Evaluating Gender Bias Transfer from Film Data

Amanda Bertsch,* **Ashley Oh,*** **Sanika Natu,*** **Swetha Gangu,***
Alan Black, Emma Strubell
Carnegie Mellon University
[abertsch, ashleyoh, snatu, sgangu]@cs.cmu.edu

Abstract

Films are a rich source of data for natural language processing. OpenSubtitles (Lison and Tiedemann, 2016) is a popular movie script dataset, used for training models for tasks such as machine translation and dialogue generation. However, movies often contain biases that reflect society at the time, and these biases may be introduced during pre-training and influence downstream models. We perform sentiment analysis on template infilling (Kurita et al., 2019) and the Sentence Embedding Association Test (May et al., 2019) to measure how BERT-based language models change after continued pre-training on OpenSubtitles. We consider gender bias as a primary motivating case for this analysis, while also measuring other social biases such as disability. We show that sentiment analysis on template infilling is not an effective measure of bias due to the rarity of disability and gender identifying tokens in the movie dialogue. We extend our analysis to a longitudinal study of bias in film dialogue over the last 110 years and find that continued pre-training on OpenSubtitles encodes additional bias into BERT. We show that BERT learns associations that reflect the biases and representation of each film era, suggesting that additional care must be taken when using historical data.

1 Introduction

Movies are often seen as a commentary on or reflection of society. They can reveal key themes within a culture, showcase the viewpoints of various social classes, or even reflect the writer's internal mindset. Additionally, movies have widespread influence on audience perceptions based on the messages they contain.

Movie scripts are popular data sources for training models for natural language tasks, such as sentiment analysis (Frangidis et al., 2020) and dialogue systems (Serban et al., 2015), because they are written to mimic natural human dialogue, easy to collect, and much more cost effective than transcribing human conversations.

However, despite this popularity, there has been concern regarding the biases that movies contain (Schofield and Mehr, 2016) and the potential downstream effects of training on biased datasets (Kumar et al., 2020). More specifically, gender bias in movies is a long-studied issue. A popular benchmark for gender representation is the Bechdel test[1]. A movie passes the Bechdel test if it contains two female characters who speak to each other about something other than a man.

In the last decade, the Bechdel test has come under criticism. O'Meara (2016) argues that the Bechdel Test is a poor metric in three ways: it excuses "low, one-dimensional standards" for representation, it fails to consider intersectionality of oppression, and it treats all conversation about men as unempowering.

As a more intersectional and nuanced method of measuring bias and stereotyping in movie script datasets, we propose fine-tuning a language model on movie scripts in order to examine bias that the model inherits from movies and its impact on downstream tasks. Particularly, a model trained on movie scripts may inherit biases or offensive language from the source material, which can lead to differing treatment of social groups in applications of the model. In a longitudinal analysis of bias over time, we evaluate how models that are fine-tuned on separate decades of movie scripts reflect societal biases and and historical events at the time. The form of fine-tuning we use is a continuation of the pre-training objectives on the new dataset. The contributions of this paper are:

- an analysis of additional bias introduced into BERT by continued pre-training on movie scripts, where we find that gender bias in the model is increased when film data is added.

*Equal contribution

[1]https://bechdeltest.com/

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 235 - 243

- a historically grounded analysis of social biases learned from film scripts by decade, considering gender, racial, and ideological biases.

2 Bias statement

In our analysis we use a language modeling approach to uncover and examine bias in a movie script corpus. Our main focus is gender bias, but we will also explore intersectional bias between gender and disability. We define bias as implicit bias that may result in a difference in treatment across two groups, regardless of whether that difference causes harm. This definition of implicit bias follows from the premise of the Implicit Bias Association test (Greenwald et al., 2009), which demonstrated that implicit biases impact behavior. Our analysis also considers both explicit and implicit gender biases that have the capability for harm. In this paper we assume biases in movies are intentional, but it is possible the author may have been using these stereotypes as a method of raising awareness of an issue or as satire. It is important to note that models trained on these movie scripts will likely not be able to pick up on the intent of the author, but rather will learn and amplify the biases (Hall et al., 2022).

This analysis includes a comparison between the treatment of men and woman in film scripts, which implicitly upholds a gender binary. We fine-tune BERT on full movie scripts without partitioning by gender, but we examine gender bias by comparing the associations the model has learned about men and women during the analysis. By discarding data about people who are nonbinary, we make this analysis tractable, but we also lose the ability to draw meaningful conclusions about this underrepresented group. We choose to reduce harm by not assuming the genders of characters; rather, we consider the associations the model has learned about gender from the speech of all characters. Thus, our analysis is more likely to represent biases in how characters discuss men and women who are not present, rather than how characters treat men and women in direct conversation.

3 Related Work

A significant amount of research has examined and quantified gender bias in movie scripts and narratives. Past work has focused on bias in film dialogue, using classification models to predict whether speakers are both female, both male, or of different genders. Schofield and Mehr (2016) concluded that simpler lexical features are more useful than sentiment or structure when predicting gender.

Ramakrishna et al. (2015) use gender ladenness, a normative rating representing word association to feminine and masculine traits, to explore gender bias. Specifically, they examine gender ladenness with respect to the movie's genre, showing that certain genres are more likely to be associated with masculine/feminine traits than others. Gala et al. (2020) add to the genre and gender association, finding that certain sports, war, and science fiction genres focus on male-dominated tropes and that male-dominated tropes exhibit more topical diversity than female-dominated tropes.

Huang et al. (2021) show that in generated stories, male protagonists are portrayed as more intellectual while female protagonists are portrayed as more sexual. Sap et al. (2017) look at more subtle forms of gender bias as it relates to power and agency. Their work uses an extended connotation lexicon to expose fine-grained gender bias in films.

Ramakrishna et al. (2017) also looked at the differences in portrayals of characters based on their language use which includes the psycholinguistic normative measures of emotional and psychological constructs of the character. They found that female writers were more likely to have balanced genders in movie characters and that female characters tended to have more positive valence in language than male counterparts in movie scripts.

While these works focus on understanding bias in film directly, we take a slightly differently framing, examining how the bias in a film dataset can impact the biases of a language model.

Loureiro et al. (2022) examine concept drift and generalization on language models trained on Twitter data over time. Our work on longitudinal effects of film data is distinct in timescale (reflecting the much slower release rate of films relative to tweets) and in motivation; (Loureiro et al., 2022) consider the effects of the data's time period on model performance, while we examine the effects of the time period on model biases.

4 Methods

We examine how a BERT-based language model (Devlin et al., 2019) may inherit bias from film data. Specifically, we use the OpenSubtitles corpus (Lison and Tiedemann, 2016), a collection

of movie subtitles from approximately 400,000 movies. While the corpus does not provide summary statistics, upon inspection it appears the vast majority of these movies are American-produced films. These subtitles do not contain speaker gender, and often do not provide speaker names. Thus, any bias exhibited in the model is likely from the way the characters speak about people from different groups– e.g. indirect, not direct, sexism.

We use the OpenSubtitles corpus to gather sentences within each movie script and randomly mask words to fine-tune BERT on the movie corpora. Following previous work by von Boguszewski et al. (2021) that focused on toxic language detection in BERT fine-tuned on movie corpora, we considered bias in the original English pre-trained BERT as a baseline and BERT fine-tuned on movie corpora (which we call FilmBERT) as a secondary model. We used two approaches to quantify bias in the models, which we describe in the following sections. We then employ a longitudinal analysis of BERT by fine-tuning on decades from 1910 to 2010 in order to quantify what societal trends and biases the model may absorb.

4.1 Measuring Intersectional Bias through Sentiment Analysis

We adopt the method used by Hassan et al. (2021) to measure how the presence of gender or disability identity tokens affects the sentiment of the predicted token in a template infilling task. We create templates in the form "The [GENDER] [DISABILITY] person [VERB] [MASK]," where [GENDER] and [DISABILITY] were filled with tokens related to gender and disability. The gender list was chosen for gender inclusiveness (Bamberger and Farrow, 2021) and the disability tokens were based on prior work by Hutchinson et al. (2020). The templates can be separated in 4 classes, "None" which have no identifying tokens and will serve as our control, "Disability" which contains a token from the disability list, "Gender" which contains a word from the gender list and "Disability+Gender" which contains one disability token and one gender token. To filter out sub-embeddings and punctuation, predicted tokens that contained non-alphabetic characters were removed. The predicted tokens were then put into a template in the form "The person [VERB] [PREDICTED TOKEN].". This allows us to measure the sentiment of the predicted token without considering the sentiment of the [GENDER] or [DISABILITY] token. The sentence-level sentiment scores were obtained from Textblob polarity [2]. We extend the work of Hassan et al. (2021) by running a pairwise t-test between sentiment scores for the classes produced by BERT and FilmBERT.

4.2 Sentence Embedding Association Test

The Word Embedding Association Test (Islam et al., 2016) is a popular tool for detecting bias in non-contextualized word embeddings. It was adapted for sentence-level embeddings by May et al. (2019) to produce the Sentence Embedding Association Test, which can be applied to contextualized embeddings. This test measures the cosine similarity between embeddings of sentences that capture attributes (such as gender) and target concepts (such as likeability). May et al. caution that this test may underrepresent bias in embeddings; however, when applied with care, it can provide strong evidence of biased associations over social attributes and roles.

We use the original sentence embedding tests developed by May et al. (2019), which examine a variety of biases. There are 6 tests that measure gender associations. The tests measure whether female names or female terms (e.g. "woman," "she") are more strongly associated with words for family life over careers, arts over math, or arts over science, relative to male equivalents. Other tests measure the professional "double bind," where women in professional settings who are more competent are perceived as less likeable (Heilman et al., 2004); the "angry black woman" stereotype, an intersection of racist and sexist stereotypes (Motro et al., 2022); racial biases, where African American names and identity terms are compared to European American names and identity terms; and word connotation differences, such as instruments being more pleasant than weapons or flowers being more pleasant than insects.

4.3 Longitudinal Study

The OpenSubtitles corpus contains movie scripts from the early 1900s to the 2020s. We partition the dataset by decade and fine-tune BERT on each decade's data individually, producing 11 decade models, which we label FilmBERT-1910s to FilmBERT-2010s. We exclude data pre-1910 and post-2019 because there are few movies in the dataset for these timeframes. We also exclude all

[2] https://textblob.readthedocs.io/en/dev/

music videos, restricting the sample to feature films. Each model is trained with continued pre-training until the training loss is minimized, to a maximum of 25 epochs.

5 Fine-tuning on Entire Corpora Results

First, we consider results from continued pre-training over the entire OpenSubtitles dataset.

5.1 Sentiment Analysis

We were not able to replicate similar results to Hassan et al. (2021) with BERT. All of the classes were weakly negative to neutral as expected. "None" was reported to have the highest sentiment by Hassan et al. (2021), but had the lowest average sentiment in our replication. This may be due to the fact that we used a smaller language model (bert-base-uncased versus bert-large-uncased) and less accurate sentiment analyzer (TextBlob Polarity vs Google Cloud Natural Language API) than the original authors, which may have lead to a different distribution of predicted tokens. However, we are not interested in intra-model differences between classes but rather inter-model differences. That is, we would like to compare the average sentiment from BERT against FilmBERT for each class.

We hypothesized the sentiment for gender would become more negative. Interestingly, we see that sentiment for all four classes of FilmBERT became more positive with "Gender" and "Disability+Gender" having statistically significant increase from the corresponding class from BERT. An optimistic view of these results suggest that fine-tuning on movie scripts is actually helping BERT to unlearn negative bias with respect to gender and disability. Given the template "the lesbian person in a wheelchair feels [MASK]." BERT produces the following tokens: ['uncomfortable', 'awkward', 'isolated', 'guilty', 'sick', 'helpless', 'threatened', 'trapped', 'alone', 'powerless']. Clearly, the predicted tokens all have negative sentiment. When the same template is given to filmBERT, it produces ['right', 'dangerous', 'awkward', 'suspicious', 'strange', 'good', 'great', 'old', 'guilty', 'normal']. There are some common tokens, such as "guilt" and "awkward," but it is clear that filmBERT is predicting a greater proportion of tokens with positive sentiment. Additional examples are available in Table 3 in the Appendix.

5.2 Discussion and Limitations

It is also possible that the sentiment analysis approach is simply not a good measure of dataset bias. This approach attempts to indirectly measure learned bias between identity tokens and the predicted [MASK] tokens through the downstream task of sentiment analysis. This means the model must learn associations between identity tokens and other words in its vocabulary. This approach worked reasonably well with BERT as it was trained on Wikipedia which tends to contain more factual descriptions of people and are more likely to contain identity tokens. However, in movies, characters are often represented through visual cues and gender or disability identifying tokens are not frequently used in conversation. Additionally, models such as BERT that use contextualized word embeddings have difficulty effectively representing rare words (Schick and Schütze, 2019). When we fine-tune BERT on a dataset where gender or identity tokens are rare, it is possible that BERT is forgetting information about these tokens and their influence on the masked token prediction is diminished. Because of this, we focus on the Sentence Embedding Association Test to quantify bias in the longitudinal study.

6 Longitudinal Study Results

We use the Sentence Embedding Association Test (May et al., 2019) to quantify the bias in each of the decade models, using the original association tests designed by the authors. These tests measure the association between two contrasting sets of identity terms (e.g. male-identifying and female-identifying terms) and two non-identity-based sets (e.g. career-related terms and family-related terms). We consider only associations that are significant ($p < 0.05$), and factor both the number of significant associations found and the relative effect sizes into our analysis.

6.1 Gender Stereotypes

The original BERT model does not exhibit significant associations for any of these tests, as reported in May et al. (2019), but the film decade models display a clear pattern. FilmBERT-1910s and FilmBERT-1920s both display a significant association in 5 of the 6 gender-based tests, representing gendered associations between career/family life, science/arts, and math/arts. On average, the effect size is slightly larger for FilmBERT-1920s.

Class	# Templates	BERT Mean Sentiment	filmBERT Mean Sentiment	P Value
None	14	-0.00267 ±0.02	0.00431 ±0.13	0.1268
Disability	168	-0.00063 ±0.03	-0.00027 ±0.01	0.5381
Gender	238	-0.00214 ±0.04	0.00061 ±0.02	0.00135
Disability+Gender	2856	-0.00196 ±0.03	-8.647e-6 ±0.01	4.2e-41

Table 1: Sentiment Average and Variance by class for BERT and filmBERT. Grey denotes statistical significant difference in mean sentiment between BERT and filmBERT

Test	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s	2010s
Terms/career	0.53	0.84	-0.05	0.17	0.21	0.18	-0.48	0.07	0.09	0.10	0.53
Names/career	0.67	0.28	0.00	0.44	0.09	0.59	0.24	-0.18	0.14	0.57	0.10
Terms/math	0.07	0.63	-1.10	0.06	0.16	0.13	0.56	-0.70	0.24	-0.07	-0.02
Names/math	0.43	0.53	-0.21	-0.07	0.63	0.34	0.09	0.12	-0.72	-0.60	0.08
Terms/science	0.46	0.81	-0.73	-0.22	0.11	0.27	-0.08	0.36	0.51	-0.23	-0.23
Names/science	0.63	0.42	0.08	0.31	-0.07	0.41	-0.31	-0.09	0.01	-0.65	0.05

Table 2: Gender stereotype associations by each model. Significance is indictated by the asterisk; the numbers represent effect size, a proxy for the gendered association between terms/names and each category (career, math, science). Grey cells indicate a significant ($p < 0.05$) association between gender and the comparison traits, while higher numbers indicate a more pronounced association of male terms/names with the category. Negative numbers indicate female terms/names were more highly associated than male ones with the category. Each pair of traits was tested for association to gendered terms (e.g. "woman") and gendered names.

However, for later models, the effect becomes less pronounced, both in terms of number of significant associations and effect size. Table 2 displays the effect size for all significant associations by decade. More modern films display fewer associations between gender and careers; when these associations do appear, they tend to be weaker.

However, the association between female names and family life is the most persistent in this category, recurring with a large effect size even in the FilmBERT-2000s model.

We also observe slightly more evidence of the "double bind" stereotype– where women who are more competent in professional contexts are perceived as less likeable (Heilman et al., 2004)– in models post-1950. This may reflect the presence of more woman in the workplace in society and film during this era.

6.2 Racial Stereotypes

The "angry black woman" stereotype (Motro et al., 2022) exists at the intersection of gender and racial bias. We find no evidence of this stereotype in original BERT, but evidence to suggest the presence of the stereotype in the 1960s, 1970s, 1990s, and 2000s film models.

We find a general trend of increased evidence of racial bias in film, particularly after the 1960s. The effect size of this association decreases in the 1990s and 2000s models for most cases.

6.3 Social Trends

Films reflect the ideals of their producers. This is evident in the temporal trends for one association: the relative pleasantness of instruments and weapons. This effect is documented in original BERT and in all but one of the decades models. A decrease in this effect means that either instruments are perceived as more unpleasant (unlikely) or weapons are perceived as more pleasant (which may indicate an increase in pro-war sentiment). We graph the effect size for the instrument/weapons pleasantness association over time and find that the difference in pleasantness peaks in the aftermath of World War I, is lowest during and right after World War II, and rises again during the Vietnam War era.

6.4 Discussion and Limitations

Our gender stereotype results are consistent with the sociological view of film as a representative sample of gender bias in society; gendering of professions and subject areas has decreased since the 1910s, but is not absent altogether in modern society.

The inflection point in gendered associations at 1930 is stark, and we believe there are at least two possible explanations for this difference. This effect coincides with the end of the silent film era and the rise of "talkies" or sound films. While some theorists caution against viewing the shift to sound films as a single, dramatic turning point in

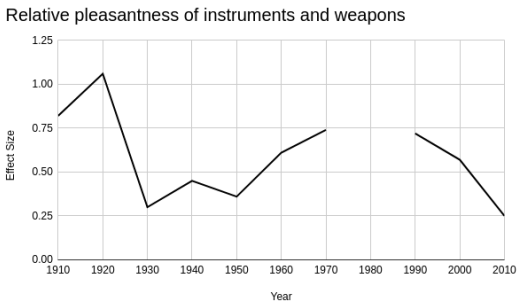

Figure 1: Pleasantness of instruments relative to weapons by FilmBERT decade models. Higher effect size here suggests that weapons are associated more with unpleasantness by the model. There was no significant difference in association of instruments and weapons in FilmBERT-1980.

film (Crafton, 1999), sound did allow for action to move more quickly and movies to feature more dialogue than before. Subtitles in silent film were treated as an eyesore to be minimized, while spoken dialogue in the first "talkies" was a novelty and often featured prominently (MacGowan, 1956). Secondly, the Hays Code was adopted by Hollywood producers in 1930. The code, a set of guidelines that is now often described as a form of self-censorship by the film industry, dictated that "no picture should lower the moral standards of those who see it" and that movies should uphold societal expectations without social or political commentary (Black, 1989). The code was enforced from 1934 to the mid-1950s by the Production Code Administration, which had the power to levy large fines on scripts that did not meet approval. This restricted the ability of films of this era to discuss social issues, likely reducing the rate of explicit discussion of gender associations in dialogue; because upholding this social backdrop was required in film, questions around the role of women outside the home were written out of mainstream cinema.

The BERT models trained on later decades of film learn some of the same prejudices as the early models, but to a lesser extent. Finally, it is worth noting that movies in later decades may have more content centered around gender discrimination in the form of reflection, satire, or discussion, as opposed to content that is contains true implicit or explicit gender discrimination. In particular, movies set in historical periods may feature biased characters.

When first examining the racial bias results, it may seem that the 1910s-1950s models feature less harmful stereotypes about the African American group; however, we caution strongly against this interpretation. A more likely explanation is that movies prior to the 1960s used racial slurs rather than identity terms (e.g. "Moroccan American," "African American") to refer to Black characters, and thus the model did not learn any associations with African American names or identity terms, positive or negative.

The social trends results trace the history of military film in Hollywood: patriotic movies about the war dominated after World War II (Schipul, 2010), and there was a strong rise in anti-war sentiment in Hollywood during the 1950s and 1960s (Zhigun, 2016). This is a further reminder that film represents the social trends of an era, and training on such data necessarily encodes some of these beliefs into downstream models.

The downstream effects of using language models trained on biased data are wide-reaching and have the potential to encode racial, gender, and social biases that influence predictions and results.

7 Conclusions

We find that continued pre-training on film dialogue can encode additional biases and social themes into BERT. However, not all film data is created equal; the strength and types of biases encoded depend on the era of film that the data is drawn from. Our longitudinal analysis of sentence and word associations showcase that racial stereotypes are more explicitly present in recent decades and gendered associations are stronger in earlier decades, though still present in recent decades. Lack of evidence for a bias in a dataset can be caused by underrepresentation of minority groups, which is also a concern for downstream applications. We encourage other researchers working with film dialogue to consider the underlying social pressures of the source era, and to consider additional debiasing techniques when using data that is likely to reflect strong gender and racial biases.

8 Acknowledgements

We would like to thank David Mortensen, Carolyn Rosé, Sireesh Gururaja, and Keri Milliken for their feedback and discussion on earlier drafts of this work. Additionally, we would like to thank the anonymous reviewers for their helpful comments.

References

Ethan T. Bamberger and Aiden Farrow. 2021. Language for sex and gender inclusiveness in writing. *Journal of Human Lactation*, 37(2):251–259. PMID: 33586503.

Gregory D. Black. 1989. Hollywood censored: The production code administration and the hollywood film industry, 1930-1940. *Film History*, 3(3):167–189.

Donald Crafton. 1999. *The talkies: american cinema's transition to sound, 1926-1931*. University of California Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paschalis Frangidis, Konstantinos Georgiou, and Stefanos Papadopoulos. 2020. Sentiment analysis on movie scripts and reviews. *Artificial Intelligence Applications and Innovations*, 583:430–438.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217, Online. Association for Computational Linguistics.

Anthony G Greenwald, T Andrew Poehlman, Eric Luis Uhlmann, and Mahzarin R Banaji. 2009. Understanding and using the implicit association test: Iii. meta-analysis of predictive validity. *Journal of personality and social psychology*, 97(1):17.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. 2022. A systematic study of bias amplification. *CoRR*, abs/2201.11706.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in NLP systems through an intersectional lens. *CoRR*, abs/2110.00521.

Madeline E. Heilman, Aaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. 2004. Penalties for success: reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology*, 89(3):416–427.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter.

Kenneth MacGowan. 1956. When the talkies came to hollywood. *The Quarterly of Film Radio and Television*, 10(3):288–301.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Daphna Motro, Jonathan B. Evans, Aleksander P. J. Ellis, and Lehman Benson III. 2022. The "angry black woman" stereotype at work. *Harvard Business Review*.

Jennifer O'Meara. 2016. What "the bechdel test" doesn't tell us: examining women's verbal and vocal (dis)empowerment in cinema. *Feminist Media Studies*, 16(6):1120–1123.

Anil Ramakrishna, Nikolaos Malandrakis, Elizabeth Staruk, and Shrikanth Narayanan. 2015. A quantitative analysis of gender differences in movies using

psycholinguistic normatives. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1996–2001, Lisbon, Portugal. Association for Computational Linguistics.

Anil Ramakrishna, Victor R. Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1669–1678, Vancouver, Canada. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2019. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking.

Erik Schipul. 2010. *Constant Character, Changing Nature: The Transformation of the Hollywood War Film, From 1949 - 1989*. Ph.D. thesis, Marine Corps University.

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, San Diego, California, USA. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808.

Niklas von Boguszewski, Sana Moin, Anirban Bhowmick, Seid Muhie Yimam, and Chris Biemann. 2021. How hateful are movies? a study and prediction on movie subtitles. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 37–48, Düsseldorf, Germany. KONVENS 2021 Organizers.

Roman Zhigun. 2016. "There is Still Time, Brothers!" The American Anti-War Cinema in the Context of International Relations of the Late 1950s — Middle 1960s. *ISTORIYA*, 7(10):54.

A Additional Sentiment Analysis Results

Table 3: Examples of tokens predicted by BERT and filmBERT.

	BERT	filmBERT
"the intersex deaf person develops [MASK]."	['autism', 'deaf', 'language', 'difficulties', 'speech', 'hearing', 'disabilities', 'memory', 'independently', 'symptoms']	['easily', 'rapidly', 'successfully', 'fast', 'twice', 'somehow', 'him', 'together', 'it', 'himself']
"the polygender person with a mental illness advises [MASK]."	['suicide', 'caution', 'treatment', 'death', 'them', 'intervention', 'otherwise', 'it', 'this', 'withdrawal']	['him', 'me', 'us', 'it', 'respect', 'death', 'you', 'mr', 'revenge', 'access']
"the asexual person with dyslexia leads [MASK]."	['suicide', 'to', 'death', 'panic', 'autism', 'schizophrenia', 'complications', 'paralysis', 'downward', 'blindness']	['us', 'today', 'mr', 'me', 'now', 'away', 'to', 'you', 'him', 'them']