What if fanfiction, but also coding: Investigating cultural differences in fanfiction writing and reviewing with machine learning methods

How has 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?

Irina-Ioana Marinescu[1]

Supervisor(s): Hayley Hung[1], Chenxu Hao[1], 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: Irina-Ioana Marinescu
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Chenxu Hao, Ivan Kondyurin, Elmar Eisemann

Abstract

This paper explores how the portrayal of female characters in fanfiction evolved in response to the #MeToo movement and fourth-wave feminism, with the aim of assessing whether the impact of the awareness of the campaign was broad enough to visibly alter how the average author portrays women in narrative contexts. To analyze these trends, fanfiction data from Archive of Our Own (AO3) spanning 2015–2019 was parsed, and two Natural Language Processing (NLP) pipelines — Word2Vec and GloVe, and BERT — were developed. The study finds that bias scores, aggregated through formulas created to compare gendered associations, show a stronger stereotypization of women before 2017 compared to after. Furthermore, a similar trend is discovered in the representation of women in fanfiction. While the BERT pipeline proved most effective for capturing contextual nuances, it is significantly limited by its reliance on binary labels and computational intensity. This further indicates the need for more inclusive and sustainable methods, making the Word2Vec/GloVe models more appropriate for this task. The paper concludes with recommendations for future work, including broader representation, longer-term analysis, and enhanced detection of evolving language patterns.

1 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 [1]. 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 [2], and therefore being an honest reflection of the beliefs and concerns that exist in the online world en masse.

A large proportion of fanfiction readers and writers are represented by women [3]. Based on this data, a relevant movement for the community was picked, that of fourth-wave feminism. Another justification for the choice is the overlap in time between the movement and the existence of fanfictions in the Internet space. This form of writing started becoming popular in the online during the early 2000s [4], at the same time that the 4th wave feminism started taking shape [5]. A remarkable event that we would take as the pivot of the study is the #MeToo movement, as it manifested itself in the same online spaces as many fanfictions, such as Twitter and Tumblr.

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. In addition, a few subquestions are addressed preliminarily: ”How to identify sexist narration?”, ”What is the appropriate NLP pipeline for retrieving the features of a gender?” and ”How to use the outcomes of the NLP pipeline to measure how misogynistic a text is?”. 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.

The rest of the paper is organized as such: Related Work will dive deeper into existing literature about the topic; Dataset will present all the data used in this study; Methodology will define more clearly the reasoning behind the research questions, the data preprocessing and the way that the pipelines were developed; Experimental Setup and Results will present the parameters used in the models and will provide an analysis of the results of the two pipelines; Responsible Research will be addressing the ethics behind conducting this research from a social sciences point of view; Conclusions and Future Work will present the answer to the research questions and will provide some recommendations for future research in the area.

2 Related Work

Zhang and Wu [6] showed that there was a significant drop in the bias against women in the way people consume books thanks to the #MeToo movement. This paper used 2 different methods to understand gender bias: one with statistical embeddings provided by Word2Vec and GloVe and the other one with Google's Bidirectional Encoder Representations from Transformers (BERT) model. There are a few issues raised by the paper related to how this research topic has been conducted so far. For example, most papers including this one omit the existence of non-binary and transgender people, which leads to further marginalization and exclusion of this group in academia. It is important to note that gender fluidity, together with the intersectionality and complexities of gender and sexuality have been an integral part in fanfiction creation from the very beginning [7], and omitting these practices limits the accuracy of the study. Another crucial point made is that many NLP systems have inherent biases from the training data, research design, pre-trained models and algorithms that might have a later impact on the outcomes of the study.[8]

Another journal paper written by Jonathan Cheng argues that body descriptions and body language in narration can be used to measure the bias against women. Women have been described throughout the past century a lot more through their bodies compared to men, this being possibly tied to Mulveyian theories of sexual objectification. This study though is not complete and only provides a suggestion of how gender bias can be measured in fiction. [9]

When related to direct character analysis in fanfictions, there are two studies that showcase a full pipeline of how character arcs can be studied. Both of these papers use Google's BERT model and have as the main data source fanfictions. A drawback of these works is that they do not focus on gendered issues and in addition, do not take at all into account the biases the models might have. [10] [11]

3 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 motivation behind choosing these canons is due to their popularity in the given time frame, which provides enough data for analysis. The data was scrapped using the A03Scraper repository [12], which consists of a simple and convenient Python script that retrieves the metadata and content of fanfictions that match a given query. The data collection is done with respect to the guidelines for scrapping given in the AO3 terms of service [13], which mostly puts a restriction on how fast data can be queried from the website. It is important to respect this guideline as to not overload the servers and keep the access to the website open and fair. This does on the other hand limit the amount of fanfictions that can be gathered due to the time constraints, as at the time the scraping for this paper's dataset was done, one could retrieve a full fanfiction once every 5 seconds.

The current dataset distribution over the years is shown in Figure 1.

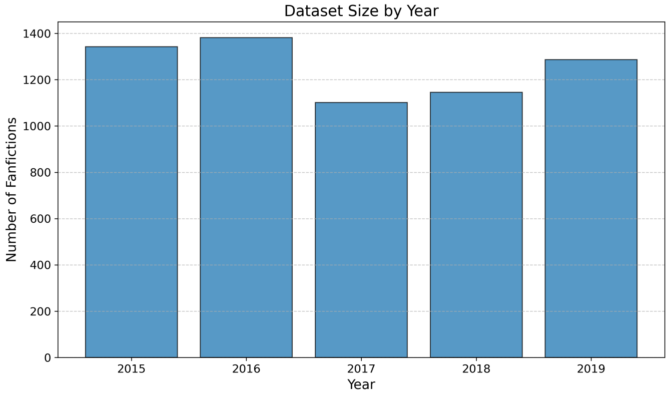


Figure 1: Dataset split over the years.

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.

4 Methodology

In this section, the reasoning behind the research questions will be presented, together with the approach to solving them. Additionally, the data preprocessing techniques will be motivated and a detailed overview of each pipeline will be provided.

4.1 Research Questions

As of November 2024, the servers of AO3 host around 13,910,000 works [14]. As to get a broad enough overview, the target is to parse as many fanfictions as possible from the period 2015-2019, this interval starting 2 years before and ending 2 years after the campaign. To achieve this, the goal is to parse as much text as possible, making the use of Natural Language Processing techniques a requirement.

For the quantification of the trends throughout the years, the notion of bias is useful. Bias has different definitions based on the context in which it is used. In this paper it represents the historical bias, that related to how machine learning mimics the bias in the real world, as the data processed in itself contains unwanted properties. The focus is further narrowed down to gender bias, which is a type of correlation bias, as it maps a potentially correlated variable (such as ”cooking”) to a demographic attribute (such as ”man” or ”woman”).[15]. By training models on biased data and afterwards assessing their bias, conclusions can be drawn about real life stereotypes. The papers referenced in related work show a few ways to calculate several types of bias through various methods.

But this begs the question, what is an appropriate pipeline for this task. Word2Vec is one of the less computationally intense models while also being good at pointing out biases, by generating vectors for each word based on the neighboring words. In addition, the GloVe model is just as widely used and its purpose is also to generate word embeddings. In order to increase accuracy, the BERT model proves to be better fitted by being able to understand the context of words, but it is costly from a computational point of view, which wastes a lot of energy for training, fine-tuning and predicting. [16]

The two pipelines will be employed as a way of validating each others results and comparing efficiency. The results will also be compared with the real life trend, as a supervised approach to validating these methods is unfeasible resources-wise. Based on this comparison, a conclusion on the pros and cons of each pipeline will be drawn.

Finally, it is important to understand how to interpret the results of the models. For this two formulas for bias aggregation are discussed in the pipelines subsection. With these results, we can plot them and assess if they answer the question on whether there was a shift in the bias between before and after 2017.

4.2 Data preprocessing

In order to make the data be an appropriate input for the Word2Vec and GloVe models, the scrapped text has to undergo lowercasing, and removal of punctuation and lemmatization. For tokenization the simple white space delimiter was used, as it is the most efficient and sufficient for the task of searching for our word lists, since these are terms that usually are not used in compound forms. Lowercasing is necessary to create uniformity, especially if some of the interest words are at the start of the sentence so they would be capitalized and therefore accidentally taken into account separately in the vector space. Next is the punctuation removal, as for these two models, there is no need for the additional information provided by these characters, and would only just clutter the data. Finally, lemmatization, is applied as some of the evaluated words are used in different forms, for example ”beautifully” instead of ”beautiful”. [17]

Additionally, in this pipeline a few lists of adjectives that represent stereotypical terms used for men and women are used. The first two lists represent terms that are usually used for identification: he, son, his, him, father, etc. and she, daughter, hers, her, mother, etc. The other lists are comprised of general stereotypical adjectives [18] [19], competency adjectives [20] and physical adjectives [21].

For the pretrained BERT model, the data is prepared by cleaning it of special characters, extra spaces and lowercasing it. Afterwards the data is ran through the BertTokenizer. [22]

The BERT model chose for bias calculation is trained on BookCorpus and English Wikipedia [23], making it appropriate for the task of analyzing fiction. To specialize the model for identifying gender bias, an additional dataset was used for fine-tuning - stereoset - which was created with the help of human annotation and contains labeled data about ethnic and gender stereotypes [24]. The dataset for fine-tuning is prepared by using the BertTokenizer as well.

4.3 Pipelines

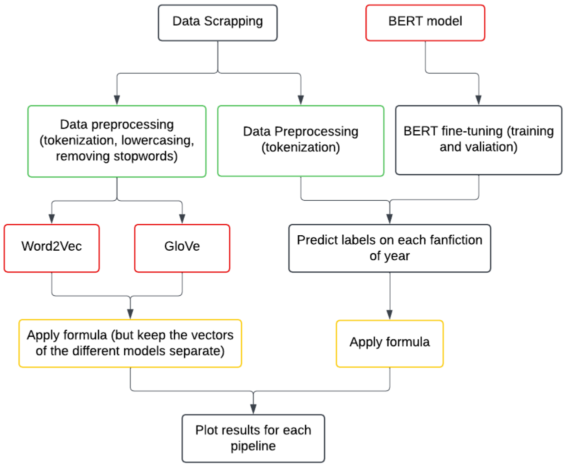

Data Scrapping
BERT model
Data preprocessing (tokenization, lowercasing, removing stopwords)
Data Preprocessing (tokenization)
BERT fine-tuning (training and valiation)
Word2Vec
GloVe
Predict labels on each fanfiction of year
Apply formula (but keep the vectors of the different models separate)
Apply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

Figure 2: Pipeline diagram.

The first pipeline has at its core the Word2Vec and GloVe models. The Word2Vec model is trained with the parameters described in the experimental setup section. Due to hardware limitations, the implementation used for Word2Vec the one found in the gensim library. The GloVe model used is the one provided by stanfordnlp [25]. Garg et al [26] has created several lists of gendered words that were previously mentioned, which were collected through surveys or websites that prove to be helpful in finding the trends of how women and men are described in fiction. These word lists are used to calculate the relative normalized distance between the valid embedded vectors of the dataset and how the trends of these words shift during time.

The preprocessed data is split into 5 groups based on the publishing date. For each year the Word2Vec skipgram model is applied in order to convert the words into embeddings. Separately the GloVe model is applied and its output vectors saved.

Taking into consideration the outputted vectors, the valid adjectives will be filtered based on whether they appear both in the word lists and the word embeddings. Afterwards, the average embedding vector for terms that represent women and for terms that represent men is calculated. This is done by taking all the words in the corresponding list and averaging their values to create only one. With all of these steps completed, the following formula is applied:

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

where v_m represents the adjective from the M adjective list, $v1$ represents the average embedding vector for women and $v2$ is the average embedding vector for men. [26]. The additional normalization is added due to the fact that different years might have a different number of valid adjectives.

The more negative the bias is, the stronger the correlation is between women and that set of adjectives, while the opposite implies it is more closely associated with men. This applies to both models.

For the second pipeline, the BERT model method used by Zhang C. and Wu B. in characterizing gender stereotypes in fiction [6] was adapted for this task, specifically by including additional normalization in the formula and limiting the labels used from the fine-tuning.

Each fanfiction is split into chunks of data of 512 tokens, as this is the limitation given by the BERT model. Using the fine-tuned model, we predict the label distributions for each of the chunks. This is done separately for each year. The fanfiction-level results are aggregated by averaging to get the general score for each year:

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

where B_{year} represents the aggregated bias for all fanfictions in a year Y, $B_{i,\text{women}}$ is the sum of the probability per each fanfiction of the 'herself' label, which is label 0, $B_{i,\text{men}}$ is the sum of the probability per each fanfiction of the 'himself' label, which is label 31, and N_{year} is the number of fanfictions per each year. This final aggregated bias is the one used for comparison over the years.

5 Experimental Setup and Results

The Word2Vec skipgram model is initialized to a vector size of 300 (the same size as the vectors created in the GoogleNews dataset), a window size of 5 and a min count of 5. The non-hardware related parameters used when running the GloVe model are: a min count of 5 (the same as for the Word2Vec model), the cutoff for the weighing function at 10, since the dataset is not too sparse but neither too large, 15 training iterations, and a vector size of 100 dimensions, as to capture enough semantic information but to not overfit the model.

The graphs showcase the gender bias trends in the word embeddings, the plots containing both the GloVe outputs of the formula, but also the Word2Vec ones. (Figures 3, 4, 5) By

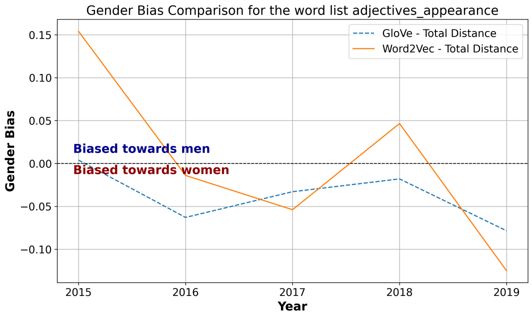


Figure 3: Gender bias for the list of adjectives pertaining appearance.

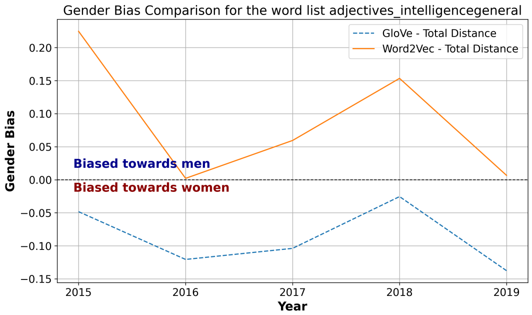


Figure 4: Gender bias for the list of adjectives pertaining intelligence.

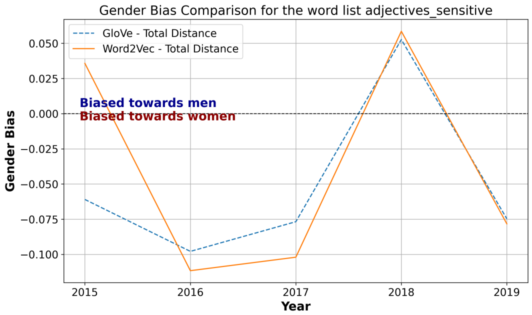


Figure 5: Gender bias for the list of adjectives pertaining sensitivity.

interpreting the formula, we understand that for each year and adjective set, we compute the distance between the averaged vectors for men and women and the neutral adjective list. The higher the distance between the two, the less is the association between the specific gender and the adjectives, while the lower the distance, the stronger it is. By substracting the distances ($distance\_woman - distance\_man$), the gender bias score is computed. If the score is positive, it indicates that the adjective list is more strongly associated with men, while if it is negative, it means it is closer to women. The graphs show the bias score over the years, plotted separately for adjectives pertaining to appearance, intelligence and sensitivity.

Figure number 3 shows the adjective category of appearance, which is especially relevant as research has shown that historically women have been described more through their bodies and this leads to reinforcing stereotypes that put the accent on physical attractiveness over other qualities. [9] In the graph, it is visible that before 2017 there was a bias towards women, with scores under 0, which matches the historical use for appearance-related terms for describing female characters. After 2017, the bias is significantly reduced for a year, which might be a reflection of the societal changes that were sparked by the #MeToo movement. To understand what happened in 2019 a study on a longer timeline should be conducted.

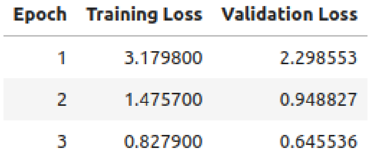
In the sensitive adjectives graph (Figure 5), the bias is more neutral, with values close to 0, but still presents some variations, which might suggest that there was a possible shift in the way society associates sensitivity to either genders. The post-2017 shift is also clearly visible here but it is still a temporary change.

Finally, in the intelligence-related adjectives plot, the largest baseline difference between the two models is noticeable, GloVe showing that intelligence attributes are overall associated with women, while Word2Vec to men. Nonetheless, the highlighted trend is the same. The bias shift in 2018 might suggest that men were described more as ”reflective”, ”thoughtful” and ”adaptable”, but also it might still reflect that the historical stereotypes of intelligence and competence being more closely associated with men over women were actually only accentuated. [27]

Taking all the plots into account, a conclusion can be drawn that fanfiction writers have shifted significantly towards using less objectifying language when it comes to appearance, and while in the other graphs the biases still remain, the trends show a slight reduction into gendered associations.

Between the two models, the trends match but there is a sizable difference between the bias values. One of the reasons that might have lead to it is because GloVe calculates the embeddings based on the global co-occurrence statistics, while Word2Vec only uses local windows to deduce context. These might have influenced the baseline distances. Another reason for this happening could be the difference in dimensions between the 2 models, as the Word2Vec embeddings have 300 dimensions and the GloVe ones only 100, and that increases the distance on average of the vectors, as adding dimensions increases the space's complexity. Nonetheless, both models capture the same similarity in the trends, which highlight the language shifts used in writing fanfiction.

BERT is loaded using the BertForSequenceClassification with the pretrained version of 'best-base-uncased'. The finetuning is completed in 3 epochs, as that seemed to be the appropriate trade-off between accuracy and overfitting, as seen in Figure 6.



Epoch	Training Loss	Validation Loss
1	3.179800	2.298553
2	1.475700	0.948827
3	0.827900	0.645536

Figure 6: The training and validation loss for the 3 epochs.

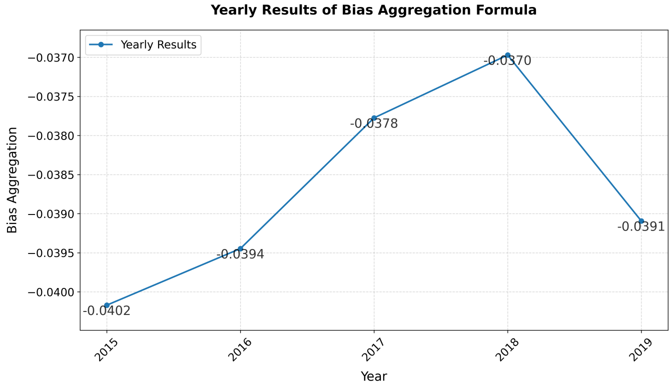


Figure 7: Bias aggregation results of the BERT pipeline plotted over the years.

Figure 7 shows the yearly aggregated bias values calculated with the formula presented in methodology. The values can be understood as follows: if the values are positive, there is a stronger association between the fanfiction content in that year and women, as the formula is calculated using the ”herself” label, and if the values are negative, there is a stronger association with men, also calculated with the ”himself” label.

Some key takeaways from the plot is that before 2017, the bias values were strongly negative, highlighting that fanfiction in this period of time was more closely tied to men. This is followed by an upwards trend that aims for gender neutrality. 2017 is the pivotal year, as the bias here is becoming less negative, showing a shift in gender representation, followed by the peak value (-0.0391) which is the closest association with women observed in the dataset. Despite it still being a negative value, the change does prove a clear shift in the narrative tone. The fact that the 2019 value is lower again might show some sort of regression, which does need further exploration to understand the trend. With these in mind, the rise between 2017 and 2018 could be attributed to the increased societal awareness of gender equality and representation thanks to the #MeToo movement.

The method applied to produce this plot does have some limitations, as the timeline might be too short to actually understand why in 2019 there was a drop again, and it makes it harder to understand if the post-2017 trend continues or stabilizes. In addition, using the binary labels ”herself” and ”himself” is not useful for capturing more nuanced gender representations, this issue being further addressed in the responsible research section. Finally, the bias values are throughout all the years negative, which suggests a permanent male-leaning narrative.

Comparing the 3 models, the trend is consistent in all of them, and do support the hypothesis that the #MeToo movement has pushed towards a more gender neutral language and evened out the associations of neutral terms and gender.

6 Responsible Research

6.1 A discussion on Gender Bias

Bias at its core is a preference or prejudice for a category over the other. In this paper the specific bias addressed is the gendered representation bias, which implies that some systems create assumptions based on the identity and representation of a gender.[28] Bias is not inherently negative, but it can enhance some already existing harmful stereotypes, which further will affect individuals.

When it comes to NLP models, there are a few challenges that arise. One of them is that gender stereotypes run deeper than just adjectives and descriptors for men and women. Stereotypes appear in roles, actions, and other narrative arcs, which cannot be captured by either of the pipelines in this paper. [29]. Another struggle for the models in this paper is that they are trained on older datasets. Language evolves over time, and while in general fiction usually terms adhere to established connotations, fanfiction is produced at a much faster pace and can show new emerging trends for expressing gender-related ideas.[30] These terms might describe bias but are not included neither in the training, fine-tuning or word list datasets of the paper. Finally, models like BERT can also struggle with understanding irony and critique, like making a character be overtly stereotypical as a way of criticizing the existence of the stereotype in the first place. NLP models are prone to misinterpreting the author’s intent, and can push the bias calculations in the wrong direction.

6.2 Non-binary gender representation

Some people do not identify themselves as neither a man nor a woman, since gender is a large spectrum. The identity of these people usually falls under the umbrella terms of genderqueerness or the non-binary gender. [31] Gender non-conforming people usually tend to use therefore non-gendered terms for self-identification, for which the Word2Vec and GloVe pipeline does not have a word list: they/them, sibling, parent, etc. Non-binary people are considered to be under the umbrella of the feminist movement, as the International Alliance of Women, for example, highlights that everyone’s rights are universal and indivisible, supporting a more inclusive approach that encompasses all gender identities. [32] Therefore not taking them into consideration when calculating the bias makes the results less accurate for evaluating the 4th feminist movement in fanfiction. The BERT pipeline on its own introduces additional challenges, as the limiting analysis of the two binary labels - ”himself” and ”herself” also do not capture properly all the forms of gender expression. This limitation stems from the way that the fine-tuned data is annotated, but it does narrow the scope of the results, and has an influence on the research outcomes.

7 Conclusions and Future Work

The purpose of this study is to observe the trends in gender bias in fanfiction over the 2015-2019 period in order to understand the societal changes that happened before and after the rise of the #MeToo movement.

Two pipelines were evaluated, one of them containing the GloVe and Word2Vec model, while the other one employs only the BERT model. Both of these pipelines manage to catch a gender bias shift, which proves that they are all useful for this task. However, BERT does come with significant computational costs, which requires a lot of energy consumption and is unsustainable.

By aggregating the yearly bias using mathematical formulas, a shift in the gender bias pre and post #MeToo movement was visible. Results from both pipelines consistently show a swing towards more women representation and a destereotypization after 2017. This answers the question by showing that the #MeToo movement brought awareness among the general public and pushed the fanfiction sphere towards gender neutrality.

There are certain limitations to the pipelines as well, as the Word2Vec and GloVe pipeline are constrained by their static nature and are not able to understand context or changing language patterns. On the other hand, BERT provided a more nuanced understanding, but it relies too much on the binary labels (”herself” and ”himself”).

For future work, it is highly encouraged to broaden the gender representation, by moving beyond the binary labels, as this is a critical point in capturing the diversity of gender expression. Word lists can also be developed more and updated to contain slang and newer gender terminology, as to improve the bias detection. Finally, larger datasets from more fandoms and a longer temporal analysis could be helpful in understanding the trends better and how they evolve after 2019 too.

In conclusion, this research shines a light on the usefulness of NLP tools to analyze gender bias in fanfiction, a genre that both reflects and challenges societal norms. While the results highlight the shift in bias following #MeToo, they also raise the need for more inclusive methodologies to account for the full complexity of gender representation. By refining the tools and datasets used for this type of analysis, future studies can deepen our understanding of how societal changes are reflected in textual narratives, leading to more responsible and equitable applications of NLP in cultural research.

References

[1] Sara L. Uckelman. *Fanfiction, Canon, and Possible Worlds*. PhilArchive. 2018. URL: https://philarchive.org/archive/UCKFCA.

[2] Bronwen Thomas. “What is fanfiction and why are people saying such nice things about it?” In: *Storyworlds: A Journal of Narrative Studies* 3 (2011), p. 6. DOI: 10.5250/storyworlds.3.2011.0001. URL: https://www.jstor.org/stable/10.5250/storyworlds.3.2011.0001.

[3] Centrum Lumina. *Gender breakdown of AO3 users*. Accessed: 2024-11-11. 2013. URL: https://centrumlumina.tumblr.com/post/62816996032/gender.

[4] Purdue University Fort Wayne Library. *Fanfiction 101: Customizing Your Superheroes*. Accessed: 2024-11-11. 2024. URL: https://library.pfw.edu/c.php?g=16316&p=89234.

[5] Encyclopaedia Britannica, Inc. *Feminism: The Fourth Wave*. Accessed: 2024-11-11. n.d. URL: https://www.britannica.com/explore/100women/rise-of-feminism/feminism-the-fourth-wave.

[6] C. Zhang and B. Wu. “Characterizing gender stereotypes in popular fiction: A machine learning approach”. In: *Online Journal of Communication and Media Technologies* 13.4 (2023), e202349. DOI: 10.30935/ojcmt/13644.

[7] J. Duggan. “Trans fans and fan fiction: A literature review”. In: *Transformative Works and Cultures* 39 (Mar. 2023). DOI: 10.3983/twc.2023.2309.

[8] Dirk Hovy and Shrimai Prabhumoye. “Five sources of bias in natural language processing”. In: *Language and Linguistics Compass* 15.8 (2021). Dirk Hovy and Shrimai Prabhumoye contributed equally, e12432. DOI: 10.1111/lnc3.12432.

[9] J. Cheng. “Fleshing Out Models of Gender in English-Language Novels (1850–2000)”. In: *Journal of Cultural Analytics* 5.1 (2020). DOI: 10.22148/001c.11652.

[10] Md Naimul Hoque et al. *Portrayal: Leveraging NLP and Visualization for Analyzing Fictional Characters*. University of Maryland, College Park and Stony Brook University. 2023. URL: https://naimulh0que.github.io/docs/dis23-42.pdf.

[11] Sriharsh Bhyravajjula, Ujwal Narayan, and Manish Shrivastava. *MARCUS: An Event-Centric NLP Pipeline that generates Character Arcs from Narratives*. 2022. URL: https://ceur-ws.org/Vol-3117/paper7.pdf#page=8&zoom=100,118,393.

[12] Jingyi Li and Sam Sterman. *AO3Scraper*. Accessed: Jan. 6, 2025. 2017. URL: https://github.com/radiolarian/AO3Scraper.

[13] Archive of Our Own. *AI and Data Scraping on the Archive*. Accessed: Jan. 6, 2025. 2023. URL: https://archiveofourown.org/admin_posts/25888.

[14] Organization for Transformative Works. *Archive of Our Own*. Accessed: 2024-11-04. 2024. URL: https://web.archive.org/web/20241104145222/https://archiveofourown.org/.

[15] T. Hellström, V. Dignum, and S. Bensch. “Bias in Machine Learning – What is it Good for?” In: *arXiv* arXiv:2004.00686 (Sept. 2020). DOI: 10.48550/arXiv.2004.00686. URL: https://arxiv.org/abs/2004.00686.

[16] Justine Calma. *AI is using up more and more electricity*. Accessed: 2025-01-23. Sept. 2023. URL: https://www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption.

[17] C. P. Chai. “Comparison of text preprocessing methods”. In: *Natural Language Engineering* 29.3 (May 2023), pp. 509–553. DOI: 10.1017/S1351324922000213.

[18] John E. Williams and Deborah L. Best. “Sex Stereotypes and Trait Favorability on the Adjective Check List”. In: *Educational and Psychological Measurement* 37.1 (1977), pp. 101–110. DOI: 10.1177/001316447703700115.

[19] John E. Williams and Deborah L. Best. *Measuring Sex Stereotypes: A Multination Study*. Revised. Vol. 6. Cross Cultural Research and Methodology. Thousand Oaks, CA: Sage Publications, 1990. ISBN: 9780803938151.

[20] Writing Recommendation Letters Online. *Superlatives*. Accessed: Jan. 6, 2025. URL: https://www.e-education.psu.edu/writingrecommendationlettersonline/node/151.

[21] Sight Words, Reading, Writing, Spelling & Worksheets. *Appearance Adjectives*. Accessed: Jan. 6, 2025. URL: https://www.sightwordsgame.com/parts-of-speech/adjectives/appearance/.

[22] Samia Khalid. *BERT Explained: A Complete Guide with Theory and Tutorial*. Accessed: 2025-01-23. Sept. 2019. URL: https://towardsml.wordpress.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/?utm_source=chatgpt.com.

[23] Tensorflow Transformers. *BERT base model (uncased)*. Accessed: Jan. 6, 2025. 2018. URL: https://huggingface.co/tftransformers/bert-base-uncased.

[24] Moin Nadeem, Anna Bethke, and Siva Reddy. *StereoSet: Measuring stereotypical bias in pretrained language models*. Accessed: Jan. 6, 2025. 2020. arXiv: 2004.09456 [cs.CL]. URL: https://huggingface.co/datasets/McGill-NLP/stereoset.

[25] Stanford NLP Group. *GloVe: Global Vectors for Word Representation*. Accessed: 2025-01-23. 2025. URL: https://github.com/stanfordnlp/GloVe.

[26] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (Apr. 2018), E3635–E3644. DOI: 10.1073/pnas.1720347115.

[27] B. A. Nosek, M. R. Banaji, and A. G. Greenwald. "Harvesting implicit group attitudes and beliefs from a demonstration web site". In: *Group Dynamics: Theory, Research, and Practice* 6.1 (Mar. 2002), pp. 101–115. DOI: 10.1037/1089-2699.6.1.101.

[28] T. Sun et al. "Mitigating Gender Bias in Natural Language Processing: Literature Review". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1630–1640. DOI: 10.18653/v1/P19-1159.

[29] Y. Hitti et al. "Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype". In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Ed. by M. R. Costa-jussà et al. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 8–17. DOI: 10.18653/v1/W19-3802.

[30] Shegufa Eram. "Sexism in Language: A Case Study on Bangladeshi Youth". Master's Thesis, Department of English and Humanities. BRAC University, 2015.

[31] C. Richards et al. "Non-binary or genderqueer genders". In: *International Review of Psychiatry* 28.1 (Jan. 2016), pp. 95–102. DOI: 10.3109/09540261.2015.1106446.

[32] Wikipedia contributors. *International Alliance of Women*. Accessed: Jan. 25, 2025. Jan. 2025. URL: https://en.wikipedia.org/w/index.php?title=International_Alliance_of_Women&oldid=1269026979.