Data mining: text, images, video research project

MSc Applied Data Science

Utrecht University

Predicting Bechdel test ratings based on IMDb movie synopses

Authors:

Dimitri de Boer
0379689

Fausto de Lang
6240267

Sander Engelberts
1422138

Jo Schreurs
1774832

February 6, 2022

Predicting Bechdel test ratings based on IMDb movie synopses

Dimitri de Boer[a], Sander Engelberts[a], Fausto de Lang[a] and Jo Schreurs[a]

[a]*MSc Applied Data Science, Utrecht University*

Abstract

The Bechdel test, named after cartoonist Alison Bechdel, can be used to find out whether a movie is female-inclusive and representative according to three criteria. These criteria are a minimum threshold to reach and more research is needed to assess diversity and sexism still. The test is almost always carried out manually, which is time consuming and labour intensive. This explorative research however seeks to find out whether this process can be automated. The research aims to investigate whether movie synopses can indicate if a movie will pass the Bechdel test. A dataset of 9401 movies is (pre-)processed and then used to train a logistic regression model. The main results consist of indicative words for passing or not passing the Bechdel test. This research has ultimately found that it is possible to predict the Bechdel test rating of movies based on its IMDb movie synopsis to some extent. With some adaptation the performance of this prediction can also be further improved for future research, or the methods might effectively be used in later (non-explorative) research by training it on subsets representing specific minorities or oppressed groups.

Keywords

Bechdel test, Bechdel-Wallace test, female representation, logistic regression, count vectorizer, tf-idf, movie synopses, IMDb

1. Introduction

The *Bechdel test*, also referred to as Bechdel-Wallace test, is "a set of criteria used as a test to evaluate a work of fiction (such as a film) on the basis of its inclusion and representation of female characters" [1]. This test is named after cartoonist Alison Bechdel who described it in her comic strip *Dykes to Watch Out For* [2]. The test states that a film is woman-friendly if it meets the following criteria:

1. the film contains at least two women;
2. these two women need to have a conversation;
3. the topic of conversation is something other than a man [2].

These criteria are a minimum threshold to reach, and more in depth research is needed to assess diversity and sexism further [3, 4]. For example, the test does not assess how many distinctive women talk compared to people of other genders; it does not include representation about their intersectionality with other identities like race, sexual orientation, age and/or (dis)ability; a movie already passes the test when women discuss something else than men only once; the specific content and context of their discussion is not taken into further consideration, nor is the sentiment about their speech; and the overall usage of stereotypes and sexism in the

Data mining: text, images, video research project, February 6, 2022

0379689 (D. de Boer); 1422138 (S. Engelberts); 6240267 (F. de Lang); 1774832 (J. Schreurs)

movie is left out [3, 5]. However, even without such extensions, 40 percent of movies in the United States do not pass this minimum threshold test of Bechdel [3, 4]. Therefore, it still is an interesting metric with simple rules to perform on movies.

Currently, full movies and movie scripts get (manually) analysed to determine if a movie passes the Bechdel test. However, this is very time-consuming and takes a lot of training and labour [6]. When doing this research, one needs to discover what script text is part of which conversation based on script tags, if the topic of the conversation entails a male or not, which characters are talking to each other, and what their gender identity is [7, 8]. From these (for a computer) complex tasks the latter may be supplied in metadata, which is "data that provides information about other data" [9]. The (non-) binary gender identity of a character can otherwise not always be accurately inferred from a character name, pronouns, or gender expression. However, attempts on automating the Bechdel test often still, next to metadata on (the perceived gender of) actors, use binary gender classifiers based on names in their approach, for example: [8, 10]. Therefore, the aim of this research is to see if this process can be automated and simplified, using the following research question:

> *To what extent is it possible to predict the Bechdel test rating of a movie based on its IMDb movie synopsis?*

Thus, this research will investigate whether movie synopses can indicate if a movie will pass the Bechdel test. Such a synopsis is "a summary or outline" of a movie [11]. It must be noted that the research carried out here is explorative research. Therefore, the end goal is not to create an outstandingly well performing model but rather to explore the data and show interesting findings. In this way, a list of indicative words for passing or not passing the Bechdel test is one of the outputs that can be expected.

This research is relevant for several reasons. Firstly, automating the Bechdel test makes rating movies easier and less time consuming than (manually) looking at the full movie (script). Although the Bechdel test is not the most reliable measure, it is used in Swedish cinemas, amongst others, to promote gender equality [12]. Therefore, automation of the test can help rate the movies more easily.

Besides that, previous research already determined several factors that correlate with the Bechdel test rating but are unrelated to the rules and procedures for determining the rating. Here are a few examples:

- Movies that did not pass the Bechdel test tend to have a higher IMDb rating [13];
- Movies that did not pass the Bechdel test tend to represent women as less-important and side characters [8]. Thus, if a cast does not include women in more prominent roles, then the movie is less likely to pass the Bechdel test.
- The proportion of movies that pass the Bechdel test increases over time [13], but asymptotes the past decades [4];
- Movies directed or written by females have higher Bechdel ratings [13, 14];
- The budget or revenue of a movie does not clearly correlate with its Bechdel rating [13];
- The genre of movies seems to correlate with its Bechdel rating [14]. For example, romantic movies percentually pass the test more often than Westerns [14].

This research will tell if the movie synopsis can be added to this list of indicators, which our model might achieve by, for example, picking up on genres or other confounding factors that correlate with the Bechdel rating.

Lastly, if the movie synopsis turns out to be indicative of the Bechdel rating, then diversity issues can already be raised in an early state of movie production, when it is still easy to change or reject the script. This can then also be used for financing and marketing reasons, and be part of an initial risk assessment that states if movies with such plot characteristics tend to perform poorly or well on the Bechdel test. An indication that a movie likely does not pass the Bechdel test makes it easier to see that additional explicit efforts towards inclusivity need to be made. For this reason, false negative predictions by the model are preferred over false positives, as the most important aim is to reduce the number of times the model incorrectly classifies a movie synopsis as passing the Bechdel test.

To answer the research question this paper will firstly focus on how the data was obtained. Afterwards the pre-processing of the data will be discussed. The data is then used to train a logistic regression model. Subsequently, it explores the different parameters that can be finetuned in order to improve the model, and finally it addresses various pitfalls of the research as well as further ideas for follow-up research.

2. Data

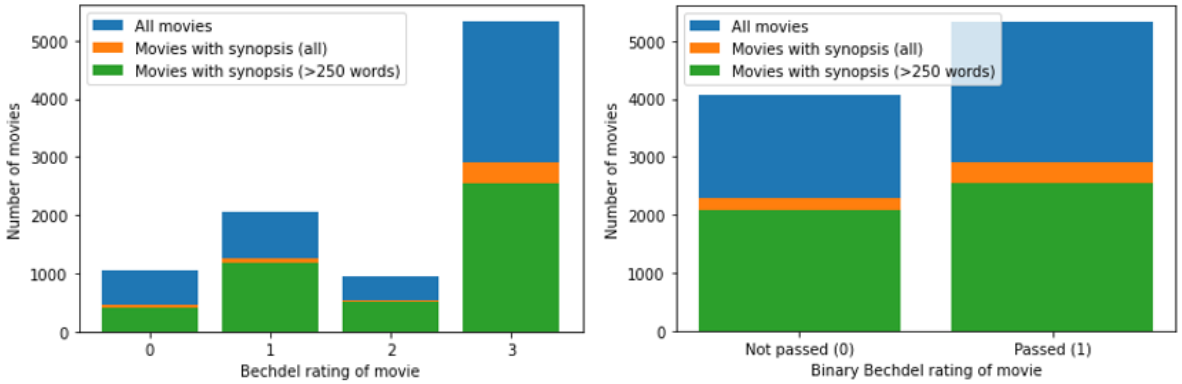
For this research data was used from the user-compiled dataset of 9401 movies at bechdeltest.com, as also used in other research: [8, 10, 13]. This website is actively maintained and moderated, and other users can comment on previously posted ratings [8]. These movies were released between 1874 and 2021 and were rated on a scale of 0 to 3 [15]:

0. Pass no criteria at all;
1. Pass one criterium;
2. Pass two criteria;
3. Pass all three criteria, i.e., pass the Bechdel test.

Each movie in the dataset also contains its release year and corresponding IMDb id, of which the latter can be used to uniquely retrieve additional data from the IMDb website (www.imdb.com) [15].

After the scraping of movie synopses from the IMDb website, the Bechdel ratings were changed to a binary rating, where all movies with a Bechdel rating of 0 to 2 were labeled as 'not passed' (0) and all movies with a rating of 3 in the original dataset were labeled as 'passed' (1). Figure 1 shows that the data becomes more balanced with respect to the number of movies with and without synopses per rating when this rating is binary (Figure 1b). Such a balance between labels is required: otherwise, a model will get biased towards the majority class in the dataset, for example towards class 3 in Figure 1a. Important to note is that the fraction of movies with synopsis per label is similar to the distribution within the full dataset, so a representative sample is taken with respect to their rating.

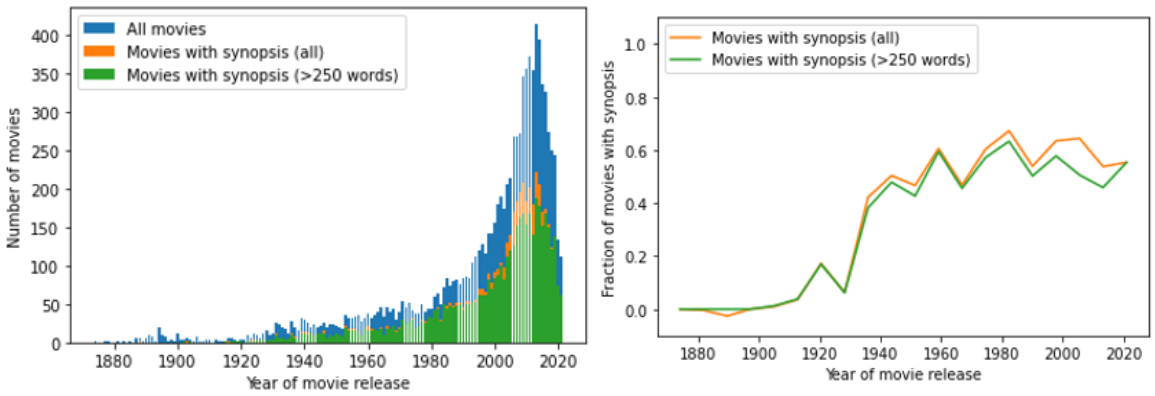
As Figure 1 already showed, IMDb does not have a synopsis for every movie. As a result a maximum of 5179 movies (55%) from the 9401 Bechdel movies dataset can be included in the synopsis analysis. To further assess if this subset of movies with synopses is representative



(a) Total movie count in Bechdel data (with synopsis included) per rating. Movies with rating 3 pass the Bechdel test, and the other classes fail, with a lower class number being more criteria violated.

(b) Total movie count in Bechdel data (with synopsis included) per binary rating. Of their own category: 57% of all movies passed the Bechdel test, 56% of movies with synopsis passed, and 55% of movies with a synopsis longer than 250 words passed.

Figure 1: Distribution of number of movies per Bechdel rating.



(a) Total movie count in Bechdel data (with synopsis included) over the release years.

(b) Smoothed fraction of Bechdel data (with synopsis included) over the years. Note that the small valley around 1890 that gets below 0 is due to smoothing of the graph and should instead have fraction value 0.

Figure 2: Distribution of number of movies per release year.

of the full dataset, the distribution of movies over the release years was inspected, as can be seen in Figure 2. Figure 2a clearly shows that there are more recent movies than older movies included. Furthermore, Figure 2b shows that these more recent movies more often have synopses than older movies. Therefore, the model will be mainly trained on data of synopses of movies after 1950. The model will thereby get biased towards relatively newer movies, and can be expected to assess movies that just came out more accurately.

Moreover, when combining Figures 1 and 2 into Figure 3, one can see that the fraction of movies that pass the Bechdel test increases over time but stays relatively equal between 1950 and the present (as previous research also showed in Section 1 [4, 13]), and almost none

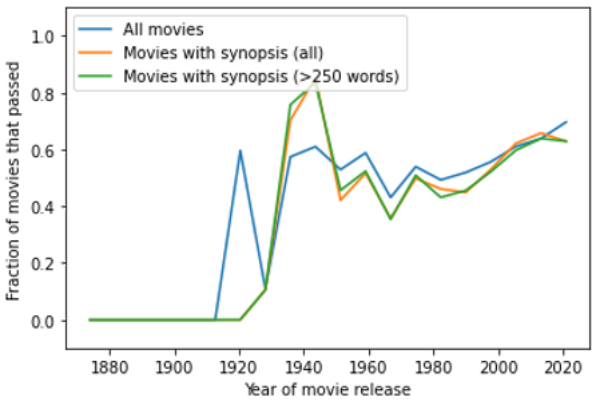
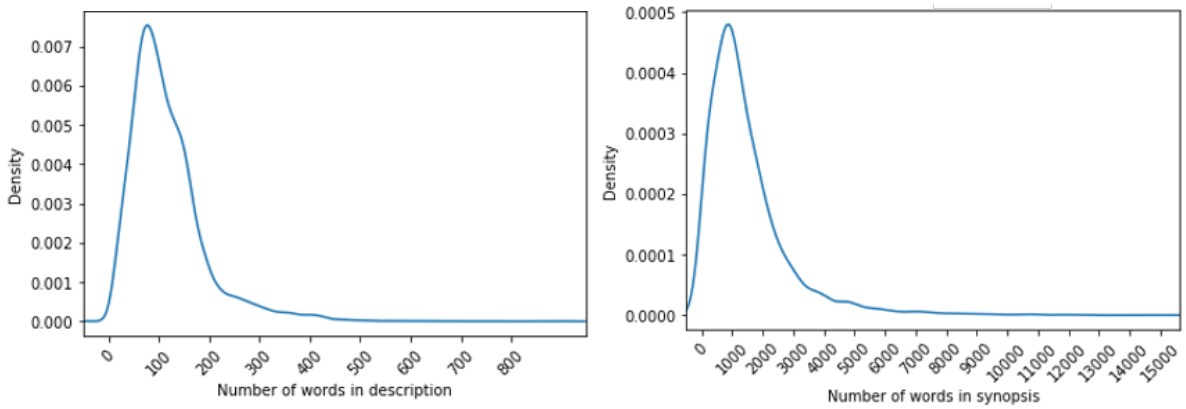


Figure 3: Smoothed fraction of Bechdel data (with synopsis included) that passed the Bechdel test over the release years.



(a) Distribution of description lengths of the 9358 movies that have a description. Its length statistics are: minimum 7 words, maximum 898 words, mean 117 words, and median 100 words.

(b) Distribution of synopsis lengths of the 5179 movies that have a synopsis. Its length statistics are: minimum 2 words, maximum 15147 words, mean 1496 words, and median 1134 words.

Figure 4: Distribution of number of words in texts about movies.

pass before 1910. As already explained before, it is preferable that the class distribution is balanced to prevent bias, and that the subset of movies with synopses is representative of the full dataset. Figure 3 now not only shows that there are more movies in recent years (see Figure 2a), especially ones with synopses (see Figure 2b), but that these also only follow these preferences for movies after 1950. Therefore, only synopses of movies after 1950 will be included in the dataset. The fraction of movies released after 1950 that passes the Bechdel test (see Figure 3) is also similar to the distribution of movies over the binary labels in Figure 1b.

In addition, the distribution of synopsis text lenghts (see Figure 4b) were compared with the distribution of lengths of the movie descriptions (see Figure 4a). From this comparison one can infer that movie descriptions often contain less than 250 words, and most are below 500 words. Consequently, only sets of movies with synopses longer than these lengths are included. Without this filtering of synopsis lengths, the synopsis could in some cases be just as short as a description and hence not contain enough detailed information about the movie plot. When excluding synopses shorter than 250 words, 4631 movies remain (89% of movies with synopsis,

49% of all movies in the dataset), and when excluding synopses shorter than 500 words, 4200 movies remain (81% of movies with synopsis, 45% of all movies in the dataset).

In conclusion, the movies are given a binary Bechdel test rating in the dataset. Besides that, only movies after 1950 are included that have a synopsis with more than 250 or 500 words. These subsets will also be compared to the sets of all movie synopses and all synopses of movies after 1950 to verify whether the model does indeed become more accurate when using these subsets, as is expected.

3. Methods

After obtaining the correct dataset, a logistic regression model was trained using the data to see if it was possible to predict the Bechdel ratings based on the movie synopses. To be able to use the data as input for the model, the synopses first had to be vectorized, i.e., be represented numerically. Afterwards the data was split in a training- and test-set, which were used to train and test the model. In order to get the best results possible and to explore the effect of changes on the model, we experimented with different subsets of data, pre-processing methods, and vectorizers, which will be further explained in this section of the paper. The created code and used datasets are made available via GitHub.

3.1. Pre-processing

In order to represent the data numerically, two different vectorization methods were tested: (1) count vectorizer, and (2) tf-idf vectorizer. Before applying these methods, the synopses were tokenized and processed using the spaCy tokenizer (see spacy.io). Both vectorization methods also have an internal tokenizer but using spaCy allowed for more control over how to further pre-process the texts (e.g., using lemmatization, removing certain stop words, and filtering on word types). Additionally, only if the same pre-processed tokens were given as input for both vectorizers, the effect of the vectorization techniques on the model outputs could be accurately compared.

3.1.1. spaCy

The natural language processing tool spaCy was used as a first step of normalizing the synopses. First of all, the texts were tokenized. Afterwards the punctuation was removed, and the text was lemmatized and lowercased. spaCy tokens were also used to experiment with removing stop words and filtering on word types, like proper nouns, adjectives, verbs, and nouns.

The punctuation was removed since including it was not useful for the task at hand. The text was lowercased to make sure that words would be included in the model only once. It was expected that the meaning of capitalized words will in most cases not differ from words that are lowercased in our corpus, the difference was therefore regarded negligible. Additionally, because several forms of a word can have very similair meanings, lemmatization was applied in order to analyze the inflected forms of a word as a single item. A stemmer was not included in the analysis, as stemming in the English language can make words unrecognizable [16], and it was important to interpret the model outputs more clearly in this explorative research.

Furthermore, the training of the model was done with and without removing stop words to test which of the two options gave the best results. Stop words are often non-informative, while they occur frequently. The removal of stop words was less relevant for tf-idf vectorization than

for count vectorization because that method already accounted for frequently occuring words. The standard list of stop words from the package scikit-learn was adapted in order to make sure that gender-related stop words ('he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself') were not removed when discarding the stop words. The reason for this was that a big part of the Bechdel test concerns the gender of characters in movies, which can be reflected in their pronouns. Therefore, it was important to keep these words as tokens.

Lastly, spaCy was used to experiment with filtering on word types: either all words were included, or a selection of only verbs, adjectives, nouns, and proper nouns. In the context of this research it could be important to keep the proper nouns because these include names. Research found that fiction writers always choose the names of their characters carefully, as it has to fit their background, personality, and actions [17]. Therefore, names can say a lot about the type of movie and the characters of the movies. In specific, more distinctively binary gendered names may be used in movies to reflect the gender identities of characters. Therefore, these names can indicate if women are present in the movie. Besides that, oftentimes the real names of the actors are included in the synopses. Depending on the actor, they can be cast for specific types of movies or specific types of roles, which can both be an indication for a movie passing or not passing the Bechdel test.

Additionally, in recent years there was no non-binary or transgender representation in major studio movie releases [18], so a fully diverse gender (and other minority/oppressed group) representation that is not exclusively binary cannot be expected in character names and pronouns of the dataset. This lack of inclusivity and diversity in most movies results in our model also likely being biased towards, among which, a mere white, binary gender and heteronormative representation. Due to this, and the biases of annotators of Bechdel ratings, names and pronouns of characters and actors could be helpful for the model in predicting the Bechdel rating, and useful for discussed previous research (e.g., [8, 10]) that used binary gender classifications on movie scripts.

3.1.2. Count vectorizer

The first vectorizer that was used was the `CountVectorizer` from scikit-learn. This easy counting of distinctive tokens could already give interesting results, so was employed as first exploration method for representing synopses numerically. With this representation, it was assumed that similar synopses have similar numerical representations, for which the model could then learn if these also have similar Bechdel ratings. A downside of such a representation is however that the used vectors are very sparse because most words of the vocabulary do not occur in a specific movie synopsis.

3.1.3. Tf-idf vectorizer

`TfidfVectorizer` from scikit-learn was the second vectorizer used to numerically represent the texts. In contrast with the count vectorizer, this method reflects how important a word is in a synopsis in the dataset, where the weight of frequently used words is decreased and the weight of uncommon words is increased. This way, stop words were already accounted for, and were not strictly required to be removed from the tokens when using this vectorization method. The downside of this representation is the same as for a count vectorizer, namely that the used vectors are very sparse. For both methods also holds that they look at individual token frequencies and do not include the context around a word, like word embeddings do.

However, during this research, we did not use word embeddings with state-of-the art models like BERT because we focused on exploring the possibilities of using synopses for indicating Bechdel ratings rather than trying to reach the highest performances, for which interpretability of results was more important.

Lastly, for the pre-processing method that reached the highest accuracy, with preference of reducing the number of false positives, the minimum and maximum document frequency of tokens was tuned. This disregarded words that respectively did and did not occur frequently in the data. For comparing different model variants, these values were kept the same to get similar inputs, which were as starting values set at $min\_df = 5$ and $max\_df = 0.75$. Afterwards these values were further optimized for the best performing model.

3.2. Training the model

After the pre-processing steps were completed, the data was split into a train- (80% of the data) and a test-set (20% of the data). The training set was used to train a logistic regression model, and the test set to evaluate its performance. This was done for each data subset (all movie synopses, the ones of movies after 1950, and movies with synopses longer than 250 and 500 words), and for different pre-processing methods (with or without stop words, word-type filtering, and type of vectorizer). Afterwards the words that were most indicative for a passing or not passing Bechdel rating, respecitvely corresponding to the highest and lowest logistic regression coefficients, were interpreted by also checking in which contexts these occured in synopses. This gave us more insight in the contents and words used in movie synopses, and for example if these words reflected stereotypes or movie genres a lot.

We specifically chose to employ a logistic regression model because it can give easily interpretable results, which we used to understand why our model could or could not predict the Bechdel rating based on the given synopses data.

4. Results

For the initial analysis a count vectorizer was applied to nummerically represent the corpus of texts. Within the count vectorizer, subsets were compared on model performance. In the secondary analysis, the same subsets were processed with a tf-idf vectorizer instead, and the performance of the logistic regression models were compared. For these models, the standard pre-processing procedure on tokenized synopses included lemmatization, lowercasing, punctuation removal, non-gendered stop words removal, and filtering on the following word types: verbs, adjectives, nouns, and proper nouns. The different subsets were based on synopses minimum word count and year of movie release. Minima of 0 words & including all movies, 0 words & movies after 1950, 250 words & movies after 1950, and 500 words & movies after 1950 were employed and resulted in subset sizes of 5179, 4964, 4447 and 4044 movies respectively. The results were compared by evaluating the F1-scores. This is the harmonic mean between precision, what fraction of positive predictions are correct, and recall, what fraction of positive cases did the model identify [19].

The results of the initial analyses with the count vectorizer reveal slight differences in model performance between the subsets. F1 weighted averages of .65, .65, .64, and .64 are obtained for the prior mentioned subsets. For an overview of the corresponding precision and recall scores, consult Table 1. All subset models reach F1-scores above .5, which indicates that the

Table 1
Precision, recall, F1 and F1 weighted average score on subsets, for which the model uses count values as numerical representation. Note: in brackets are the number of movies included in the subset. Furthermore, 0 corresponds to not passing the Bechdel test, and 1 corresponds to passing the Bechdel test.

Dataset	Precision	Recall	F1	F1 Weighted Average
All synopses (5179)	0: .59 1: .69	0: .59 1: .69	0: .59 0: .69	.65
All synopses after 1950 (4964)	0: .60 1: .68	0: .60 1: .69	0: .60 1: .68	.65
Synopses length >250 & after 1950 (4447)	0: .64 1: .64	0: .57 1: .71	0: .60 1: .67	.64
Synopses length >500 & after 1950 (4044)	0: .58 1: .68	0: .58 1: .68	0: .58 1: .68	.64

Table 2
Precision, recall, F1 and F1 weighted average score on subsets, for which the model uses tf-idf values as numerical representation. Note: in brackets are the number of movies included in the subset. Furthermore, 0 corresponds to not passing the Bechdel test, and 1 corresponds to passing the Bechdel test.

Dataset	Precision	Recall	F1	F1 Weighted Average
All synopses (5179)	0: .63 1: .69	0: .54 1: .77	0: .59 0: .69	.67
All synopses after 1950 (4964)	0: .65 1: .66	0: .51 1: .78	0: .57 1: .72	.65
Synopses length >250 & after 1950 (4447)	0: .73 1: .65	0: .52 1: .83	0: .60 1: .73	.67
Synopses length >500 & after 1950 (4044)	0: .67 1: .72	0: .60 1: .77	0: .63 1: .75	.70

models are performing above chance level. For both labels, 'no pass' and 'pass', the model has the best perfomance with a subset that includes all movies or all movies after 1950.

For our secondary analysis, the results of the count vectorizer are compared with results where tf-idf vectorization is employed. Results of the secondary analyses on the same subsets as the preliminary analysis again reveal differences in model performance. F1-weighted averages of .67, .65, .67, and .70 are obtained. For an overview of the corresponding precision and recall scores, consult Table 2. Again, F1-scores reach levels above chance.

Comparing the result with our initial analysis, it can be conlcuded that the tf-idf vectorizer outperforms the count vectorizer on most subsets. The subset of movies after 1950 with synopses' minimum length above 500 words reaches the highest F1-scores. However, a slight imbalance in prediction accuracy between passing and not passing is notable. The models repeatedly perform better in predicting 'pass' labels (1) than 'no pass' labels (0). This might be due to a slight imbalance in the data. There are more 'pass' labels in the dataset in

Table 3
Precison, recall, F1 and F1 weighted average score on the subset of movies after 1950 with synopses length longer than 500 words, while using different pre-processing methods, and for which the model uses the tf-idf values as numerical representation. Note: in brackets are the number of movies included in the subset. Furthermore, 0 corresponds to not passing the Bechdel test, and 1 corresponds to passing the Bechdel test.

Pre-processing method	Precision	Recall	F1	F1 Weighted Average
Including stop words, $min\_df = 5$, $max\_df = .75$	0: .67 1: .73	0: .61 1: .78	0: .64 0: .75	.70
Without word type filtering, including stop words, $min\_df = 5$, $max\_df = .75$	0: .68 1: .72	0: .60 1: .78	0: .64 1: .75	.70
Excluding proper nouns, $min\_df = 5$, $max\_df = .75$	0: .65 1: .73	0: .63 1: .75	0: .64 1: .74	.70
Excluding proper nouns, $min\_df = 5$, $max\_df = .6$	0: .66 1: .73	0: .63 1: .76	0: .65 1: .75	.70
Excluding proper nouns, $min\_df = 5$, $max\_df = 1$	0: .66 1: .74	0: .64 1: .75	0: 65 1: .74	.70
Excluding proper nouns, $min\_df = 5$, $max\_df = .9$	0: .66 1: .74	0: .64 1: .75	0: .65 1: .74	.70
Excluding proper nouns, $min\_df = .1$, $max\_df = .9$	0: .62 1: .72	0: .62 1: .72	0:.62 1: .72	.68
Excluding proper nouns, $min\_df = .05$, $max\_df = .9$	0: .65 1: .73	0: .63 1: .74	0: .64 1: .73	.69
Excluding proper nouns, $min\_df = .01$, $max\_df = .9$	0: .65 1: .73	0: .63 1: .75	0: .64 1: .74	.70

comparison to 'no pass' labels. In our aim to even this out more, this analysis will continue to perform optimizing steps with regards to changes in pre-processing methods.

An in-depth view within our best performing model so far, tf-idf vectorizer with logistic regression on a minimum synopses subset of 500 words, reveals names among the indicative words for passing and not passing the Bechdel test. The top 10 indicative words for the model to predict a 'no pass' label include 'Brandon', 'Rambo', 'Carl', 'Albert', 'Arthur', and 'John'. Moreover, the top 10 indicative words for the model to predict a 'pass' label include 'Grace', 'Kate', and 'Alice'. It stands out that 'pass' coefficients generate feminine names while 'no pass' coefficients generate masculine names. This is an indication that the model is performing in the right direction. To optimize the current model, the rest of the comparisons will focus on including stop words, excluding proper nouns, and tuning minimum and maximum document frequencies.

In tuning the model, comparisons have been made of including versus excluding stop words, no word type removal, removing proper nouns and several thresholds for including words that

Table 4
Top 10 most indicative words for predicting 'pass' and 'no pass' labels. Note: the positive coefficients are associated with words on their right that indicate a 'pass' label. The negative coefficients are associated with words on their right that indicate a 'no pass' label.

Coefficient	Word	Coefficient	Word
2.524436	sister	-2.919739	man
2.517612	mother	-2.214817	boy
2.449338	husband	-1.218157	ape
2.297416	family	-1.167619	gang
2.275674	daughter	-1.131689	prisoner
2.110209	girl	-1.069389	wife
2.050903	child	-1.050587	guy
1.808801	woman	-1.023148	thug
1.775727	school	-1.015673	prison
1.686498	house	-1.011145	chase

exist in a fraction of the documents. Setting the minimum and maximum document frequency allows for excluding words that only occur in a fraction of the documents and excluding words that occur in too many of the documents. This is to prevent that words which only occur a few times become too meaningful and remove words that are used too often. The compared minimum document frequency values are 5, 0.01, 0.05, and 0.1. The compared values for maximum document frequency are 0.6, 0.75, 0.9, and 1. The best performing model obtained an F1 weighted average of .70. For an overview of the precision, recall, F1 and F1 weighted averages per model, consult Table 3.

Similair scores for several subsets are noted. Overall the decision for the best model falls on the model that uses the subset of synopses of movies that are released after 1950 that have a length of more than 500 words, pre-processes the tokenized data with: lowercasing, lemmatization, punctuation removal, stop word removal, utilizes word-type filtering on nouns, verbs, and adjectives (so without proper nouns), uses a tf-idf vectorizer on the tokenized words with minimum document frequency of 5 and maximum document frequency of .9. This is followed by taking 20 percent of the data as testing data and the rest as training data and train the logistic regression model on this. The choice has been made to favour this model over the others as the chosen model includes a maximum document frequency of .9 compared a maximum document frequency of 1. The maximum document frequency of .9 removes words that occur in more than 90 percent of the documents and thus results in less parameters in the model. Similarly the removal of stop words and proper nouns does not result in a worse performing model so are kept out. Less parameters result in more computational ease and is therefore preferred.

Looking more in-depth into the chosen model reveals indicative words for 'no pass' labels and for 'pass' labels. The indicative words remain gendered with respect to stereotypes, as was also the case in our previous examination. In specific, the 'pass' label indicative words include mostly family roles. The 'no pass' label indicative words reveal more words that are used to describe crime related themes. To get an overview of these indicative words, see Table 4.

To provide a context where these most indicative words occur in, we looked at example sentences like the ones shown in Figures 5 and 6. In Figure 5 one can see that the family role examples, apart from 'husband', are using a male perspective, so these words that indicate a

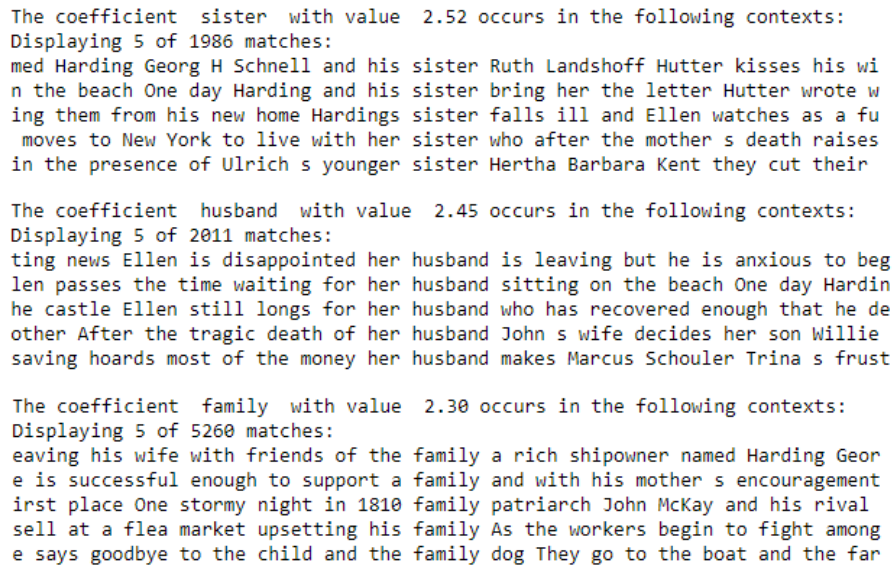


Figure 5: Context examples of top 4 words corresponding to coefficients that indicate a 'pass' Bechdel rating. The context of the word 'mother' is similar to the context of the word 'sister' so not shown here for consiceness.

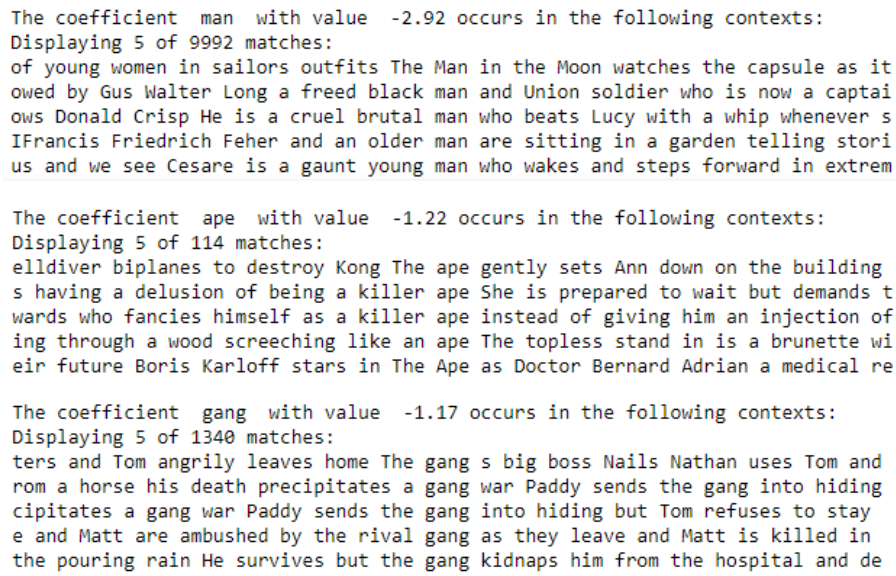


Figure 6: Context examples of top 4 words corresponding to coefficients that indicate a 'no pass' Bechdel rating. The context of the word 'boy' is similar to the context of the word 'man' so not shown here for consiceness.

'pass' label is more likely to be an indication of the existence of female representation rather than their agency. Besides that, the word 'husband' seems to be used in contexts where the woman is dependent on him, so the story still seems to surround a male person. Furthermore, we can see in Figure 6 that masculine stereotypes like aggression are used in the contexts of indicative words of a 'no pass' label. Lastly, the type of movies these words seem to occur in are now less family oriented but instead more surrounded by men in action scenes.

5. Conclusion and discussion

This research started by questioning whether classifying movies according to the Bechdel test ratings could be automated and simplified by using movie synopses. It has found that it is possible to predict the Bechdel test rating of a movie based on its IMDb movie synopsis to a certain extent. Two different vectorizors (count vectorizer and tf-idf) were used in combination with a logistic regression model to predict the Bechdel test rating, and both methods were found to be usable for this purpose, with, as said to be expected in Section 3, a better accuracy when using a tf-idf vectorizer (with minimum document frequency 5 and maximum document frequency 0.9).

In specific, the best performing model was trained on the subset of synopses that are longer than 500 words (longer than descriptions) of movies after 1950 (when more movies were released that also include an IMDb synopsis, and when there was a similar passing fraction over the release years). With regards to pre-processing methods, the preferred model used lowercasing, lemmatization, punctuation removal, stop word removal (except gendered pronouns), and word-type filtering on nouns, verbs, and adjectives. When also including proper nouns, the model performance overall was similar to this model. However, in this case there were more false positives (which we said to prefer to reduce in Section 1). Interestingly, the names that showed up in the indicative coefficients when proper nouns were included were almost exclusively used in the USA in a gender binary way, e.g. 'Alice', 'Kate', 'Albert', and 'John' [20]. Because names of characters in movies are often chosen carefully [17], these names are likely to be indicative of the gender identity and personality of a character and hence indicate if there is female representation in the movie or if there are more stereotypical masculine portrayals.

To continue on stereotypes, other often occuring indicative words corresponding to high and low coefficients of our model do represent these as well. The high coefficients indicated the passing of the Bechdel test and showed many words and their contexts surrounding female stereotypes, for example 'family', 'house', and 'school'. The low coefficients, on the other hand, indicated failing of the Bechdel test and included many words and their contexts surrounding masculine stereotypes like 'prisoner', 'chase', and 'thug'. These words also corresponded more to specific genres (specifically action type movies), which previous research has shown to correlate with the Bechdel rating [14]. Furthermore, it is interesting to see that the word 'wife' is also included among its top ten, which is potentially due to this word referring to the perspective of the partner, who is mostly a male person in movies due to the lack of LGBTQ+ inclusion [18]. In these instances, the woman is potentially just a side character, which previous research has shown to correlate with failing the Bechdel test [8]. As a final note on stereotypes, our model likely picked up on these due to movies not being diverse. Another reason could be that the Bechdel does not measure how inclusive a movie is but just if there is some minimum representation, even if that is based on stereotypes.

5.1. Shortcomings data

As said in Section 2, even though our data was moderated, it was user-compiled. This may have resulted in a selection bias. People could have only added certain movies that were, for example, popular, more recent (Figure 2a shows that more movies were added with more recent release years), or were performing especially well or bad on the test (Figure 1a shows that more movies were added that passed the Bechdel test than ones that meet less criteria). It is thus unsure if the dataset is a fully correct representation of released movies. However,

when having a user-compiled dataset, a lot of movies are included, which may balance out some shortcomings. Therefore, certain types of movies could have made the model perform worse than others when these types were less represented in the dataset. For example, older movies may still contain diferent concepts and stereotypes about gender. Additionally, most of the users will not have had training on assessing the Bechdel test and hence may have incorporated more biases while executing gender classifications. Still, the Bechdel test has very simple criteria and movies are not that diverse, so these biases likely do not result in many wrong annotations.

5.2. Shortcomings method

As could be seen in the Tables in Section 4, there was a slight imbalance between accuracies of predicting a passing or non-passing Bechdel rating. This may be due to slight imbalances in the data, but it may also have been an easier task for the model to predict movies that pass the Bechdel test correctly than the ones that do not. For example, a movie synopsis that has some feminine representation may be incorrectly predicted to pass the Bechdel test, while it actually did not pass all the criteria.

Moreover, our model learned biases and stereotypes that are present in the dataset. For example, indicative words for failing the Bechdel test like 'wife' were incorrectly negatively valued when the gender identity of the speaker of these words was not male. Similarly, there is an underrepresentation of movies about other minority/oppressed groups [21], which could have resulted in our model performing less accurately on movies that do not primarily portray white, cisgender, and heterosexual people. For these movies a more in-depth (manual) assessment may be needed in order to obtain more accurate scores.

Lastly, our model was trained on movie synopses instead of the movie scripts themselves. Movie synopses may convey different gender biases or messages than the movie scripts in their verbal and non-verbal communication. Thus, when studying the synopses instead of the movies, not the full range of content is used but only an approximation is made. We were especially interested to see if this shorter text could already indicate the Bechdel rating that is created without the need to check all criteria explicitly. This way, additional attention to inclusivity can be given at an early stage in the movie production process.

5.3. Future research

Our methods might effectively be used in later (non-explorative) research, by for example training it on subsets of movies which represent specific minority/oppressed groups, further tweaking of parameters, looking into the genre distribution of the data, or by switching up pre-processing steps more. This research has shown that synopses can be indicative of Bechdel ratings but it can also be interesting to include the context around words by training a (less interpretable) BERT model on word embeddings. This model may lead to higher accuracies than the logistic regression model in this research. When the best model is then created, with potentially different versions for specific types of movies, this model can be employed as risk assessment tool when movies are proposed. The model results can then give an indication whether more efforts may need to be made to increase the female representation in a movie.

Additionally, when a first version of the movie script is completed, another model that is trained on movie scripts can be used as a second assessment before the movie gets filmed. Here it could be interesting to research if there is a difference in performance between models that

are using similar methods as the ones in this research, and ones that specifically evaluate each criterium on all conversations by employing gender classifiers or, preferably, by using metadata that accurately specifies the gender identities of characters.

Moreover, it is interesting to research different metrics that test the representation of other minority/oppressed groups, or ones that are more complex extensions of this minimum threshold test. For this it is also important to look at the way in which groups are represented, e.g., without stereotypes, and not only at the existence of some representation like the Bechdel test does. Examples of different tests are:

- The DuVernay test which assesses racial diversity in movies by checking if there are people of colour who live fully realized lives instead of merely serving as side/background characters [22];
- The Chavez Perez test which assesses if two non-white characters with names talk to each other about something else than a crime [23];
- The Johanson analysis which rates movies based on the type of representation of women in movies, including the following categories: female protagonist, agency, male gaze, stereotypes [24];
- The Vito Russo test which evaluates LGBTQ+ representation in a way that characters are not only defined by their sexual- or gender identity and are significant characters in the plot [25]. For this it is specifically interesting to explicitly focus on transgender, asexual, and aromantic representation, as well as intersections with multiple minority/oppressed identities, and consensual non-monogamous relationship types.

Lastly, we would like to mention that we hope that the Bechdel test will become obsolete in the future because all movies pass it, without using stereotypes while doing this. In this case, our model would also start performing worse because it cannot pick up on those stereotypes anymore. Therefore, it would need to be trained again on the newer movies, or rather on mentioned extensions of the test that assess more representation of diversity and inclusivity.

References

[1] Merriam-Webster, Bechdel Test, n.d. https://www.merriam-webster.com/dictionary/Bechdel%20Test.

[2] A. Bechdel, The rule, in: Dykes To Watch Out For, Firebrand Books, 1986, p. 22. https://dykestowatchoutfor.com/the-rule/.

[3] K. Gray Bouchat, Testing the Bechdel test, University Honors Theses (2019). URL: https://doi.org/10.15760/honors.731. doi:10.15760/honors.731.

[4] R. Smith, Sizing up Hollywood's gender gap, 2017. https://researchblog.duke.edu/2017/08/04/sizing-up-hollywoods-gender-gap/.

[5] J. O'Meara, What "the Bechdel test" doesn't tell us: examining women's verbal and vocal (dis)empowerment in cinema, Feminist Media Studies 16 (2016) 1120–1123. URL: https://doi.org/10.1080/14680777.2016.1234239. doi:10.1080/14680777.2016.1234239.

[6] S. L. Smith, M. Choueiti, K. Pieper, Inequality in 1,300 Popular Films: Examining Portrayals of Gender, Race/Ethnicity, LGBTQ & Disability from 2007 to 2019, USC Annenberg Inclusion Initiative (2020). URL: https://assets.uscannenberg.org/docs/aii-inequality_1300_popular_films_09-08-2020.pdf.

[7] K. Faith Lawrence, SPARQLing Conversation: Automating The Bechdel-Wallace Test (2011). URL: http://nht.ecs.soton.ac.uk/2011/papers/12-flawrence.pdf.
[8] A. Agarwal, J. Zheng, S. Vasanth Kamath, S. Balasubramanian, S. Ann Dey, Key female characters in film have more to talk about besides men: Automating the Bechdel test, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (2015) 830–840. URL: https://aclanthology.org/N15-1084.pdf.
[9] Merriam-Webster, Metadata, n.d. https://www.merriam-webster.com/dictionary/metadata.
[10] E. Irhamy, Predicting Bechdel test score using machine learning, 2020. https://ai.plainenglish.io/predicting-bechdel-test-score-using-machine-learning-7253618a3f8.
[11] Merriam-Webster, Synopsis, n.d. https://www.merriam-webster.com/dictionary/synopsis.
[12] T. Guardian, Swedish cinemas take aim at gender bias with bechdel test rating, 2013. https://www.theguardian.com/world/2013/nov/06/swedish-cinemas-bechdel-test-films-gender-bias.
[13] N. Selvaraj, The Bechdel test: Analyzing gender disparity in Hollywood, 2020. https://towardsdatascience.com/the-bechdel-test-analyzing-gender-disparity-in-hollywood-263cd4bcd9d.
[14] bechdeltest.com, Bechdel test charts, 2013. https://imgur.com/a/612eD#9PzuDib.
[15] bechdeltest.com, bechdeltest.com API documentation, n.d. https://bechdeltest.com/api/v1/doc#getMovieByImdbId.
[16] C. D. Manning, P. Raghavan, H. Schütze, Stemming and lemmatization, 2008. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html.
[17] S. Wilcox, B. Blackand, Sense and serendipity: Some ways fiction writers choose character names, Names 59 (2011) 152–163.
[18] GLAAD, 9th annual GLAAD studio responsibility index: Growth in racial diversity and screen time for LGBTQ characters but zero transgender characters in 2020 wide release films, 2021. https://www.glaad.org/releases/9th-annual-glaad-studio-responsibility-index-growth-racial-diversity-and-screen-time-lgbtq.
[19] T. Wood, What is the f-score?, n.d. https://deepai.org/machine-learning-glossary-and-terms/f-score.
[20] S. S. A. USA, Popularity of name, n.d. https://www.ssa.gov/cgi-bin/babyname.cgi.
[21] C. Iasiello, Underrepresentation of minorities in Hollywood films: An agent based modeling approach to explanations, Proceedings of the 2017 Winter Simulation Conference (2017) 4582–4583. URL: https://www.informs-sim.org/wsc17papers/includes/files/445.pdf.
[22] C. Moore, Have we cleared the intersection yet? Black women in comic film adaptations, ImageText: Interdisciplinary Comics Studies 11 (2020). URL: https://imagetextjournal.com/have-we-cleared-the-intersection-yet-black-women-in-comic-film-adaptations/.
[23] I. C. Perez, The Chavez Perez-test celebrates its 10th anniversary, 2021. https://www.intichavezperez.se/the-chavez-perez-test-celebrates-its-10th-anniversary/.
[24] M. Johanson, Where are the women? rating criteria explained (updated!), 2016. https://www.flickfilosopher.com/2016/04/where-are-the-women-rating-criteria-explained.html.
[25] GLAAD, GLAAD introduces 'studio responsibility index,' report on LGBT images in films released by 'big six' studios, 2013. https://www.glaad.org/releases/glaad-introduces-studio-responsibility-index-report-lgbt-images-films-released-big-six.