Exploring Genre Preferences and Audience Engagement in Multilingual Fanfiction

A Study of Popularity and Preferences

Javina Qian Qian Ye[1]

Supervisor(s): Hayley Hung[1]**, Chenxu Hao**[1] **and Ivan Kondyurin**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Javina Qian Qian Ye
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Chenxu Hao, Ivan Kondyurin, Elmar Eisemann

Abstract

This study investigates how genre preferences and sentiment influence fanfiction popularity across multiple languages, focusing on English, Mandarin, Russian, and Spanish datasets. Leveraging advanced natural language processing techniques, including multilingual sentiment analysis, genre classification, and topic modeling, this research explores the interplay between cultural and linguistic factors in storytelling. Preprocessing steps, such as translation and named entity recognition, ensured consistency and reduced noise across the multilingual dataset. Key findings reveal cross-linguistic patterns, such as the popularity of genres like *Alternate Universe* and *Romance*, alongside cultural distinctions in sentiment and engagement. This work contributes to computational fan studies by demonstrating how linguistic and cultural factors influence storytelling trends and audience preferences in fanfiction.

1 Introduction

Fanfiction, a form of user-generated content where fans reinterpret or expand upon existing fictional universes, represents a rich and diverse dataset for computational analysis. Beyond its cultural significance as a participatory storytelling medium, fanfiction provides a unique opportunity to explore how language, sentiment, and genre preferences interact across linguistic and cultural boundaries. Despite its global appeal, most computational studies of fanfiction have focused predominantly on English-language works, leaving other languages and their associated cultural storytelling traditions underexplored [15, 3].

Language and culture significantly shape narrative conventions and preferences. While linguistic analysis has traditionally focused on formal texts, the informal and community-driven nature of fanfiction offers a less constrained but equally compelling window into cultural values and storytelling norms. For example, narrative structures in Western contexts often emphasize linear progression and individualism, whereas Eastern storytelling may prioritize cyclical narratives and collectivist themes [2]. These distinctions suggest that computational tools, such as natural language processing (NLP) and sentiment analysis, can reveal cross-linguistic patterns in how stories are told, received, and adapted.

This study applies computational techniques, including topic modeling, sentiment analysis, and genre classification, to investigate how genre preferences in fanfiction vary across languages. Genres such as *Angst*, *Fluff*, or *Alternate Universe* (AU) are inherently tied to specific emotional tones and reader expectations. Sentiment, which captures the emotional tone of a text, is therefore deeply intertwined with genre. For instance, *Angst* often conveys a negative or melancholic sentiment, while *Fluff* is typically associated with positive, comforting emotions. Understanding the sentiment profile of each genre can provide deeper insights into how readers engage with specific types of stories and what emotional experiences they seek within different cultural contexts.

Sentiment also plays a crucial role in audience engagement, as emotional resonance often determines a story's appeal. By examining sentiment distribution alongside engagement metrics—such as kudos, comments, and bookmarks—this study explores the interplay between emotional tone and audience response. For instance, fanfictions with complex or emotionally intense narratives may generate higher reader engagement in some linguistic communities, reflecting cultural attitudes toward vulnerability or emotional depth [21, 7]. The relationship between sentiment and genre also sheds light on how cultural norms influence storytelling preferences. Cultures that value emotional expression may favor genres with strong emotional contrasts, such as *Angst*, while others may gravitate toward lighter or plot-driven genres, like *Action/Adventure*.

In addition to its cultural and linguistic implications, this research highlights the computational challenges and opportunities of working with multilingual datasets. Translation pipelines, named entity recognition (NER), and thematic analysis are adapted to reduce noise, address the prevalence of character-specific terminology, and improve the consistency of results across languages. This preprocessing step ensures that genre extraction and sentiment analysis yield meaningful insights rather than being confounded by language-specific artifacts.

The main research question driving this study is: *How do genre preferences and audience engagement in fanfiction vary across languages and cultural contexts?* To address this, the study focuses on the following sub-questions:

1. How do genre preferences differ across languages, and which genres dominate within each language?
2. How do sentiment distributions vary across languages?
3. What thematic patterns emerge in fanfiction?
4. Which factors influence audience engagement in fanfiction across languages?

This study contributes to computational fan studies by presenting a multilingual framework that combines preprocessing, topic modeling, genre classification, and sentiment analysis to investigate how storytelling differs across linguistic and cultural contexts. The framework leverages NLP techniques, such as machine translation and named entity recognition (NER), to enable robust cross-linguistic comparisons. By analyzing large-scale fanfiction datasets from multiple languages, this study aims to provide insights into how genre preferences reflect cultural values, how sentiment shapes audience engagement, and how multilingual fanfiction data can be used to study narrative trends and reader reception. Through this exploration, the study advances the field of multilingual NLP and aims to provide a framework for understanding fanfiction as both a cultural artifact and a computational challenge.

2 Related Work

The computational analysis of fanfiction has emerged as a rich field of study, offering insights into narrative structures, cultural values, and audience engagement. This section reviews key contributions in genre and sentiment analysis, multilingual natural language processing (NLP), and computational tools relevant to fanfiction, situating the present study within this evolving landscape.

2.1 Genre and Sentiment Analysis in Fanfiction

Fanfiction presents unique opportunities to study user-generated narratives at scale. Sourati Hassan Zadeh et al. [17] conducted a quantitative analysis of fanfiction popularity, focusing on the relationship between story characteristics and reader engagement metrics. Their work highlighted the importance of extracting genres from stories, though their approach utilized traditional literary genres like romance and adventure. This research builds upon their approach by exploring fanfiction-specific genres, such as ”Angst” and ”Fluff,” which are inherently tied to emotional tones and unique to the medium, offering deeper insights into how these categories influence engagement across linguistic and cultural contexts.

Fanfiction presents unique opportunities to study user-generated narratives at scale. Kim and Klinger [10] reviewed sentiment and emotion analysis methods for computational literary studies, emphasizing the potential of lexicon-based and deep learning models to capture nuanced emotional tones in creative writing. Their findings inform the sentiment analysis framework used in this research, where multilingual fanfiction texts are translated to standardize sentiment scoring and ensure consistent comparisons across languages.

2.2 Multilingual and Cross-Cultural Studies

Multilingual analysis is crucial for understanding how linguistic and cultural contexts shape storytelling. The Stanford Literary Lab's *Multilingual Fanfic* project [18] explored fanfiction in multiple languages, revealing how cultural norms influence genre preferences and narrative structures. This project aligns with the current study's focus on cross-linguistic comparisons, specifically how genre and sentiment preferences differ across English, Spanish, Mandarin, and Russian fanfiction datasets.

Huang et al. [8] explored the capabilities of multilingual NLP models for analyzing diverse text datasets, highlighting the importance of cross-lingual embeddings in enabling linguistic comparability. This study similarly employs machine translation and multilingual embeddings to map fanfiction narratives into a shared linguistic space, facilitating robust cross-cultural analyses of genre and sentiment.

2.3 Tools and Methodologies for Fanfiction Analysis

Specialized tools have been developed to address the challenges of processing informal, community-driven texts like fanfiction. Yoder et al. [22] introduced FanfictionNLP, a pipeline for extracting key features such as character dynamics and narrative arcs. This tool underscores the importance of adapting NLP techniques to the unique characteristics of fanfiction. While FanfictionNLP is tailored to English texts, this study extends similar methods to multilingual data, incorporating preprocessing steps such as named entity recognition (NER) and thematic content filtering.

Thematic analysis also benefits from recent advancements in topic modeling. Neugarten [12] applied aspect-based sentiment analysis to fanfiction based on Greek mythology, illustrating how computational techniques can uncover nuanced reader evaluations of narrative elements. This study builds on Neugarten's methodology by using BERTopic to extract themes from multilingual fanfiction datasets, ensuring that cultural and linguistic variations are accounted for in topic modeling.

3 Methodology

This study investigates how genre preferences influence fanfiction popularity across different languages through a structured computational pipeline involving data collection, preprocessing, thematic analysis, genre classification, and sentiment analysis. The methodological steps are outlined as follows:

3.1 Data Collection

he fanfiction dataset for this study comprises multilingual texts collected from the Archive of Our Own (AO3) platform using the AO3Scraper tool developed by [13]. This tool facilitated the automated extraction of metadata, engagement metrics, and the full text of fanfiction entries. The dataset includes metadata fields such as the title, language, and engagement metrics (e.g., kudos, comments, bookmarks, and hits), as well as user-generated tags and the complete body text of each fanfiction.

The dataset spans four languages—English, Spanish, Mandarin, and Russian—and includes fanfictions from three major fandoms: Harry Potter, Marvel, and Good Omens. A balanced representation was prioritized across languages to ensure meaningful cross-linguistic comparisons. The dataset (currently) consists of 4997 English fanfictions, 3866 Spanish fanfictions, 4305 Mandarin fanfictions, and 3899 Russian fanfictions.

The user-generated tags associated with each fanfiction were used for genre classification and thematic analysis, as they capture key narrative elements, emotional tones, and character dynamics. This tagging system enabled a detailed examination of genre-specific patterns and their relationship to reader engagement across different languages and cultural contexts.

3.2 Data Preprocessing

Given the complexity and multilingual nature of the dataset, preprocessing was an essential step to ensure consistency and accuracy in downstream analyses. Data preprocessing involved cleaning and standardizing various columns. Missing data in key fields, such as the body and additional tags, was handled through imputation or removal of incomplete entries. Engagement metrics (kudos, comments, bookmarks, and hits) were converted from text to numeric values to simplify statistical analysis.

For genre classification, the `additional tags` column was mapped to a predefined list of consolidated genres. This mapping relied on keyword matching, with the addition of a translation step for non-English tags using the `MarianMT` model [5]. For example, a Spanish tag like *aventura* was translated to its English equivalent, *adventure*, before being categorized. This approach ensured a uniform genre representation across languages.

Named Entity Recognition (NER) using SpaCy's en_core_web_sm model [1] filtered out proper nouns (e.g., PERSON, ORG, and GPE) to anonymize text and focus on thematic content. A stopword list, combining standard stopwords from sklearn and custom additions such as common character names that were left despite using NER, was applied to filter uninformative words during thematic analysis.

Non-English texts were translated into English using Helsinki-NLP/opus-mt-mul-en [20], enhanced by pre-translation cleaning and post-translation deduplication of repeated phrases to improve readability. Batch processing was implemented for translation and NER to handle large datasets efficiently, ensuring high performance and scalability. The final preprocessed dataset was saved for downstream tasks, such as topic modeling and sentiment analysis, ensuring robust and consistent results across analyses.

3.3 Thematic Analysis

To explore recurring themes in fanfiction, the study employed BERTopic [6], a state-of-the-art topic modeling framework. BERTopic uses sentence embeddings, dimensionality reduction, and clustering to extract topics from text data. Sentence embeddings were generated using the sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 model [14], which supports multilingual text and ensures that fanfictions in different languages are represented in the same linguistic space.

Once embeddings were generated, the study used UMAP [11] to reduce the dimensionality of the data and HDBSCAN [4] to cluster fanfictions into topics. The extracted topics were refined by removing irrelevant terms such as character names and filler words. To enhance interpretability, topic labels were translated into English using the MarianMT model [5]. The results were visualized in two-dimensional space, providing a clear representation of thematic relationships across fanfictions.

3.4 Genre Classification and Sentiment Analysis

Translation also supported genre classification and sentiment analysis by standardizing linguistic expressions across languages. User-provided tags were mapped to predefined genres, with non-English tags translated into English to ensure consistent categorization.

To enable cross-linguistic comparisons, genre counts were calculated as proportions of the total fanfictions within each language dataset. This proportional approach ensures that differences in dataset sizes across languages do not overshadow the analysis. Additionally, normalization was applied to popularity metrics (e.g., kudos, comments, bookmarks, and hits) by calculating the average engagement per fanfiction within each genre. This allowed for a fair evaluation of smaller or niche genres (e.g., Time Travel) alongside larger, more common genres (e.g., Romance), ensuring that engagement trends were not biased by varying sample sizes. The use of proportional and normalized metrics highlights the relative performance and preference of each genre across linguistic and cultural contexts.

For sentiment analysis, VADER [9] was applied to English texts, while non-English fanfictions were translated prior to analysis to maintain consistent polarity scoring. Studies such as [19] corroborate the effectiveness of lexicon-based sentiment analysis on translated texts, highlighting its ability to capture nuanced emotional tones.

3.5 Quantitative Analysis

To analyze the factors influencing fanfiction popularity, variables such as genre and relationship category were examined for their impact on engagement metrics (hits, kudos, comments, and bookmarks). Weighted least squares (WLS) regression analysis [16] was applied to account for heteroscedasticity, ensuring more reliable parameter estimates by giving less frequent genres and relationship categories meaningful influence without being overshadowed by more frequent ones.

Regression coefficients estimate the effect of each genre or relationship category on engagement metrics, relative to a baseline category (e.g., a reference genre or no relationship category). For instance, a positive coefficient for the "Fluff" genre indicates higher engagement compared to the baseline genre, holding other variables constant. Similarly, coefficients for relationship categories (e.g., Male/Male or Female/Male relationships) reflect their specific contribution to engagement metrics.

Genres and relationship categories were treated as separate dimensions because they capture distinct aspects of fanfiction. **Genres** describe narrative themes (e.g., "Angst," "Alternate Universe"), while **relationship categories** represent the primary dynamics between characters (e.g., Male/Male, Female/Male). Separating these dimensions allows for a more granular analysis of how thematic elements and relational dynamics independently influence engagement metrics.

To prevent multi-genre fanfictions from inflating results, engagement metrics were normalized by dividing them by the number of genres associated with each story. Variance inflation factors (VIF) were calculated to ensure minimal multicollinearity among predictors. Separate regression models were fitted for English, Mandarin, Russian, and Spanish datasets to examine how genre and relationship category effects vary across linguistic and cultural contexts.

All analyses were implemented in Python using the statsmodels library for regression and pandas for data preprocessing. This approach ensures robust and interpretable findings, enabling cross-linguistic comparisons of how genres and relationship categories influence engagement.

4 Results

This section presents the results of genre distribution, sentiment analysis, engagement metrics, and thematic analysis across languages. It addresses the following research questions:

1. How do genre preferences differ across languages, and which genres dominate within each language?

2. How do sentiment distributions vary across languages?
3. What thematic patterns emerge in fanfiction?
4. Which factors influence audience engagement in fanfiction across languages?

4.1 Genre and Engagement Analysis

This section explores how genre preferences differ across languages and identifies dominant genres in each linguistic context, addressing Research Question 1. Additionally, it presents the results of sentiment analysis to address Research Question 2, focusing on the variation in sentiment distributions across linguistic contexts

Cross-Language Genre Distribution. Figure 1 illustrates the genre distribution across all datasets, highlighting both universal trends and linguistic nuances. Romance, Fluff, and Alternate Universe consistently dominate across languages, reflecting a shared global interest in relational and imaginative storytelling. However, genre preferences also vary by linguistic context.

English fanfiction exhibits a balanced distribution, with significant representation in relational genres (Romance, Fluff), emotional genres (Angst, Hurt/Comfort), and themes like Friendship and Family. Mandarin fanfiction leans toward Alternate Universe, Fluff, and Romance, with a notable emphasis on Humor and Hurt/Comfort, suggesting lighter and reflective narratives. Russian fanfiction stands out for its focus on Romance and Humor, alongside significant representation of Alternate Universe and Drama. Spanish fanfiction mirrors Russian trends, with Romance dominating, followed by Alternate Universe and Angst, indicating a preference for relational and emotionally intense storytelling.

While genres like Horror and Science Fiction are underrepresented across all languages, Action/Adventure sees moderate engagement in English and Russian. These results underscore both the universal appeal of relational and emotionally engaging genres and the unique storytelling preferences within each language.

Audience Engagement by Genre. Figure 3 shows the normalized kudos for key genres across languages. Time Travel consistently garners the highest engagement. Spanish datasets highlight high engagement for Action/Adventure and Fantasy, while Mandarin datasets feature strong engagement for Family and Mystery. Despite their high frequency, genres like Romance and Fluff generally yield lower kudos, suggesting broader but diluted recognition.

Sentiment Analysis. Sentiment distributions across languages are visualized in Figure 2. English fanfiction skews positive, with 4-star ratings being the most frequent. In contrast, Mandarin, Russian, and Spanish datasets lean toward neutral or mildly negative sentiments, with 2-star ratings most common. Despite these trends, positive genres like Fluff remain prevalent, reflecting nuanced engagement with bittersweet or emotionally complex narratives.

4.2 Topic Extraction Results

This subsection examines the thematic patterns emerging in fanfiction, answering Research Question 3 through a detailed analysis of the key topics in each dataset.

Topic modeling was conducted in the fanfiction datasets to explore thematic patterns. For each dataset, coherent topics were identified along with a group of outliers that did not align with the dominant themes. Tables 1, 2, 3 and 4 summarize the key topics, their representative keywords, and the number of fanfictions associated with each. The corresponding scatter plot visualization of these topics has been included in the appendix (see Appendix figures 4, 5, 6, and 7).

Table 1: Summary of Topics with Keywords (English)

Topic ID	Keywords (Top Terms)
0	*career, cheap, liquid*
1	*sixth, casting, delicate*
2	*just, time, know*
3	*think, time, know*
4	*interns, badges, labs*
5	*think, time, know*
6	*think, know, time*
7	*fate, delicate, savior*
8	*delicate, strip, noon*
9	*think, know, just*
10	*delicate, realm, liquid*

Table 2: Summary of Topics with Keywords (Mandarin)

Topic ID	Keywords (Top Terms)
0	*weather, snow, road*
1	*life, humans, evil*
2	*students, life, legs*
3	*life, marriage, child*
4	*legs, throat, lips*
5	*boyfriend, nipples, talking*
6	*read, company, reading*
7	*dream, wakes, woke*
8	*shop, coffee, coffee shop*

Table 3: Summary of Topics with Keywords (Spanish)

Topic ID	Keywords (Top Terms)
0	*close, clothes, tired*
1	*order, stars, fear*
2	*rights, omegaverse, base*
3	*family, boyfriend, born*
4	*book, drink, dinner*

Key Findings:

- **English Dataset:**
  - **Topic 2 (*just, time, know*):** Narratives with reflective and introspective tones.
  - **Topic 4 (*interns, badges, labs*):** Stories set in workplace or institutional environments.

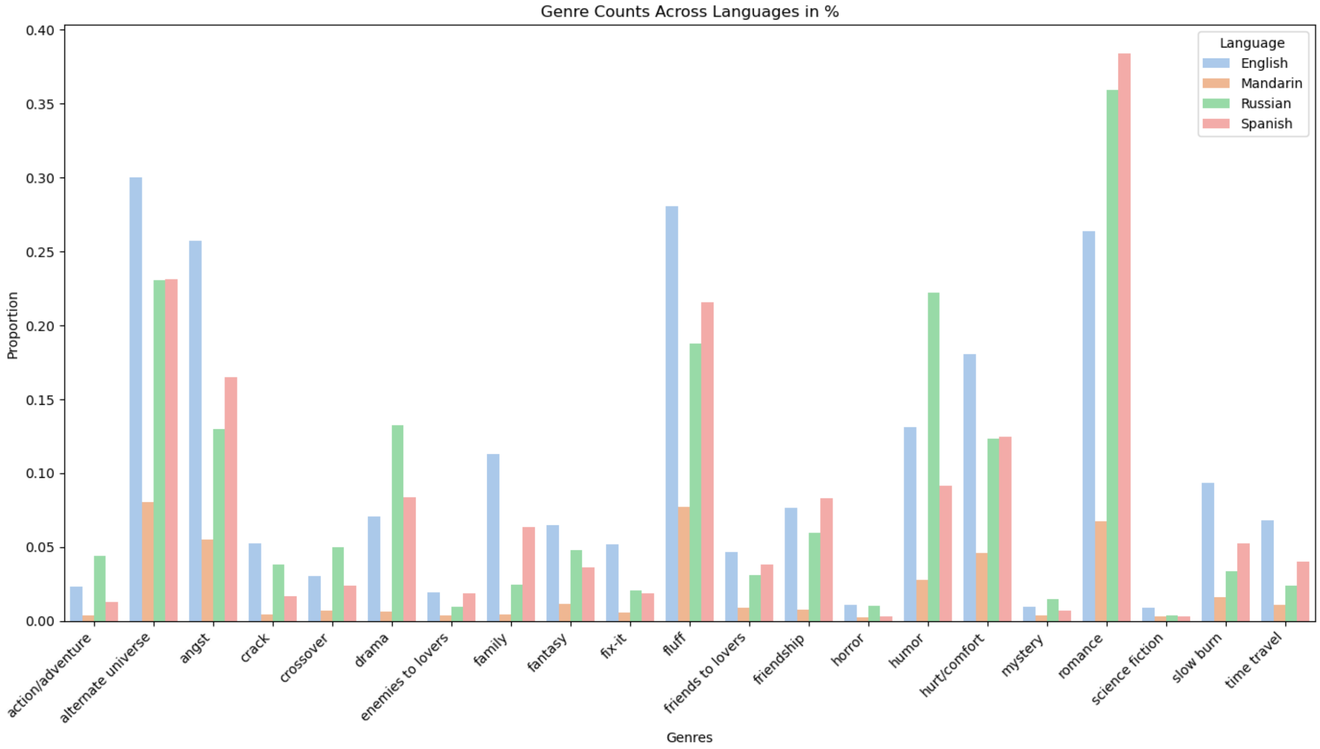


Figure 1: Proportional genre distribution across languages.

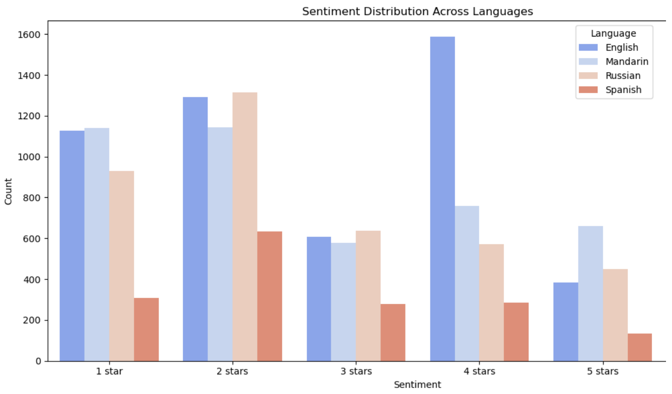


Figure 2: Sentiment distribution across languages.

- **Topic 7 (*fate, delicate, savior*):** Themes of heroism and destiny, potentially involving transformative character arcs or significant events.

- **Mandarin Dataset:**
  - **Topic 3 (*life, marriage, child*):** Narratives exploring family dynamics, love, and significant life events, such as parenthood or partnerships.
  - **Topic 5 (*boyfriend, nipples, talking*):** Stories emphasizing interpersonal relationships and potentially romantic or explicit conversations.
  - **Topic 8 (*shop, coffee, coffee shop*):** A representation of the ”coffee shop AU” trope, focusing on everyday social interactions in casual settings.

Table 4: Summary of Topics with Keywords (Russian)

Topic ID	Keywords (Top Terms)
0	*rain, soho, cottage*
1	*students, education, summer*
2	*stars, power, shadows*
3	*kisses, soft, wake*
4	*family, dating, pick*
5	*glasses, smell, decades*
6	*gets number, awkward questions, bitch*

- **Spanish Dataset:**
  - **Topic 2 (*rights, omegaverse, base*):** This topic highlights the significance of the Omegaverse subgenre, showing its strong thematic focus on rights and familial roles, a popular fanfiction-specific trope.
  - **Topic 3 (*family, boyfriend, born*):** Family-centered narratives are a recurring theme, with stories exploring relationships, life milestones, and generational dynamics, aligning with cultural emphasis on familial connections.
  - **Topic 4 (*book, drink, dinner*):** Leisure-focused narratives reflect a significant subset of stories centered on casual, intimate settings, indicating a pref-

erence for relatable and grounded experiences.

- **Russian Dataset:**
  - **Topic 1 (*students, education, summer*):** Stories highlighting academic life and personal growth, often set in school or university contexts.
  - **Topic 3 (*kisses, soft, wake*):** Romantic and intimate narratives, emphasizing emotional and physical connections between characters.
  - **Topic 4 (*family, dating, pick*):** Narratives focused on family relationships and dating, often exploring themes of love, humor, and personal decisions.

Outlier groups were observed in all datasets, comprising fanfictions that did not cluster into specific thematic groups. These outliers reflect diverse narrative styles or ambiguous language and are not the focus of this analysis.

4.3 Cross-Linguistic Regression Analysis Results

The regression analysis provides insight into the factors influencing audience engagement across languages, addressing Research Question 4 by identifying significant predictors like genres and relationship categories.

The coefficient comparison across languages (see Appendix Figure 8) highlights the influence of genres and relationship categories on audience engagement, measured through normalized kudos. Regression models for each language yielded adjusted R-squared values ranging from 0.05 (English) to 0.15 (Russian and Spanish), suggesting that while some variance in engagement is explained by genres and categories, audience preferences remain complex and influenced by additional factors.

Genres and Categories as Predictors. In this study, genres refer to thematic tags such as *Time Travel* or *Alternate Universe*, while categories primarily capture relationship dynamics (e.g., *Male/Male (M/M)*, *Female/Male (F/M)*). These dimensions were analyzed separately to examine their unique contributions to audience engagement. Genres often reflect narrative themes or plot structures, while categories address relational focus, making both valuable but distinct predictors of audience preferences.

Key Findings:

- **Universal Patterns:**
  - Certain genres, such as *Time Travel* and *Alternate Universe*, consistently show strong positive coefficients across multiple languages, indicating their widespread appeal.
  - For instance, *Time Travel* is highly significant in all datasets, particularly in Spanish, suggesting that speculative and plot-driven narratives resonate broadly with diverse audiences.
- **Cross-Linguistic Differences:**
  - In the English dataset, genres like *Crossover*, *Humor*, and *Alternate Universe* are positively associated with engagement, reflecting a preference for imaginative and comedic narratives.
  - The Russian dataset reveals negative coefficients for emotional genres such as *Angst* and *Fluff*, suggesting a less favorable reception of intense or comforting themes.
  - Mandarin fanfictions demonstrate strong positive effects for relationship categories like *M/M* and *F/M*, whereas *Romance* has a negative coefficient, potentially indicating cultural nuances in the portrayal of romantic themes.
- **Relationship Categories:**
  - Categories such as *M/M* and *F/M* consistently exhibit positive coefficients in the Mandarin, Russian, and Spanish datasets, highlighting the audience's engagement with relationship-focused narratives.
  - The *General (Gen)* category, however, shows weaker or non-significant effects, suggesting that stories without a clear relational focus may attract less engagement.

Multicollinearity and Model Limitations. The variance inflation factor (VIF) analysis for multicollinearity flagged *Romance* in the Mandarin dataset as a potential concern (VIF >10). This suggests overlapping relationships with other genres or categories, which may affect interpretability. Future models could address this by further feature engineering or excluding highly collinear predictors.

Interpretation of Findings. The results underscore the interplay between thematic genres and relational categories in shaping audience engagement. Genres like *Time Travel* or *Alternate Universe*, which often involve speculative or reimagined narratives, appear to have a universal appeal. Meanwhile, relationship categories highlight the importance of interpersonal dynamics, though their influence varies across cultural contexts.

Table 5 summarizes the key regression statistics for each dataset, while Appendix Figure 8 provides a visual comparison of significant coefficients across languages.

5 Discussion

5.1 Use of User-Generated Genres

This study utilized user-generated genres rather than training a model to extract them due to the complexity and subjectivity associated with certain genres. For instance, genres like *Alternate Universe* are highly nuanced and can vary significantly in interpretation, making automated classification prone to inaccuracies. Preliminary attempts at training a genre classification model yielded inconsistent results. User-generated genres provide a more reliable foundation, as they reflect community consensus on narrative themes, ensuring that genre assignments align with readers' expectations and cultural norms.

5.2 Translation for Thematic Analysis

This study translated non-English fanfiction texts into English before performing thematic analysis to ensure consistency and comparability across languages. Constructing stopword lists to remove character names and fan-specific terms

Table 5: Summary of Regression Statistics Across Languages

Language	Observations	Adjusted R-squared	F-statistic (p-value)	Significant Features
English	7939	0.052	19.98 ($p < 0.001$)	Time Travel, Crossover, Humor, M/M
Mandarin	1607	0.065	5.84 ($p < 0.001$)	F/M, M/M, Alternate Universe, Fix-it
Russian	4437	0.146	34.00 ($p < 0.001$)	Time Travel, Angst (negative), Fluff (negative)
Spanish	2010	0.143	16.22 ($p < 0.001$)	Time Travel, Action/Adventure, F/M

in multiple languages would have been inefficient and error-prone, particularly for languages in which the researcher is not proficient. Translation allowed the use of a single, custom stopword list applied uniformly across all texts.

Analyzing texts in a single language also ensured that sentence embeddings for topic modeling were generated by the same pre-trained model, reducing semantic inconsistencies. Although translation can introduce minor distortions, the high-quality Helsinki-NLP Opus-MT models minimized these risks, preserving the core narrative structure. This approach enabled robust cross-linguistic comparisons, free from language-specific artifacts, while maintaining focus on narrative trends and reader engagement across cultural contexts.

5.3 Choice of Engagement Metric

This study uses kudos as the primary engagement metric because it directly reflects reader appreciation with minimal ambiguity. Unlike hits, which can count repeat views and accidental clicks, kudos indicate intentional positive feedback from unique users. Bookmarks, while also meaningful, may reflect the intent to read later rather than immediate approval. By focusing on kudos, the analysis captures a clearer sign of audience engagement and story appeal.

5.4 Using Linear Regression

Linear regression was chosen to analyze the influence of genres on engagement metrics as it allows for simultaneous estimation of the effect of each genre relative to a reference category. This method provides effect sizes and avoids multiple pairwise comparisons required by methods like ANOVA.

Unlike ANOVA, which only detects differences in means across groups, regression offers a more detailed understanding of the relationship between genres and engagement metrics while accounting for potential covariates. However, this approach assumes linearity and independence, which may not fully capture more complex relationships in the data.

5.5 Model Limitations and Interpretability

The regression model for normalized kudos shows a relatively low R^2 (5.5%), indicating that much of the variance in engagement remains unexplained. However, this is expected in studies involving user-generated content, where engagement is influenced by many external factors, such as author popularity, timing, and community trends, which were not included in the model.

Despite this, the regression coefficients offer valuable insights into the relative impact of genres and categories on reader engagement. The model's purpose is not precise prediction but to highlight significant patterns in engagement across different genres. Future research could improve explanatory power by including additional variables, such as fandom-specific data or publication timing.

6 Responsible Research

This study adheres to ethical research and reproducibility principles. The AO3Scraper tool was used for data collection with an appropriate header identifying the automated requests and their purpose, ensuring compliance with ethical web scraping practices. Only publicly available metadata and text were collected, with no attempts to deanonymize authors or readers, preserving user anonymity. Additionally, care was taken to avoid excessive requests to the Archive of Our Own (AO3) platform to minimize disruption to its operations.

6.1 Design Choices and Limitations

The choice of languages—English, Spanish, Mandarin, and Russian—was guided by a balance between the availability of fanfiction data in these languages on AO3 and the representativeness of different linguistic and cultural contexts. However, this selection introduces limitations, as other significant linguistic groups (e.g., Arabic, French, or Japanese) are not included. Future studies could expand the analysis to additional languages to provide a more comprehensive cross-linguistic perspective.

Cross-lingual representation was addressed using multilingual tools such as MarianMT for translation and BERTopic with sentence-transformer embeddings to project texts into a shared linguistic space. While these tools facilitate comparisons across languages, they may introduce biases related to translation quality and embedding alignment, particularly for languages with less robust computational resources. These biases could affect the accuracy of genre classification and sentiment analysis. To mitigate this, pre- and post-translation cleaning steps were implemented, and multilingual models with strong performance across diverse languages were chosen.

6.2 Reproducibility and Transparency

All analysis steps, including sentiment analysis, topic modeling, and regression, were implemented in Python with reproducible workflows. Key scripts, such as those for genre extraction, translation, and regression weighting, were version-controlled using Git to ensure traceability. The datasets were preprocessed using documented settings, and all code dependencies are specified in a requirements file to facilitate replication.

The study acknowledges that some preprocessing steps, such as the removal of named entities or the mapping of non-

English tags to English genres, involve subjective decisions that may impact reproducibility. To enhance transparency, these mappings and preprocessing configurations are documented and will be made publicly available upon publication, allowing future researchers to evaluate and replicate the methodology.

6.3 Ethical Considerations

While this study relies solely on publicly available data, the sensitive and participatory nature of fanfiction necessitates additional ethical considerations. Fanfiction represents creative work shared within a community, and its analysis must respect the intentions and privacy of its authors. By anonymizing text, excluding identifying information, and refraining from sharing raw datasets, the study ensures that the focus remains on aggregate trends rather than individual works.

7 Conclusions and Future Work

This study analyzed fanfiction datasets in English, Mandarin, Russian, and Spanish to uncover universal trends and language-specific variations in engagement, sentiment, and thematic preferences. While genres like *Romance*, *Fluff*, and *Alternate Universe* are consistently popular across languages, plot-driven genres such as *Time Travel* stand out for their higher engagement. Cultural nuances were evident, with Spanish fanfiction favoring *Action/Adventure*, Mandarin emphasizing *Family*, and Russian prioritizing *Humor*.

Sentiment analysis revealed a more positive tone in English fanfiction, contrasting with the neutral-to-bittersweet sentiment observed in the other languages. Topic modeling further highlighted thematic distinctions, such as family narratives and casual social settings in Mandarin, Omegaverse tropes in Spanish, and academic life and romance in Russian, reflecting the interplay between cultural contexts and storytelling.

Regression analysis identified *Time Travel* and relational categories (e.g., M/M, F/M) as strong predictors of engagement, particularly in non-English datasets, while demonstrating that genres and categories alone offer limited explanatory power. Factors like fandom-specific trends and writing quality likely play significant roles in shaping audience preferences, emphasizing the complexity of fanfiction engagement.

Future research could explore additional languages, such as French, German, or Japanese, to provide a more comprehensive understanding of global fanfiction trends. Incorporating other predictors, such as character dynamics, temporal trends, narrative complexity, or the impact of crossover fandoms, could yield deeper insights. Further exploration of sentiment progression within fanfictions, such as shifts between positive and negative tones, could illuminate emotional engagement patterns.

Additionally, studying reader feedback mechanisms, such as the role of reviews and comments in influencing popularity, could provide a richer understanding of audience interaction. Furthermore, examining temporal trends, including how genre preferences evolve over time, or analyzing fanfiction in response to major events in canon works, could offer a new perspective on fan engagement.

References

[1] spacy: Industrial-strength natural language processing in python. https://spacy.io.

[2] A. M. Aiswarya. Fanfiction as an academic tool for advanced language fluency: A study. *Turkish Journal of Computer and Mathematics Education*, 12(4):367–372, 2021.

[3] R. W. Black. Language, culture, and identity in online fanfiction. *E-Learning and Digital Media*, 3(2):170–184, 2006.

[4] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Proceedings of PAKDD*, pages 160–172, 2013.

[5] Hugging Face. Marianmt: Pretrained models for machine translation, 2020. https://huggingface.co/transformers/model_doc/marian.html.

[6] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint*, 2022.

[7] S. B. Heath. Taking a cross-cultural look at narratives. *Topics in Language Disorders*, 7(1):84–94, 1986.

[8] K. Huang, F. Mo, and others. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*, 2024.

[9] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[10] E. Kim and R. Klinger. A survey on sentiment and emotion analysis for computational literary studies. *Journal of Digital Humanities*, 4(1), 2019.

[11] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint*, 2018.

[12] J. Neugarten, T Dejaeghere, et al. Catching feelings: Aspect-based sentiment analysis for fanfiction comments about greek myth, 2024.

[13] Radiolarian. Ao3scraper: A tool to scrape metadata and content from archive of our own, 2021. https://github.com/radiolarian/AO3Scraper.

[14] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of EMNLP-IJCNLP*, 2019.

[15] S. Sauro and B. Sundmark. An unexpected journey: Using fanfiction to foster second language development and literary competence. In *Paper presented at the annual conference of the American Association for Applied Linguistics*, Toronto, Canada, March 23 2015.

[16] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, pages 57–61, 2010.

[17] Z. Sourati Hassan Zadeh, N. Sabri, et al. Quantitative analysis of fanfictions' popularity. *Social Network Analysis and Mining*, 12(1):42, 2022. Available at https://doi.org/10.1007/s13278-021-00854-9.

[18] Stanford Literary Lab. Multilingual fanfic, 2019. Available at https://litlab.stanford.edu/projects/multilingual-fanfic/.

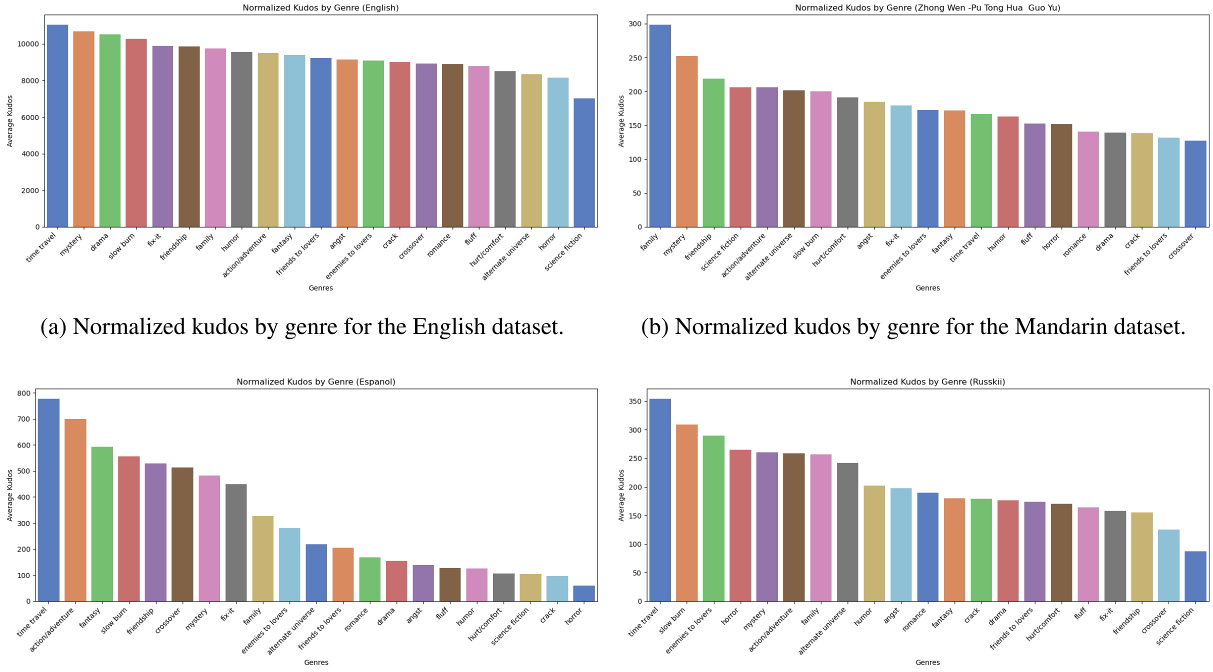
[19] M. Taboada, J. Brooke, et al. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011. [Online]. Available: https://aclanthology.org/J11-2001.

[20] J. Tiedemann and S. Thottingal. Opus-mt – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, 2020.

[21] J. L. Tsai, F. F. Miao, et al. Influence and adjustment goals: Sources of cultural differences in ideal affect. *Journal of Personality and Social Psychology*, 92(6):1102–1117, 2007.

[22] M. Yoder, S. Khosla, and others. Fanfictionnlp: A text processing pipeline for fanfiction. In *Proceedings of the Third Workshop on Narrative Understanding*, 2021.

Appendix



(a) Normalized kudos by genre for the English dataset.

(b) Normalized kudos by genre for the Mandarin dataset.

(c) Normalized kudos by genre for the Spanish dataset.

(d) Normalized kudos by genre for the Russian dataset.

Figure 3: Comparison of genre counts in different language datasets.

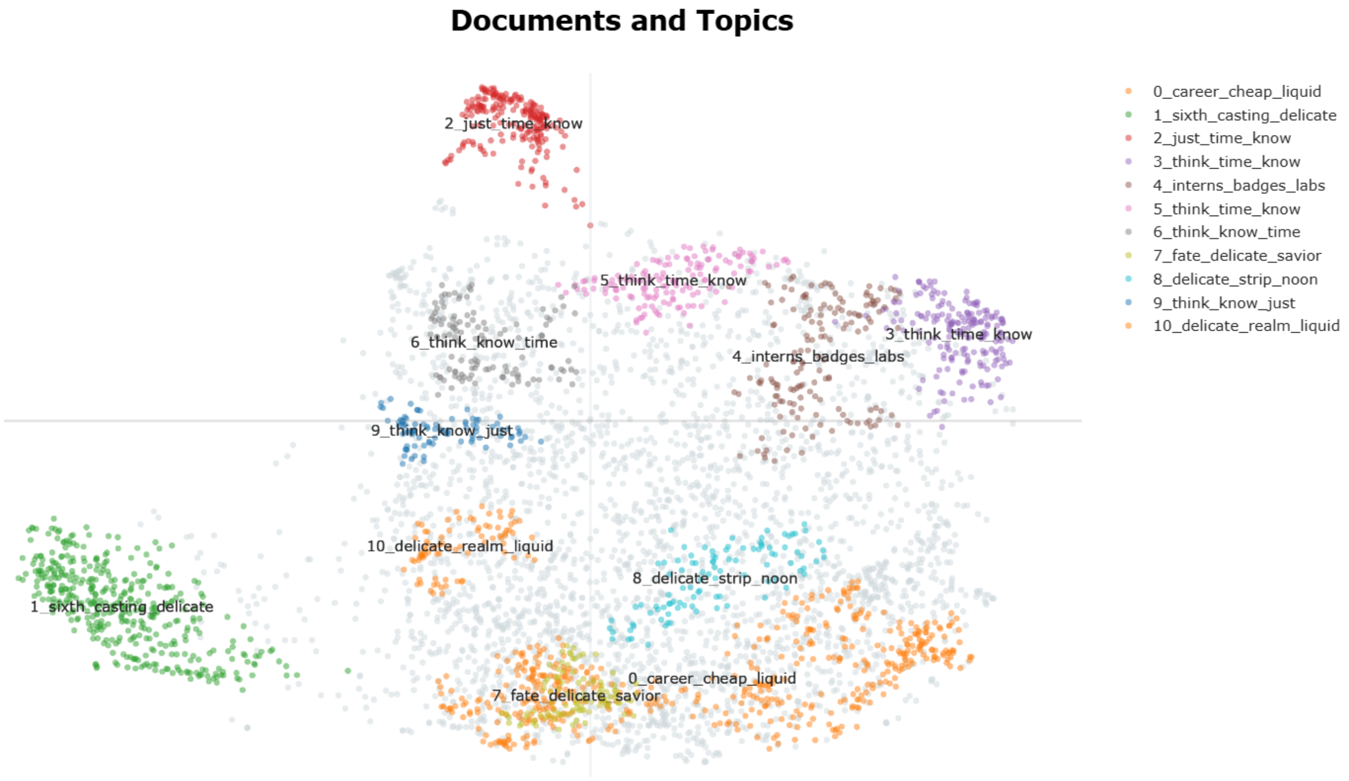


Figure 4: Scatter plot of fanfictions by topic distribution in the English dataset. Each point represents a fanfiction, with color-coding indicating its corresponding topic.

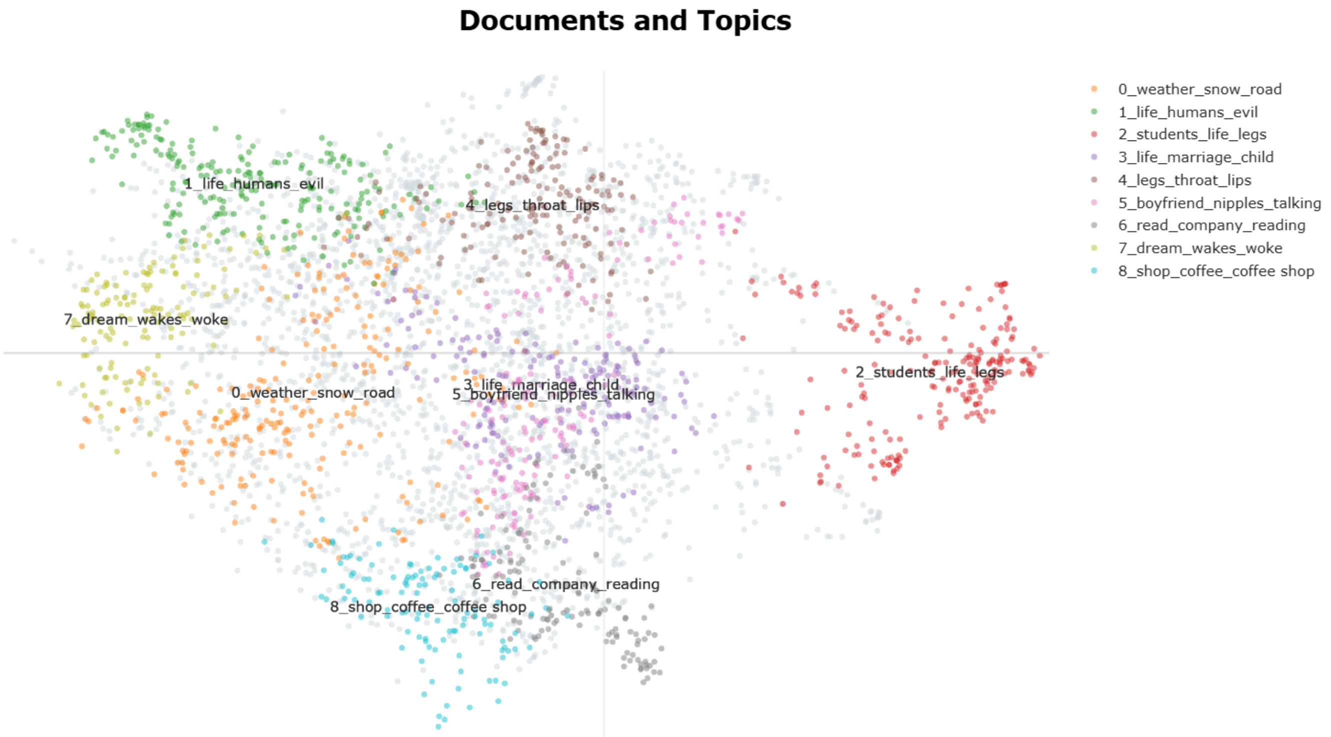


Figure 5: Scatter plot of fanfictions by topic distribution in the Mandarin dataset. Each point represents a fanfiction, with color-coding indicating its corresponding topic.

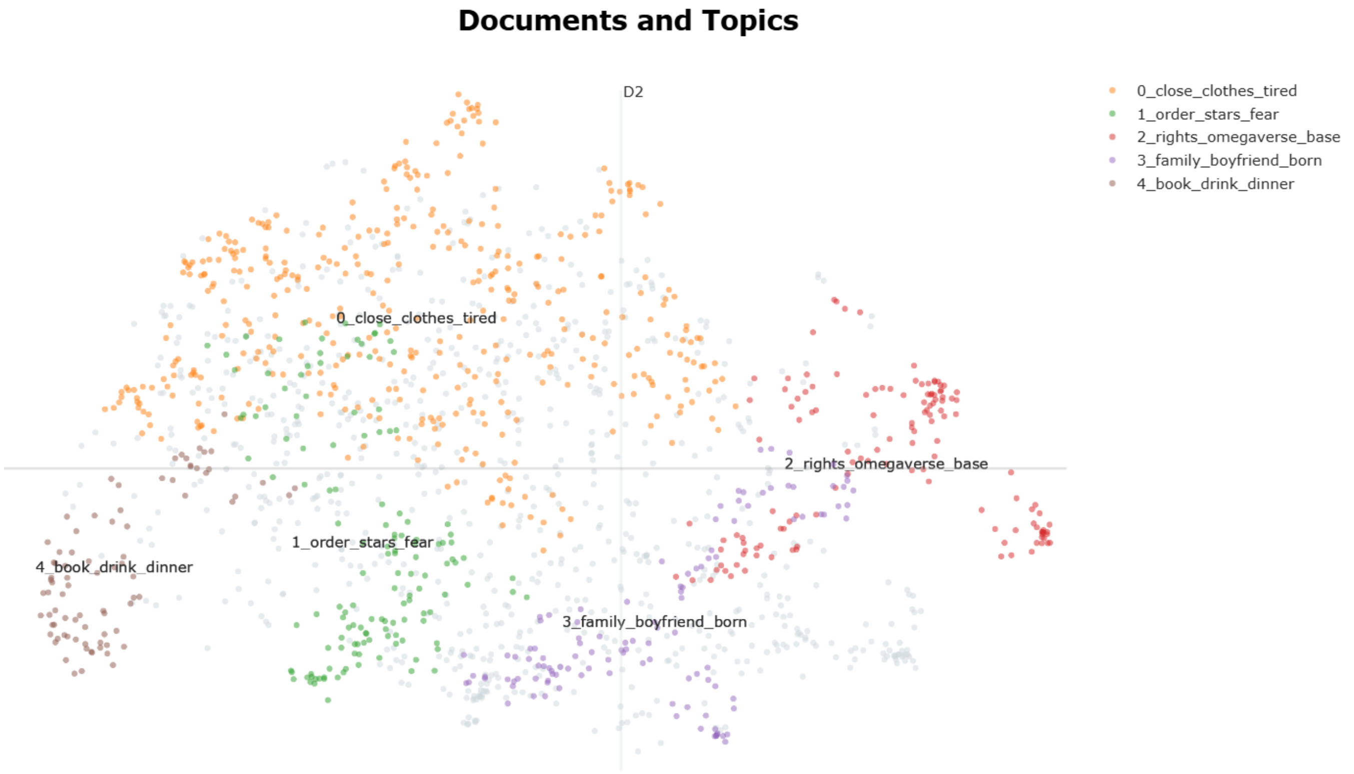


Figure 6: Scatter plot of fanfictions by topic distribution in the Spanish dataset. Each point represents a fanfiction, with color-coding indicating its corresponding topic.

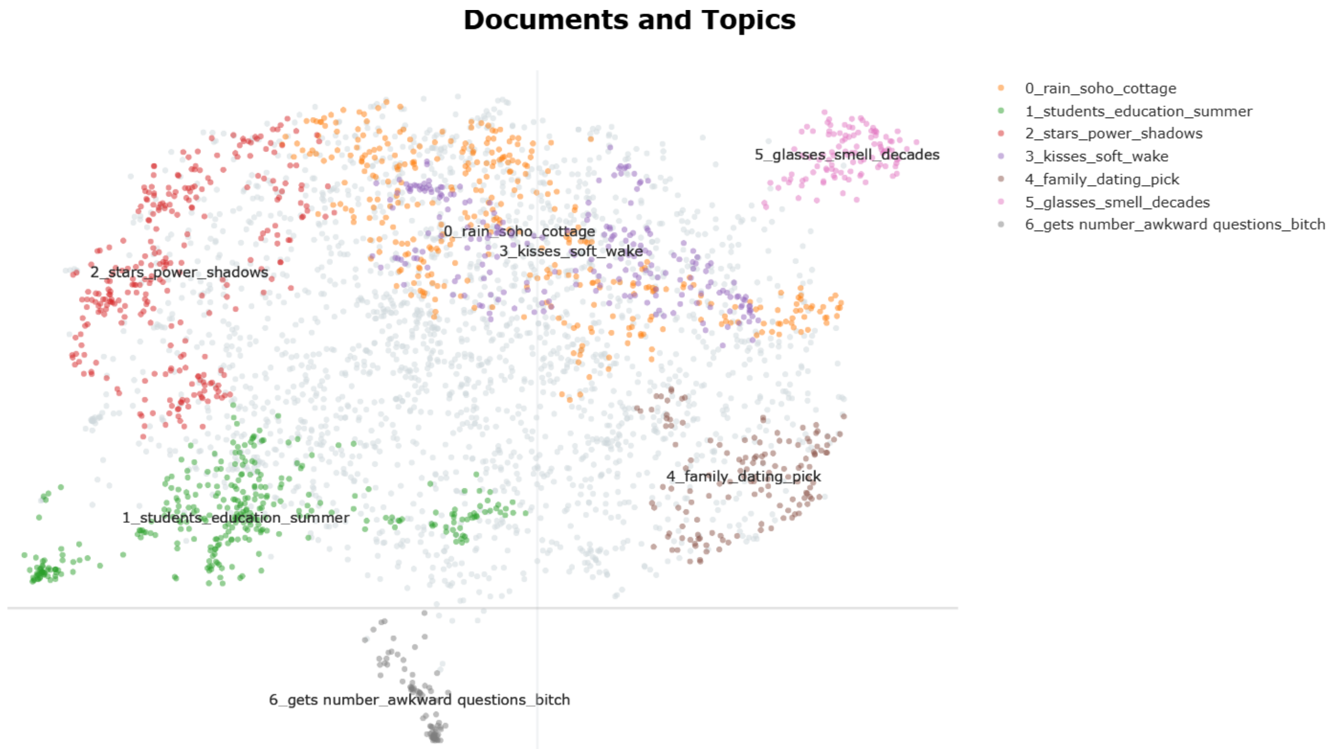


Figure 7: Scatter plot of fanfictions by topic distribution in the Russian dataset. Each point represents a fanfiction, with color-coding indicating its corresponding topic.

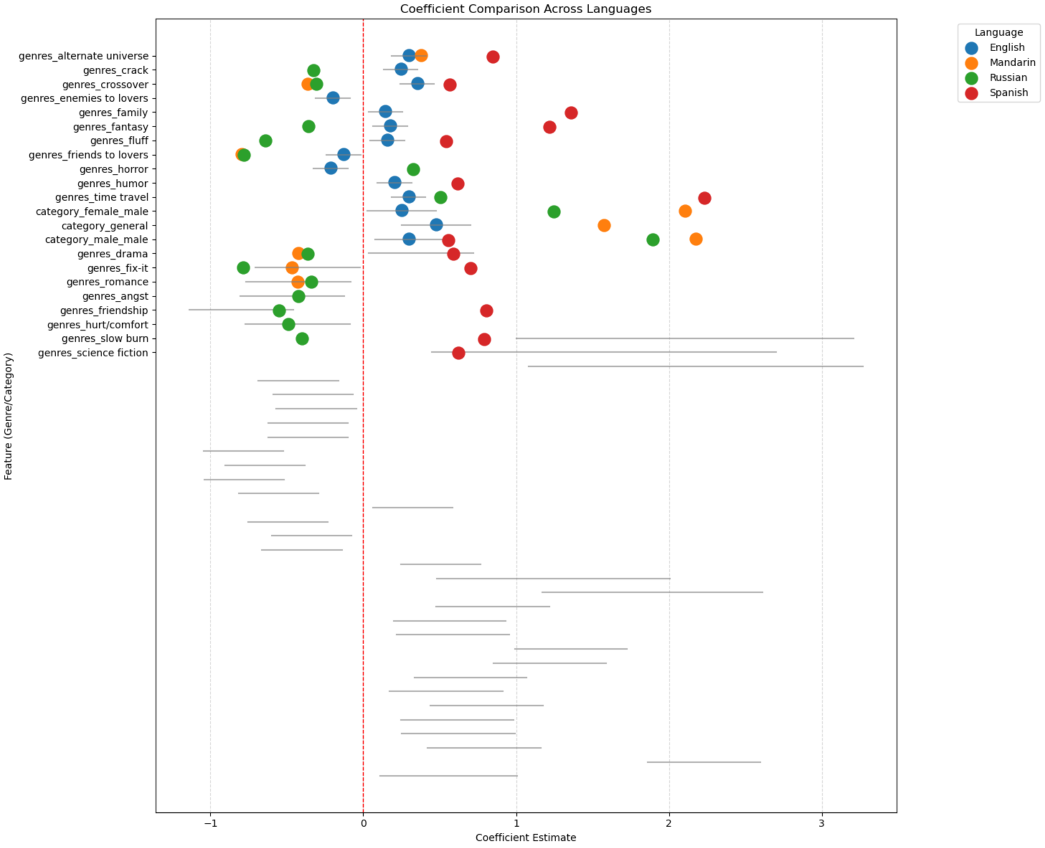


Figure 8: Significant Coefficient Comparison Across Languages. The dots represent estimated coefficients for genres and categories, with horizontal lines indicating confidence intervals. Only significant features (confidence intervals that do not cross zero) are included.