette analysis [42] we determine that the arcs, independently from the genres they belong to, create 2 to 4 meaningful clusters, with the strongest separation between two macro-clusters.

As visualized in Figure 3, most of these narrative shapes result in an "end point" that is more positive than their "starting point" even if this is not always the highest point of the story.

Table 4 describes the distribution of these four clusters across genres. Notably, the cluster with the highest percentage of *Angst* is also the only cluster displaying an overall downwards-bound shape (Cluster 2), while the cluster with the highest percentage of *Hurt/comfort* features the most distinctly "rags to riches" narrative path (Cluster 3), meaning an upward trajectory. Finally, the most frequent shape in *Fluff* includes a long central stretch of low feelings, but guarantees an extremely happy ending (Cluster 1).

In other words, the most frequent clusters in *Fluff* and *Hurt/Comfort* contain very "low lows" - especially in Cluster 1 - but end on a much more positive note, with respect to both their beginning and their average valence, than the *Angst*-related Cluster 2.

These results seem in accordance with the cognitive finding that the inner ordering, and especially the ending, of any experience matters greatly for its remembrance and categorization. Our findings support the idea that the valence at the end of a (reading) experience influences the way

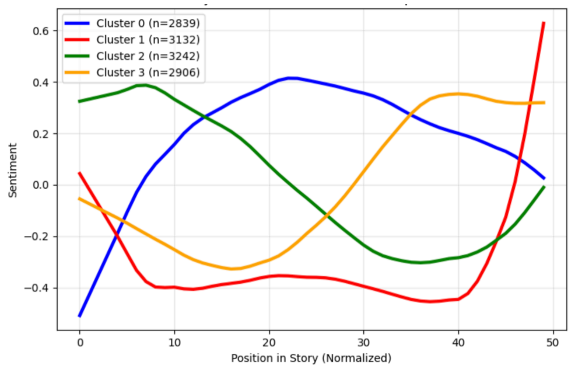


Figure 3: Abstracted shapes of the four main narrative clusters. Clusters 0 and 1 are the most distinct. All but Cluster 2 end "better" – meaning more positively – than they started. Cluster 1 has most *Fluff*, Cluster 2 has most *Angst*, Cluster 3 has most *Hurt/comfort*.

Genre	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Total
Angst	274 (26.0%)	257 (24.4%)	284 (**26.9%**)	240 (22.7%)	1055
Fluff	291 (21.3%)	396 (**29.0%**)	335 (24.6%)	342 (25.1%)	1364
Hurt/comfort	113 (28.5%)	91 (23.0%)	74 (18.7%)	118 (**29.8%**)	396
Other	2161 (23.2%)	2388 (25.7%)	2549 (**27.4%**)	2206 (23.7%)	9304
Total	2839 (23.4%)	3132 (25.8%)	3242 (26.8%)	2906 (24.0%)	12119

Table 4: Distribution of story genres across four narrative arc clusters using `Syuzhet`. Absolute counts and percentage (in parenthesis) of the genre for that sentiment arc cluster. Highest percentage per genre indicated in bold.

this experience is remembered [33], so a happy ending leads to the genre-classification of a happy story, despite low sentiments throughout the text: a positive ending makes a positive story, no matter how hard things got before, and vice versa.[8]

Genre	Uprising	Downfalling	Neutral	Total
Angst	528 (50.0%)	359 (26.3%)	212 (20.1%)	1,055
Fluff	765 (56.1%)	315 (29.9%)	240 (17.6%)	1,364
Hurt/Comfort	243 (61.4%)	91 (23.0%)	62 (15.7%)	396
Other	4,872 (52.4%)	2,516 (27.0%)	1,916 (20.6%)	9,304
Total	6,408 (52.9%)	3,281 (27.1%)	2,430 (20.1%)	12,119

Table 5: Narrative trajectory patterns by genre. Percentages refer to the genre (row).

Despite these differences, as shown in Table 5, all genres display a consistent majority of "happy endings".[9] The upward direction in narrative sentiments is preferred across all genres, holding the smallest proportion in *Angst*, where it represents half of the population. Thus, the dramatic *Angst* genre appears associated with negativity mainly in a relative sense, having the largest subset of tragic structures. The steepest and most frequent positive inclines characterize

[8] While in fanfiction communities the author – not the readership – applies genre-labels, these findings suggest that authors themselves strongly weigh the valence of their endings when applying labels.

[9] We computed these happy endings as arcs where the final 5% is at least 0.15 degrees higher than its first 5%. We defined these parameters to represent the beginning and ending of a story and a sentiment shift distinct enough to constitute a significant variation, but different parameters return similar results. See Appendix 11 for different ranges.

the *Hurt/comfort* genre, followed by *Fluff*. Meanwhile, on average, *Fluff* has the most positive peaks. These findings again indicate the relevance of the stories' shape and ending to their assigned genre category (see also Table 6). The fact that the upward direction or happy ending remains the single dominant structure throughout all of these clusters suggests that the authors have a marked tendency to provide emotional satisfaction – through a happy ending – for readers. This aligns with previous research suggesting that within the controllable and familiar environment of fanfiction communities, readers read about unpleasant or negative feelings to "resolve their own potentially overwhelming emotional states" [8, p. 15]. Our findings thus suggest that fanfiction, both on the production and consumption side, prefers happy endings for their capacity to ameliorate or work through difficult emotional trajectories. We call this fanfiction's therapeutic potential.

Genre	Method	Ending Avg	Begin → End	Overall → End	Max Positivity
Angst	`Syuzhet`	0.044	0.083	0.062	0.092
	`xlm-roberta`	0.064	0.193	0.136	0.100
Fluff	`Syuzhet`	0.060	0.117	0.109	0.113
	`xlm-roberta`	0.102	0.256	0.178	0.145
Hurt/Comfort	`Syuzhet`	0.164	0.298	0.180	0.056
	`xlm-roberta`	0.164	0.394	0.243	0.035
Other	`Syuzhet`	0.036	0.062	0.068	0.126
	`xlm-roberta`	0.075	0.192	0.138	0.120

Table 6: Sentiment Arcs' average ending, incline, and point of highest positivity

Finally, the distribution of stories from different fandoms over the four clusters (Table 7) generates some insight into the interplay between sentiments and fandoms (**RQ 3**) – particularly, the potential impact of the source material on the sentiment arcs of fanfiction. The AGRL and HP fandoms appear to have a relative preference for the "tragic shape" (Cluster 2). It intuitively makes sense that fanfiction in the AGRL fandom (stories rewriting Greek myth) is most prevalent in the cluster we call "tragic" here (Cluster 2). Simultaneously, this suggests fanfiction about AGRL may be replicating rather than transforming the sentiment arcs of its – admittedly quite diverse and diffuse – source material, for example by rewriting the tragic narrative of Achilles death in battle or Persephone's kidnapping at the hands of Hades.

Looking at the distribution of the sentiment arc clusters across fandoms, stories from the PJ and LOTR fandoms most often have happy ending arcs (Cluster 1), although differences between fandoms are small. This is especially striking since PJ and LOTR were more or less polar opposites in the comparison of mean sentiment above, with LOTR characterized by the most positive sentiment range or bandwidth and PJ by the most negative one. Nonetheless, here we find that these two fandoms tend towards similar sentiment arc shapes, those characterized by happy endings. This cluster intuitively makes sense for LOTR, which had the highest mean sentiment, but is surprising for PJ, where works in general have lower sentiment and *Angst* stories were most characterized by negative sentiment out of all groups. We conclude that negative sentiment in a story does not necessarily indicate an unhappy ending. In future work, the differences between fandoms could be analyzed further by comparing the sentiment arcs of fanfiction to those of its source material.

6 Conclusion

What does the statistical analysis and clustering of sentiment arcs reveal about the workings of emotionally-inflected fanfiction genres? Firstly, genre differences between the arcs (**RQ 1**) can be clearly detected for *Angst*, which has a significantly lower mean sentiment, and *Fluff*, which has

Fandom	N	Cluster 0	Cluster 1	Cluster 2	Cluster 3
AGRL	3584	832 (23.2%)	907 (25.3%)	1049 (**29.3%**)	796 (22.2%)
HP	2886	719 (24.9%)	705 (24.4%)	801 (**27.8%**)	661 (22.9%)
PJ	2877	692 (24.1%)	771 (**26.8%**)	673 (23.4%)	741 (25.8%)
LOTR	2772	596 (21.5%)	749 (**27.0%**)	719 (25.9%)	708 (25.5%)

Table 7: Percentage of each cluster in the different fandoms (`Syuzhet` method). Highest percentage per fandom in bold. Results for the `Roberta` method, which are similar, are in Appendix B

a significantly higher mean sentiment. No such effect was found for *Hurt/Comfort*. Additionally, *Angst* was relatively more associated with negative sentiments, since it was the largest subset of stories with a downfalling or tragic sentiment arc. In terms of the development of arcs (**RQ 2**), *Hurt/comfort* also conformed to expected patterns, since it was the largest subset in the cluster characterized by a development from negative to positive sentiments. Finally, *Fluff* was characterized by extremely happy endings. These findings all conform to our existing understanding of the genres.

More surprisingly, we also found evidence of an ending effect in all genres. In other words: all stories tend to end more positively than they begin, even those categorized as *Angst*, a genre known for its portrayal of sad or negative feelings and plot events. While research shows that reading about such negative experiences is often sought after and desired in fanfiction communities [8], we now have empirical evidence that the fanfiction texts themselves facilitate a kind of therapeutic process by having even the most miserable stories end more positively than they begin. Through this recurring narrative structure, we hypothesize that the stories let readers work through difficult emotional experiences vicariously.

The comparison of different fandoms (**RQ 3**) also generated some interesting insights into the workings of emotion in different subgroups of the fanfiction community. Out of all four inspected fandoms, *Angst*-stories about *Percy Jackson* were saddest, provided we take negative sentiment as a proxy for sadness. This is remarkable considering the PJ books could be considered the least tragic of the stories being rewritten in our data – compared to the other source materials, *Percy Jackson* contains little explicit violence or deaths of central characters. Additionally, we found that LOTR *Angst* is not at all sad when examined relative to the other fandoms. In other words: each fandom stays on its respective levels of happiness versus sadness. Every fandom has its own scale of emotions. While it has been argued that fanfiction exists to heighten or intensify the existing emotions of the original text [24], others argue that a primary motivation for fanfiction writing is the opportunity to explore what is *missing* from the text, especially as it pertains to characters' emotions [2]. Our findings support this second argument, that these "unseen mental states and emotions" [2, p. 76] come to dominate fanfiction in specific fandoms.

To conclude: fandoms can be distinctly characterized by their sentiment setpoint. Different fandoms each present a unique emotional "baseline" that seems to permeate readers' and writers' expectations. Additionally, from the clustering of sentiment arcs, we generally find expected patterns of narrative structure across genres. Computational sentiment analysis thus emerges as a valuable tool for examining narrative affect, contributing to literary reception studies, fan studies and computational humanities. A surprising finding was the positive ending effect across genres, even for the otherwise negative *Angst* genre. Additionally, our study supports theories from qualitative research on fanfiction, which argue that fanfiction writing is often motivated by the emotions that are missing from the source. Likewise, we find that fanfiction tends to fill emotional gaps, especially since fanfiction about the lighthearted *Percy Jackson* novels tends to be characterized by comparatively negative sentiment. Future research could extend these methods to other genres

or explore how higher-level textual features influence stories' emotional impact and engagement.

7 Limitations

One limitation of this research is that the subset of *Hurt/comfort* stories was relatively small, which may have influenced results. Since we were interested in pure genre effects, we designated texts with multiple genre-tags as *Other*. This probably led to an artificially small subset of *Hurt/comfort*. As mentioned previously, *Hurt/comfort* can be understood as stories going from *Angst* to *Fluff*, and fans might therefore decide to add these additional genres tags along with *Hurt/comfort*. As studies of genre are limited within fanfiction research, this controlled corpus was necessary here. The lack of "pure" *Hurt/comfort* and the overlapping use of genre-tags lends itself well to future work on these genres and their interaction with style and content.

The method of sentiment analysis also has limitations. Sentiment can be measured in various ways – both via different technical means, such as model types, and via different conceptual routes, such as polarity, intensity, etc. [21] – each with its own advantages and disadvantages. In this paper, we used two different methods: `Syuzhet` and `xlm-roberta` for continuous valence scoring. We show that these tools align most closely with human ratings (Appendix A).[10] Still, it is worth questioning to what extent machine-annotations can approximate the complexity of reader experiences [7]. Additionally, we have made the interpretive leap of connecting sentiment-scores to the level of happiness or sadness in stories. The fact that many of our findings confirm the difference between the typically sad genre of *Angst* and the typically happy genre of *Fluff* supports this decision. Nonetheless, sentiment analysis methods are still only one possible operationalization of a single aspect of a complex set of reader-text interactions.

Acknowledgements

Part of this research was supported by the Dutch ministry of Education, Culture and Science (OCW) through the Dutch Research Council (NWO), as part of the Anchoring Innovation Gravitation Grant research agenda of OIKOS, the National Research School in Classical Studies, the Netherlands (project number 024.003.012). Part of this research was supported by a Christine Mohrmann Stipend awarded by Radboud University.

References

[1] Barbieri, Francesco, Espinosa Anke, Luis, and Camacho-Collados, Jose. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, 2022, pp. 258–266. DOI: `10.48550/arXiv.2104.12250`.

[2] Barnes, Jennifer L. "Fanfiction as Imaginary Play: What Fan-written Stories Can Tell Us about the Cognitive Science of Fiction". In: *Poetics 48* (2015), pp. 69–82. DOI: `10.1016/j.poetic.2014.12.004`.

[3] Bizzoni, Yuri, Feldkamp Moreira, Pascale, Öhman, Emily, and Nielbo, Kristoffer L. "Comparing Transformer and Dictionary-based Sentiment Models for Literary Texts: Hemingway as a Case-study". In: *NLP4DH (forthcoming)*. Tokyo, Japan, 2023.

[10] For the persistently good performance of `Syuzhet`, also when detrending arcs, see [3].

[4] Bizzoni, Yuri, Moreira, Pascale, Thomsen, Mads Rosendahl, and Nielbo, Kristoffer. “Sentimental Matters - Predicting Literary Quality by Sentiment Analysis and Stylometric Features”. In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 11–18. DOI: 10.18653/v1/2023.wassa-1.2.

[5] Bizzoni, Yuri, Peura, Telma, Nielbo, Kristoffer, and Thomsen, Mads. “Fractality of Sentiment Arcs for Literary Quality Assessment: The Case of Nobel Laureates”. In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*. Taipei, Taiwan: Association for Computational Linguistics, 2022, pp. 31–41. DOI: 10.18653/v1/2022.nlp4dh-1.5.

[6] Bizzoni, Yuri, Peura, Telma, Thomsen, Mads Rosendahl, and Nielbo, Kristoffer. “Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences”. In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NIT Silchar, India: NLP Association of India (NLPAI), 2021, pp. 1–6. URL: https://aclanthology.org/2021.nlp4dh-1.1.

[7] Boot, Peter, Daza, Angel, Schnober, Carsten, and Hage, Willem van. “In the Context of Narrative, we Never Properly Defined the Concept of Valence”. In: *CHR 2024: Computational Humanities Research Conference*, Aarhus, Denmark, 2024, pp. 740–760. URL: https://ceur-ws.org/Vol-3834/paper67.pdf.

[8] Bruns, Cristina Vischer. “Stinging or Soothing: Trigger Warnings, Fanfiction, and Reading Violent Texts”. In: *Journal of Aesthetic Education 55*, no. 3 (2021), pp. 15–32.

[9] Buechel, Sven and Hahn, Udo. “EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 578–585. DOI: 10.48550/arXiv.2205.01996.

[10] Chun, Jon. “SentimentArcs: A Novel Method for Self-Supervised Sentiment Analysis of Time Series Shows SOTA Transformers Can Struggle Finding Narrative Arcs”. Oct. 2021. DOI: 10.48550/arXiv.2110.09454.

[11] Coppa, Francesca. *The Fanfiction Reader: Folk Tales for the Digital Age*. Ann Arbor: University of Michigan Press, 2017.

[12] Eatough, Matthew. “” Are They Going to Say This Is Fantasy?”: Kazuo Ishiguro, Untimely Genres, and the Making of Literary Prestige”. In: *MFS Modern Fiction Studies 67*, no. 1 (2021), pp. 40–66. DOI: 10.1353/mfs.2021.0002.

[13] Feldkamp, Pascale, Kostkan, Jan, Overgaard, Ea, Jacobsen, Mia, and Bizzoni, Yuri. “Comparing Tools for Sentiment Analysis of Danish Literature from Hymns to Fairy Tales: Low-Resource Language and Domain Challenges”. In: *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, ed. by Orphée De Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 186–199. DOI: 10.18653/v1/2024.wassa-1.15.

[14] Feldkamp, Pascale, Lindhardt, Ea Overgaard, Nielbo, Kristoffer L., and Bizzoni, Yuri. “Sentiment Below the Surface: Omissive and Evocative Strategies in Literature and Beyond”. In: *Proceedings of the Computational Humanities Research Conference*, ed. by Wouter Haverals, Marijn Koolen, and Laure Thompson. Vol. 3834. 2024, pp. 681–706.

[15] Floegel, Diana. ""Write the story you want to read": World-queering through Slash Fanfiction Creation". In: *Journal of documentation 76*, no. 4 (2020), pp. 785–805. DOI: 10.1108/JD-11-2019-0217.

[16] Fowler, Megan Justine. "Rewriting the School Story through rRacebending in the Harry Potter and Raven Cycle Fandoms". In: *Transformative Works and Cultures 29* (2019). DOI: 10.3983/twc.2019.1492.

[17] Gao, Jianbo, Jockers, Matthew L, Laudun, John, and Tangherlini, Timothy. "A Multiscale Theory for the Dynamical Evolution of Sentiment in Novels". In: *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*. IEEE. Durham, NC, USA, 2016, pp. 1–4. DOI: 10.1109/BESC.2016.7804470.

[18] Grossberg, Lawrence. "Is there a Fan in the House?: The Affective Sensibility of Fandom". In: *The Adoring Audience: Fan Culture and Popular Media*. London: Routledge, 1992, pp. 50–65.

[19] Hall, Edith. "Putting the Class into Classical Reception". In: ed. by Christopher Stray Lorna Hardwick. Hoboken: Wiley-Blackwell, 2008, pp. 386–397. DOI: 10 . 1002 / 9780470696507.ch29.

[20] Hu, Qiyue, Liu, Bin, Thomsen, Mads Rosendahl, Gao, Jianbo, and Nielbo, Kristoffer L. "Dynamic Evolution of Sentiments in Never Let Me Go: Insights from Multifractal Theory and its Implications for Literary Analysis". In: *Digital Scholarship in the Humanities 36*, no. 2 (2020), pp. 322–332. DOI: 10.1093/llc/fqz092.

[21] Hutto, Clayton and Gilbert, Eric. "Vader: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225. DOI: 10.1609/icwsm.v8i1.14550.

[22] Jacobsen, Mia, Bizzoni, Yuri, Moreira, Pascale Feldkamp, and Nielbo, Kristoffer L. "Patterns of Quality: Comparing Reader Reception Across Fanfiction and Commercially Published Literature". In: *Proceedings of the Computational Humanities Research Conference 2024*. Vol. 3834. 2024, pp. 718–739.

[23] Jacobsen, Mia and Kristensen-McLachlan, Ross Deans. "Admiration and Frustration: A Multidimensional Analysis of Fanfiction". In: *Proceedings of the Computational Humanities Research Conference 2024*. Aarhus, Denmark, 2024, pp. 93–112. URL: https://ceur-ws.org/Vol-3834/paper57.pdf.

[24] Jenkins, Henry. *Textual Poachers: Television Fans and Participatory Culture*. New York: Routledge, 1992.

[25] Jianbo Gao, Sultan, H., Jing Hu, and Wen-Wen Tung. "Denoising Nonlinear Time Series by Adaptive Filtering and Wavelet Shrinkage: A Comparison". In: *IEEE Signal Processing Letters 17*, no. 3 (2010), pp. 237–240. DOI: 10.1109/LSP.2009.2037773.

[26] Jockers, Matthew. "A Novel Method for Detecting Plot". Matthew L. Jockers Blog. 2014. URL: https://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/.

[27] Jockers, Matthew. "Syuzhet: Extract Sentiment and Plot Arcs from Text". 2015. URL: www.matthewjockers.net/2015/02/02/syuzhet/.

[28] Al-Laith, Ali, Degn, Kirstine, Conroy, Alexander, Pedersen, Bolette, Bjerring-Hansen, Jens, and Hershcovich, Daniel. “Sentiment Classification of Historical Danish and Norwegian Literary Texts”. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, ed. by Tanel Alumäe and Mark Fishel. Tórshavn, Faroe Islands: University of Tartu Library, 2023, pp. 324–334. URL: https://aclanthology.org/2023.nodalida-1.34/.

[29] Magnifico, Alecia Marie and Jones, Karis. “Theorizing Fanfiction: the Importance of Remixed Social Genres Composed on the Internet”. In: *Computers and Composition 75* (2025), p. 102916. DOI: 10.1016/j.compcom.2025.102916.

[30] Mattei, Andrea, Brunato, Dominique, and Dell’Orletta, Felice. “The Style of a Successful Story: a Computational Study on the Fanfiction Genre”. In: *Computational Linguistics CLiC-it 2020*. Bologna, Italy, 2020.

[31] Mendes, Gonçalo Azevedo and Martins, Bruno. “Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers”. In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*. Berlin, Heidelberg, 2023, pp. 84–100. DOI: 10.1007/978-3-031-28244-7_6.

[32] Mohammad, Saif. “Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 174–184. DOI: 10.18653/v1/P18-1017.

[33] Müller, Ulrich W. D., Witteman, Cilia LM, Spijker, Jan, and Alpers, Georg W. “All’s bad that ends bad: There is a Peak-end Memory Bias in Anxiety”. In: *Frontiers in Psychology 10* (2019), p. 1272. DOI: 10.3389/fpsyg.2019.01272.

[34] Neugarten, Julia. “Delicious Angst and Tooth-rotting Fluff: Distant Reading Community Discourses of Emotion in Harry Potter Fanfiction Comments”. In: *Journal of Fandom Studies 11*, no. 2 & 3 (2023), pp. 205–28. DOI: 10.1386/jfs_00082_1.

[35] Neugarten, Julia. “MythFic Metadata: Gendered Power Dynamics in Fanfiction about Greek Myth”. In: *Digital Humanities Benelux Journal 6* (2024), pp. 133–153.

[36] Neugarten, Julia. “Using Riveter to Map Gendered Power Dynamics in Hades/Persephone Fan Fiction”. In: *Transformative Works and Cultures 46* (2025). DOI: 10.3983/twc.2025.2643.

[37] Neugarten, Julia and Meijerink, Thijs. “Giddy Gods and Happy Heroes: Detecting Character-Emotions in Fanfiction about Greek Myth with Vector Space Models”. In: *Digital Humanities 2025: Accessibility & Citizenship*. Lisboa, Portugal, 2025.

[38] Nguyen, Duy, Zigmond, Stephen, Glassco, Samuel, Tran, Bach, and Giabbanelli, Philippe J. “Big Data Meets Storytelling: Using Machine Learning to Predict Popular Fanfiction”. In: *Social Network Analysis and Mining 14*, no. 1 (2024), p. 58. DOI: 10.1007/s13278-024-01224-x.

[39] Pugh, Sheenagh. *The Democratic Genre: Fan Fiction in a Literary Context*. Bridgend: Seren Books, 2005.

[40] Reagan, Andrew J., Mitchell, Lewis, Kiley, Dilan, Danforth, Christopher M., and Dodds, Peter Sheridan. “The Emotional Arcs of Stories Are Dominated by Six Basic Shapes”. In: *EPJ Data Science 5*, no. 1 (2016), pp. 1–12. DOI: 10.1140/epjds/s13688-016-0093-1. (Visited on 09/06/2022).

[41] Rebora, Simone. “Sentiment Analysis in Literary Studies. A Critical Survey”. In: *Digital Humanities Quarterly 17*, no. 2 (2023).

[42] Rousseeuw, Peter J. “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. In: *Journal of computational and applied mathematics 20* (1987), pp. 53–65. DOI: 10.1016/0377-0427(87)90125-7.

[43] Samutina, Natalia. “Emotional Landscapes of Reading: Fan Fiction in the Context of Contemporary Reading Practices”. en. In: *International Journal of Cultural Studies 20*, no. 3 (2016), pp. 253–269. DOI: 10.1177/1367877916628238.

[44] Sourati Hassan Zadeh, Zhivar, Sabri, Nazanin, Chamani, Houmaan, and Bahrak, Behnam. “Quantitative Analysis of Fanfictions’ Popularity”. In: *Social Network Analysis and Mining 12*, no. 1 (2022), p. 42. DOI: 10.1007/s13278-021-00854-9.

[45] Yang, Xiaoyan and Pianzola, Federico. “Exploring the Evolution of Gender Power Difference through the Omegaverse Trope on AO3 Fanfiction”. In: *Proceedings of the Computational Humanities Research Conference 2024*. 2024, pp. 906–916. URL: https://ceur-ws.org/Vol-3834/paper27.pdf.

A Sentiment Analysis

As noted, we compared 4 models for sentiment score retrieval: 2 dictionary-based models and 2 transformer-based models.[11] Models were selected based on previous studies, showing both the good performance of xlm-models and their improvement over monolingual (English) models [3; 13; 31].

These were tested against human ratings on a dataset of fiction – *Fiction4* [14] – and mixed nonfiction and fiction – *EmoBank* [9]. Each dataset contains sentences scored for valence by, in the case of *Fiction4* at least 2 annotators, and in the case of *EmoBank*, at least 10 annotators per sentence.

Models tried were:

Transformer-based ↓	`twitter-xlm` [1]
	`xlm-roberta` [1]
Dictionary-based ↓	`VADER` [21]
	`Syuzhet` [27]

For the full transformer-model names, see Table 9.

Result of the model comparisons can be found in Table 8. Note that transformers consistently outperform other tools, both on Danish and historical fiction, as well as other categories in the contemporary *EmoBank*. `xlm-roberta` shows a slightly better performance overall than `twitter-xlm`. Note also that `Syuzhet` appears better than `VADER` on the overall *EmoBank*, as well as its “Fiction” category. That is why we chose `Syuzhet` and `xlm-roberta` for the current study of fanfiction.

A.1 Transformer-based Sentiment

A.1.1 Continuous scale sentiment from transformer-predictions

As [3; 13] have shown, converting discrete predictions to continuous scores via model confidence values tends to outperform dictionary-based sentiment tools with continuous output when doing SA

[11] For the transformer-based scores, all code and data for testing is available here: https://anonymous.4open.science/r/literary_sentiment_benchmarking-D6E6

	Fiction4		EmoBank			
	Multilingual	**English texts**	**Overall**	**Blog**	**Newspaper**	**Fiction**
Publication dates	*1798-1965*	*1952-1965*	*1990-2008*	*1990-2008*	*1990-2008*	*1990-2008*
N. sentences	6,300	3,500	8,870	1,378	1,381	2,893
Human IRR $\rightarrow$	0.67	0.60	0.34	0.31	0.29	0.35
Models $\downarrow$						
`vader`	-	0.51	0.43	0.41	0.42	0.37
`syuzhet`	-	0.50	0.46	0.37	0.42	0.43
`twitter-xlm`	0.55	0.60	0.64	0.65	0.61	**0.57**
`xlm-roberta`	**0.60**	**0.61**	**0.65**	**0.65**	**0.65**	0.56

Table 8: Spearman correlations of sentiment models' scores with a human gold standard on the *Fiction4* (left) and *EmoBank* (right) datasets. At the top, information on each set (years & number of sentences), as well as Inter Rater Reliability (IRR)(measured in Krippendorff's α) per set. Columns from left to right: Overall evaluation on Multilingual (Danish and English) `Fiction4Sentiment` sentences ($n = 6,300$), evaluation of the exclusively English set of sentences ($n = 3,500$). Evaluation on overall *EmoBank* ($n = 8,870$) and 3 subgenres. The best model performance per Dataset setting is in bold. Note: All p-values < 0.01.

Type	**Shorthand, Modelname & URLs**
Shorthand	twitter-xlm
Name	`cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual`
URL	https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual
Shorthand	xlm-roberta
Name	`cardiffnlp/xlm-roberta-base-sentiment-multilingual`
URL	https://huggingface.co/cardiffnlp/xlm-roberta-base-sentiment-multilingual

Table 9: Full model names & details.

for literary texts. In this paper, following the same method, standard three-way transformer output (positive, neutral, negative) was mapped to a continuous scale by using the model's confidence as a proxy for intensity: a sentence classified as *positive* with a confidence of 0.67 is converted to +0.67, and similarly for *negative* predictions. *Neutral* outputs were assigned a value of 0, in line with the observation that most human annotations tend to cluster near the neutral midpoint (see [13]).

A.1.2 Truncation

To ensure compatibility with different transformer architectures, our script retrieved maximum input length dynamically from the tokenizer for the model via `tokenizer.model_max_length`.[12] For the `xlm-roberta` model, max input length was 514. Note that very few sentences were actually truncated, i.e., $< 0.02\%$, where the majority (>90%) were split into only 2 chunks.

For example, given the sentence:

```
This book is brilliant but the pacing is terrible.
```

If we imagine that this would exceed the token limit, it would be split into chunks of max length:

[12] Note, if this value is unrealistically high (e.g., $> 20,000$), it is capped the common standard of 514 tokens to avoid inefficient behavior.

- **Chunk 1**: `This book is brilliant`
- **Chunk 2**: `but the pacing is terrible.`

Each chunk is passed to the sentiment model. Suppose it returns:

- **Chunk 1**: `label = positive` & `confidence score = 0.9` $\rightarrow +0.9$
- **Chunk 2**: `label = negative` & `confidence score = 0.8` $\rightarrow -0.8$

The overall sentence score is then computed as the mean of the chunk scores:

$$\text{Sentiment} = \frac{(+0.9) + (-0.8)}{2} = 0.05$$

B Distribution of Clusters across Genres with Roberta

Genre	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Total
Angst	250 (23.7%)	230 (21.8%)	**300 (28.4%)**	275 (26.1%)	1,055
Fluff	270 (19.8%)	**390 (28.6%)**	320 (23.5%)	384 (28.2%)	1,364
Hurt/comfort	**140 (35.4%)**	80 (20.2%)	90 (22.7%)	86 (21.7%)	396
Other	2,100 (22.6%)	2,400 (25.8%)	**2,550 (27.4%)**	2,254 (24.2%)	9,304
Total	2,760 (22.8%)	3,100 (25.6%)	3,260 (26.9%)	2,999 (24.8%)	12,119

Table 10: Distribution of story genres across four narrative arc clusters with `Roberta` method, with substantial variation. Absolute counts and percentages (in parentheses) of each genre for that sentiment arc cluster. Highest percentage per genre indicated in bold.

C Model Outputs

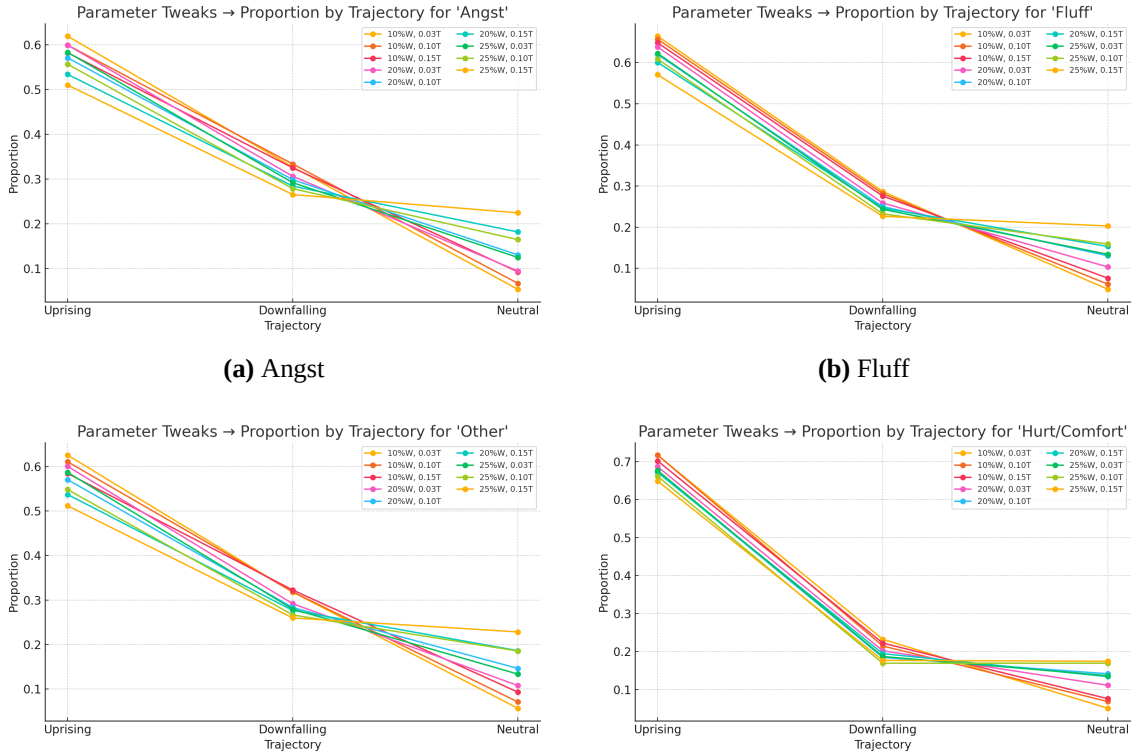
Mean `Syuzhet`		β	SE	t-value	p-value
	Angst	-0.062	0.01	-5.97	<0.001*
	Fluff	0.11	0.011	10.03	<0.001*
	Hurt/comfort	-0.022	0.0185	-1.17	0.24
	HP	0.00094	0.0053	.177	0.86
	LOTR	0.044	0.0056	7.91	<0.001*
	PJ	-0.0074	0.00559	-1.32	.19
	Published	0.0000012	0.0000012	0.97	0.33
	Word Count	-0.00012	0.0016	-0.074	0.94
	HP:Angst	-0.012	0.015	-0.77	0.44
	LOTR:Angst	0.010	0.016	0.66	0.51
	PJ:Angst	-0.032	0.015	-2.06	<0.05*
	HP:Fluff	-0.018	0.015	-1.25	0.21
	LOTR:Fluff	-0.0018	0.016	-0.11	0.91
	PJ:Fluff	-0.017	0.014	-1.26	0.21
	HP:Hurt/comfort	-0.017	0.027	-0.62	0.54
	LOTR:Hurt/comfort	-0.023	0.024	-0.967	0.33
	PJ:Hurt/comfort	-0.042	0.025	-1.65	0.099
Mean `xlm-roberta`		β	SE	t-value	p-value
	Angst	-0.041	0.0066	-6.28	<0.001*
	Fluff	0.0688	0.0068	10.12	<0.001*
	Hurt/comfort	-0.012	0.012	-1.08	0.28
	HP	-0.0053	0.0034	-1.56	0.11
	LOTR	0.0097	0.0035	2.73	<0.01*
	PJ	-0.0075	0.0035	-2.13	<0.05*
	Published	0.000001	0.00000076	1.32	0.19
	Word Count	-0.00049	0.001	-0.49	0.62
	HP:Angst	-0.011	0.0096	-1.17	0.24
	LOTR:Angst	0.0081	0.0099	0.82	0.41
	PJ:Angst	-0.029	0.0096	-3.02	<0.01*
	HP:Fluff	-0.0071	0.0092	-0.77	0.44
	LOTR:Fluff	-0.0065	0.010-	-0.64	0.52
	PJ:Fluff	-0.012	0.0086	-1.39	0.17
	HP:Hurt/comfort	-0.000085	0.017	-0.005	0.99
	LOTR:Hurt/comfort	-0.0087	0.015	-0.58	0.56
	PJ:Hurt/comfort	-0.018	0.016	-1.12	0.26

Table 11: All estimates for model (1) for each sentiment analysis method (`Syuzhet`/`xlm-roberta`)

SD Syuzhet		β	SE	t-value	p-value
	Angst	0.025	0.0090	2.77	<0.01*
	Fluff	-0.016	0.0092	-1.75	0.08
	Hurt/comfort	-0.0019	0.015	-0.12	0.90
	HP	-0.041	0.0051	-8.0	<0.001*
	LOTR	0.024	0.0055	4.38	<0.001*
	PJ	-0.074	0.0054	-13.68	<0.001*
	Published	0.0000018	0.0000011	1.55	0.12
	Word Count	0.0018	0.0014	1.25	0.21
	HP:Angst	-0.010	0.014	-0.74	0.46
	LOTR:Angst	-0.010	0.014	-0.74	0.46
	PJ:Angst	-0.016	0.013	-1.19	0.23
	HP:Fluff	-0.0046	0.013	-0.36	0.72
	LOTR:Fluff	0.011	0.014	0.80	0.42
	PJ:Fluff	-0.0035	0.012	-0.30	0.76
	HP:Hurt/comfort	-0.0041	0.024	-0.18	0.86
	LOTR:Hurt/comfort	0.0099	0.020	0.48	0.63
	PJ:Hurt/comfort	0.0062	0.022	0.29	0.78
SD xlm-roberta		β	SE	t-value	p-value
	Angst	0.0054	0.0027	2.0	<0.05*
	Fluff	0.0053	0.0028	1.89	0.059
	Hurt/comfort	0.0019	0.0048	0.41	0.68
	HP	-0.019	0.0014	-12.98	<0.001*
	LOTR	-0.010	0.0015	-6.77	<0.001*
	PJ	-0.018	0.0015	-11.65	<0.001*
	Published	0.00000097	0.0000032	3.03	<0.01*
	Word count	0.000090	0.00042	0.21	0.83
	HP:Angst	-0.0040	0.0040	-0.99	0.32
	LOTR:Angst	0.0020	0.0041	0.49	0.62
	PJ:Angst	-0.0049	0.0040	-1.24	0.21
	HP:Fluff	0.0020	0.0038	0.52	0.61
	LOTR:Fluff	0.0011	0.0042	0.27	0.79
	PJ:Fluff	-0.0040	0.0036	-1.13	0.26
	HP:Hurt/comfort	0.0021	0.0070	0.30	0.77
	LOTR:Hurt/comfort	-0.0025	0.0062	-0.40	0.69
	PJ:Hurt/comfort	0.0017	0.0066	0.26	0.80

Table 12: All estimates for model (2) for each sentiment analysis method (Syuzhet/xlm-roberta)

D Uprising versus Downfalling Arc Categories per Genre Across Parameters



(a) Angst **(b)** Fluff

(c) Other **(d)** Hurt/Comfort

Figure 4: Proportions of Uprising, Downfalling, and Neutral trajectories under different parameter settings for the four genres. Each line corresponds to one ⟨window%, threshold⟩ combination, showing how genre-level classification shares shift depending on the sampling window (first/last α of story) and the threshold T. As can be seen, changing the parameters does not change the percentage of classes, although in most cases accepting a too large window has the Neutral class rise (as larger windows will include different sections of the story, more likely to have important fluctuations: it is relatively difficult to have a constantly increasing/decreasing sentiment line).