Fanfiction & Feminism & AI: Unveiling Sociopolitical Reflections in Character Portrayals

Author: Irina-Ioana Marinescu (i.i.marinescu@student.tudelft.nl)
Responsible Professor: Hayley Hung
Examiner: Elmar Eisemann
Supervisors: Chenxu Hao, Ivan Kondyurin

Fandoms Browse Search About Search

Introduction

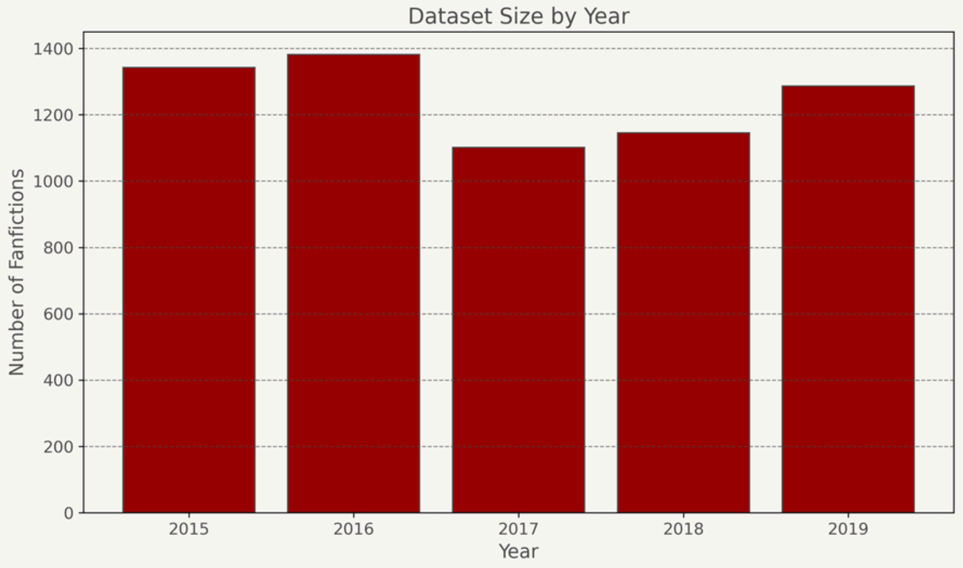
Fanfiction is a form of media that is derived from previously existing canon work, such as books, movies, and popular culture, and is created by fans with the purpose of exploring alternative narratives. The Internet, specifically websites like AO3 and Fanfiction.net, represents the space where fanfictions are created and shared, making this hobby accessible to the large public and, therefore being an honest reflection of the beliefs and concerns that exist in the online world en masse.

This paper will focus on answering how the portrayal of female characters in fanfiction evolved in response to the #MeToo movement and fourth-wave feminism, as analyzed with the help of NLP techniques. The final objective is to understand whether the impact of the awareness campaign was broad enough and had visible consequences in the way the average author portrays women in narration. By achieving this, the paper would bring new insights into NLP research, showing ways of how computational methods can analyze modern cultural expression. Moreover, this study contributes to ongoing discussions about gender and power dynamics in media.

Dataset

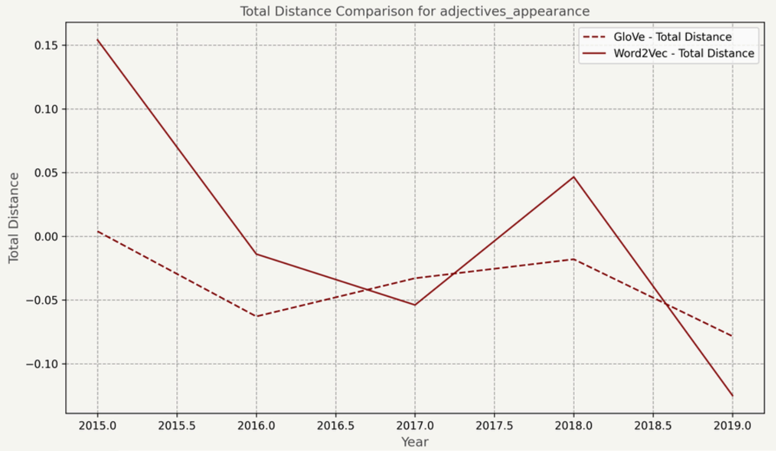
The full dataset contains **6255** fanfictions from the Twilight, Hunger Games, and Good Omens canons posted in the years 2015-2019, as to include enough samples for trend determination around the #MeToo movement. The data was scrapped using the AO3Scraper repository, which consists of a simple and convenient Python script that retrieves the metadata and content of fanfictions that match a given query.

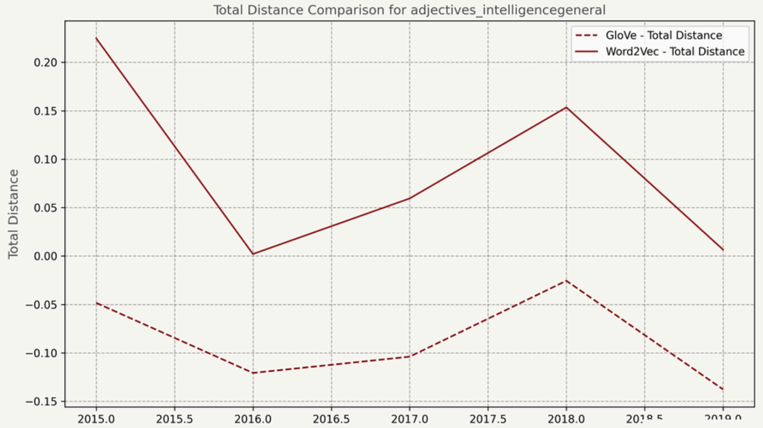
For the Word2Vec and GloVe pipeline, 1101 fanfictions were used for each year, as to avoid an additional need for normalization of the results. On the other hand for the BERT pipeline the full available dataset was included in the calculations.



Results

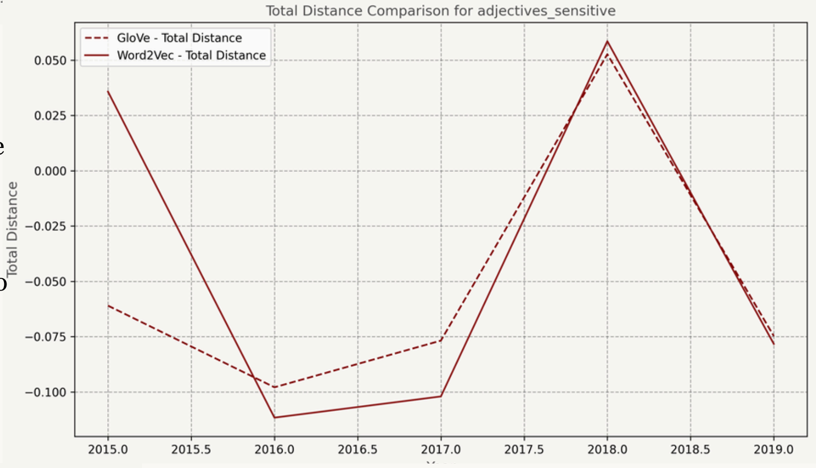
The figure on the right highlights the appearance-related adjective category, showing how women have historically been described more through their bodies, reinforcing stereotypes that prioritize physical attractiveness over other qualities. Before 2017, the graph shows a bias towards women, with scores below 0, reflecting this trend. After 2017, the bias decreased, possibly due to societal changes sparked by the #MeToo movement. The shift in 2019, however, encourages further study over a longer timeline.

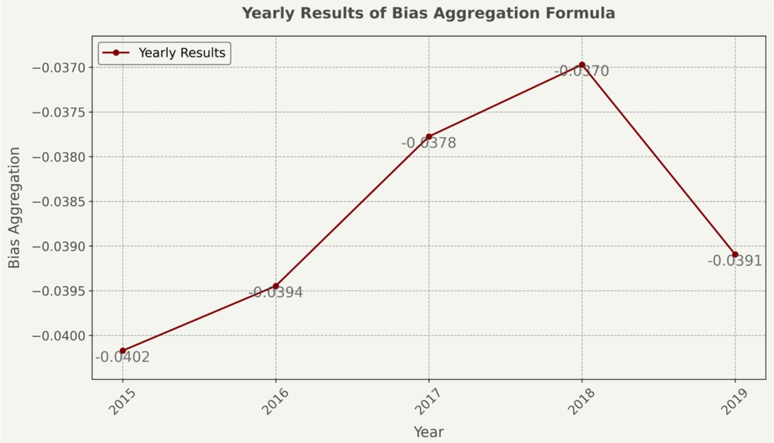




The intelligence-related adjectives plot on the left shows the largest baseline difference between models, with GloVe associating intelligence more with women and Word2Vec with men. Despite this, both models follow the same trend. The bias shift in 2018 could either suggest men were described more as "reflective," "thoughtful," and "adaptable," or the reinforcing of historical stereotypes linking intelligence and competence more closely to men.

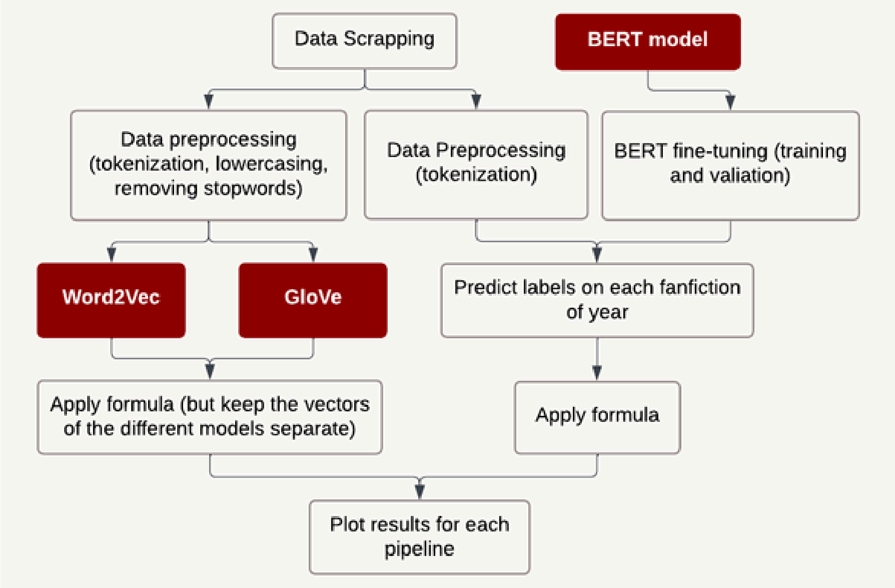
In the sensitive adjectives graph (plot on the right), the bias is more neutral, with values close to 0, but still presents some variations, which might suggest that there was a possible shift in the way society associates sensitivity to either genders. The post-2017 shift is also clearly visible here but it is still a temporary change.





Before 2017, the bias values were strongly negative for the BERT bias plot, indicating fanfiction represented men more. A shift towards gender neutrality begins in 2017, marked by the pivotal year when bias becomes less negative and peaks at -0.0391, the closest association with women in the dataset. Though still negative, this shift highlights a clear change in narrative tone. The lower value in 2019 suggests a possible regression, requiring further exploration.

Method


Data Scrapping
BERT model
Data preprocessing (tokenization, lowercasing, removing stopwords)
Data Preprocessing (tokenization)
BERT fine-tuning (training and valiation)
Word2Vec
GloVe
Predict labels on each fanfiction of year
Apply formula (but keep the vectors of the different models separate)
Apply formula
Plot results for each pipeline

>> Word2Vec/Glove pipeline

This pipeline uses two core embedding models — Word2Vec (trained via gensim) and GloVe (from stanfordnlp). We split the preprocessed data into five groups by publishing date, train Word2Vec for each group, and also apply GloVe separately. Garg et al.'s gendered word lists help identify adjectives in the word embeddings that can be used for stereotypification. We calculate average vectors for women and men, then apply the relative normalized distance formula to measure bias. A more negative value means a stronger correlation with women; while a positive one means stronger correlation with men.

$$\text{relative norm distance} = \frac{1}{|M|} \sum_{v_m \in M} \left(\|v_m - v_1\|_2 - \|v_m - v_2\|_2 \right)$$

>> BERT pipeline

We adapt Zhang C. and Wu B.'s BERT-based method by adding extra normalization and limiting the fine-tuning labels. Each fanfiction is split into 512-token chunks, and we predict label distributions for each chunk—done separately by fanfiction. We then average these predictions at the fanfiction level and compute the yearly bias score:

$$B_{\text{year}} = \frac{\sum_{\text{fanfiction}_i \in Y} B_{i,\text{women}} - \sum_{\text{fanfiction}_i \in Y} B_{i,\text{men}}}{N_{\text{year}}}.$$

Conclusion

This study set out to track gender bias in fanfiction from 2015 to 2019 and see how the #MeToo movement affected it. Two pipelines were tested — one with GloVe and Word2Vec, the other with BERT — and both spotted a noticeable shift toward more female representation and reduced stereotyping after 2017. BERT gave a more nuanced view, while Word2Vec/GloVe were static but simpler to run. Formulas used to aggregate bias scores confirmed that #MeToo indeed pushed fanfiction towards gender neutrality. Still, Word2Vec and GloVe don't account for context, and BERT's binary labels ("herself"/"himself") can't capture all gender identities. Looking ahead, researchers should expand beyond the binary, update word lists with modern terms, and analyze broader datasets across more fandoms and years.