Hidden sexism in the songs we all know and love

Jago Muhlehner
2666821
j.muhlehner@student.vu.nl

Ilayda Bilgin
2668922
i.bilgin@student.vu.nl

Sofia Fraile
2659410
s.m.fraile@vu.nl

Kilian Diederix
2673989
k.diederix@student.vu.nl

Abstract

Throughout the centuries music has taken place in almost everyone's life, while studies have shown that music can have a big influence on our everyday behaviour and ideologies. This paper analyzes this influence by focusing on sexist content in contemporary music. Using cross-domain classification sentiment analysis a model was created trained on identifying sexist tweets and applied to song lyrics. The results of this model was then tested on a different lyric dataset and tested for accuracy.

1 Introduction

Music has become part of our everyday life. We listen to it on the bike to work, in the metro, when we're sad and when we're happy, we might hum to the beat or scream to the lyrics. But recent studies by Ruth & Schramm have shown that lyrics of music might change our everyday behaviour (Ruth and Schramm, 2020). This is particularly concerning when the theme of said lyrics is a troublesome subject like racism or sexism (Neff, 2014).

In research done in 2011 by Brummert et al., violent lyrics were shared with participants. Participants reported feeling angrier and more aggressive after being exposed to these lyrics for a prolonged period of time. These results were similar across three different music genres (Brummert Lennings and Warburton, 2011).

Sexism has been a recurrent and present theme in our society for centuries. In Britain alone, countless women who work in the music industry testify to frequent cases of sexual misconduct and misogyny (Claire, 2017). This problem is of course, not only existing in Britain and not only in the music industry but extends worldwide in all sorts of forms and in all sorts of circumstances, think about social media, catcalling etc.

The fact that listening to sexist or misogynistic music could possibly lead to more sexism in any type of form is something nobody wants. But before that can be tested, a classifier has to be made that can identify 'sexism' in songs.

1.1 Research Question

To dive deep into the issue of sexism, using text mining techniques we analyze the level of sexism of popular songs and attempt to resolve the question *"Can a model trained on social media posts be used successfully on song lyrics ?"* by analyzing the number of sexist songs, versus non-sexist using a sentiment analysis classifier. We expect that this is possible, but it won't be a surprise if the accuracy of the model is around 0.5, because social media posts and songs can be vastly different in i.e. sentence structure, length and meaning.

2 Methodology

The following section will present the methods that were used during the text mining process of this project. The project (Group13, 2022) can be found here: https://github.com/KilianDiederix/TMprojectGroup13

2.1 Resources

Text mining consists of different stages with different steps to be followed in each of them (Isaeva and Aldarova, 2021). Finding datasets and defining the problem that will be solved by the text mining technique is the first step to take. We chose the techniques in this project as Sentiment Analysis and Cross-Domain Evaluation.

We trained our model based on the EXIST (sEXism Identification in Social neTworks) challenge database; a shared task that proposes two challenges: sexism identification and sexism categorization of tweets and gabs, both in Spanish and English. For the first task the data is categorized

as either sexist or not sexist, while for the second task the sexist tweets are divided in the type of sexism they represent, such as stereotyping-dominance, ideological and inequality, objectification and (non) -sexual violence (Rodríguez-Sánchez et al., 2021).

This model was then used in the songlyric data set which has lyrics from 6 different genres found on Kaggle (Neisse, 2020), however this dataset turned out to be quite unhelpful for the proposed task, as we had no efficient way of evaluating the results. The dataset contained a large number of songs, however once the model predicted whether a song was sexist or not it would be extremely inefficient to check the accuracy of the predictions and perfom a proper analysis in order to confirm or negate the research question. For this reason we decided to construct our own small lyric dataset to test our model based on popular songs. We tried to draw a sample of songs (both sexist and non-sexist) from various genres that would allow us to properly represent the effectiveness or drawbacks of our approach with regards to the identification of sexism, as can be seen further on. Even though the dataset is very small it does manage to highlight some key issues relating to the proposed task. All of our code files and sources can be further reviewed on our code uploads.

2.2 Data Exploration and Preparation

The datasets that we obtained had to be prepared before they could be used in the Text Mining process. After the data exploration process and getting to know what our datasets precisely consisted of, we started with the pre-processing steps. In order to make a more accurate analysis, we decided to focus only on English lyrics as the EXIST dataset contained Spanish lyrics as well. We thus dropped all Spanish tweets, so only those in English remained. On the remaining tweets, we first cleaned them from commonly found symbols (such as '@' and '#' with usernames, links etc.) after which we proceeded to remove all the stopwords and transfrom words to their common base form using the SnowballStemmer. After the stemming and tokenization, we obtained a final version of the lyrics that we used.

2.3 Training the models

After the data preparation process, we trained a Logistic Regression model in order to have an initial idea of how accurate the lyrics would be classified as sexist or non-sexist. Having trained that, we converted the raw frequency counts to TF-IDF and trained the model again. We then experimented with various document frequency thresholds, namely four, six and eight. We used the confusion matrix in order to observe the ratio of the correctly or incorrectly classified lyrics. Furthermore we trained the tweet and gab data on a classifier model using Multinomial Naive Bayes both with bag of words and TF-IDF vectorization to see the accuracy score and observe how that would perform. All the logistic regression models were trained with an ngram_range of (1,2).

2.4 Songs dataset

The next step after training models on the EXIST dataset, was to create a new dataset based on songs that we know to be sexist or not sexist. We gathered songs, along with their lyrics and manually appointed gold labels (labeled sexist based on articles and analyses found online) and after applying the same preprocessing steps as the previous database, we used the initial logistic regression model to predict labels for the new data such that the models used on the domain dataset are now being used to predict the labels of our target dataset.

3 Results

The results from the various models trained on the EXIST database can be visualized in the table below, where LR stands for Logistic Regression and MNB stand for Multinominal Naive Bayes:

Model	Settings	Accuracy	F1-score
LR	Bag of words	0.7237	0.72
LR	TF-IDF	0.7196	0.72
LR	TF-IDF, min_df=4	0.7160	0.72
LR	TF-IDF, min_df=6	0.7151	0.71
LR	TF-IDF, min_df=8	0.7219	0.72
MNB	Bag of words	0.6834	0.68
MNB	TF-IDF	0.6784	0.68

Table 1: Accuracy and F1-score of all models.

As we can see, the first Logistic Regression model with bag of words representation performed the best with an accuracy of 0.7237. For this reason we chose this model, alongside the Naive Bayes model (for comparison), to use for the cross domaim sentiment analysis. The results for the label predictions on the song dataset are both 0.65 (albeit with different classifications for some of the songs) on a total of 20 songs. However, this value does not properly reflect the model as the songs were chosen specifically to gauge what the

model can and cannot handle and where it performs best/worst, and thus do not represent a normal distribution of the types of songs that are popular today (see discussion for further details). With regards to the classifier, after training the model we further investigated the features which are most important for both sexist and not sexist songs (see figure(1) in the appendix). This resulted in two lists of words and their importance score, which help to understand why certain songs are classified in a certain way. Finally, we also experimented with the second task of the sexism dataset, or rather predicting the specific type of sexism represented in a given tweet. This resulted in an accuracy of 0.5978, which doesn't seem very high, however, identifying types of sexism is not a simple task. From the confusion matrix and the classification report (see figures (2) and (3) in the appendix) we can identify which types the algorithm struggles most with, mainly misogyny and non sexual violence, and what the algorithm identifies with more consistency, for instance ideological inequality.

To analyse the results an ANOVA 2 way test was performed. The null hypothesis being that there is no difference between the performance of logistic regression and Multinomial naive bayes in this classification problem. The test gave a result of 0.002975 which is smaller than alpha = 0.05, therefore we can reject the null hypothesis and have sufficient evidence to say that there's a statistically significant difference between the mean of the accuracy scores. Confirming that logistic regression has a better outcome.

4 Discussion

As mentioned above, the accuracy of the predicted values is relative and strongly depends on the selected songs, which were chosen to point out the strengths and issues pertinent to classifying sexism in popular song lyrics. In figure (4) of the appendix we can see each song with the correct label, attributed to each song by reading the lyrics and researching blogs about the song (with an IAA of 0.6875), alongside the predicted values from both the Logistic Regression and Naive Bayes models. Seven of the eleven songs identified as being sexist were correctly classified by both models. This is due to the fact that these songs are quite blatantly sexist, particularly when looking at the choices of words, such as the excessive use of derogatory terminology and the overall high presence of words that were found to be important features in detecting sexism. For the non sexist songs, only four of the nine were correctly labeled by both models simultaneously. The key drawbacks arise from the mislabeling of sexist or non sexist songs, which we will analyse in more detail: First of all, with the songs labeled as non sexist which were classified as sexist we found three possible reasons for the misclassification. Firstly, with the song 'Girls like you' by Maroon5, as the song contains the word 'girl' many times. While this word is both an important feature for sexist and non sexist songs, it has a higher importance in sexist songs and the high repetition of this word likely impacts the overall label. Secondly, the song 'Hard out here' by Lily Allen seems to be sexist at first glance, however it is a sarcastic nod to sexism in pop music, and thus technically portrays the opposite message. One could argue however that the song was labelled correctly as the lyrics, taken out of social context, are in fact sexist. Lastly, the song 'I like it' by Cardi B is labelled as sexist due to the high presence of 'vulgar' words, which together in the song don't allude to any form of sexism however these words are often used in a defamatory and misogynistic manner. On the other hand there are songs that were mistakenly classified as not sexist when in reality they were. This pertains to songs such as 'Baby it's cold outside' and 'Hotline bling'. While at first sight these songs seem harmless and free of misogyny, when analysing the lyrics one can uncover various sexist implications and insinuations. In the first song the man ignores his date's protests and pressures her to stay, giving off the sentiment of sexual coercion, or another example where at one point the woman asks, "Say, what's in this drink?" which could allude to roofying, in a present day context. These subtleties are not explicitly sexist and thus are more often overlooked by a machine learning algorithm, which couldn't possibly understand (at the present time) every aspect of what is the multilayered and multifaceted reality of sexism.

5 Conclusion

The answer to the research question "Can a model trained on social media posts be used successfully on song lyrics ?" is not as straightforward as we would like. As we have seen it is possible to detect sexism in music lyrics to a certain extent, how-

ever we have experienced various cases in which the current text mining and machine learning techniques struggle to apprehend the more implicit side of sexism given how deeply it is rooted in our society. Nonetheless, the results of our analysis showed that it was possible to train on a dataset from a completely different domain and use that in our target dataset, providing valuable insight as to which types of sexism are more easily detectable by a machine and which still need some work. However, the model did perform better than we expected.

5.1 Limitations and further research

The main limitations on our approach is in regards to time. We would like to take a neural network approach to this cross-domain sentiment analysis. This would give us a much deeper understanding of the text in the lyrics by understanding patterns, therefore hidden sexism, that could potentially be missed with a classifier and solve the problem of incorrect classification based on sarcasm or non-explicit lyrics, by allowing it to take into account the context of it. We were also limited by fact that there was no database containing sexist song lyrics which we could use to properly analyse the effectiveness of a cross domain analysis for this particular sentiment. With more time we would have also liked to expand our project to include historical data (which is unfortunately not available), this historical data would need to be retrieved from services like genius and Spotify to then use our trained model in these hit songs throughout the years. That would then provide a great outlook into whether the sexism in songs is prevalent or it has decreased with the new awareness brought to the topic.

6 Division of work and Appendix

The division of work was as follows:

	Coding	Analysis	Reporting
Jago	40%	25%	25%
Sofia	10%	25%	25%
Ilayda	20%	30%	25%
Kilian	30%	20%	25%

Table 2: Division of work.

Kilian started with classifying the sEXIST dataset, he made the classifier, while Ilayda helped with the pre-processing phase. Jago then used the classifier to test on our own created dataset. He also trained a Multinominal Naive Bayes model. Ilayda and Sofia helped with our coding predominantly by creating the dataset and analyzing the results. ANOVA tests were done on our analysis which was done by Sofia. Kilian and Jago were responsible for the classification reports and confusion matrices. Everyone then reported on their own part, and worked together to get a well-flowing report. We all agree that everyone did their part in this project. However, just like with any group project, there will always be one or two people who do slightly more than the rest.

Figure 1: Word Cloud for most important words in 'sexist' classification

Figure 2: Confusion matrix, type of sexism in songs.

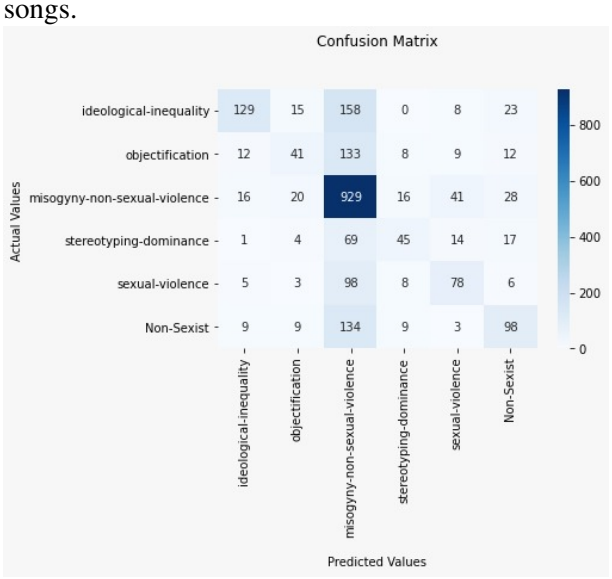

Figure 3: Classification Report for task 2

	precision	recall	f1-score	support
ideological-inequality	0.75	0.39	0.51	333
misogyny-non-sexual-violence	0.45	0.19	0.27	215
non-sexist	0.61	0.88	0.72	1050
objectification	0.52	0.30	0.38	150
sexual-violence	0.51	0.39	0.44	198
stereotyping-dominance	0.53	0.37	0.44	262
accuracy			0.60	2208
macro avg	0.56	0.42	0.46	2208
weighted avg	0.59	0.60	0.56	2208

Figure 4: Predictions on songs

Song	Gold	LR-pred	NB-pred
Some Girls	Sexist	Sexist	Sexist
kim	Sexist	Sexist	Sexist
Young girl	Sexist	Sexist	Sexist
Blurred lines	Sexist	Sexist	Sexist
girls like you	Not sexist	Sexist	Sexist
Gods plan	Not sexist	Not sexist	Not sexist
bitches ain't shit	Sexist	Sexist	Sexist
The A team	Not sexist	Not sexist	Not sexist
Little bad girl	Sexist	Sexist	Sexist
Baby it's cold outside	Sexist	Not sexist	Not sexist
Still loving you	Not sexist	Not sexist	Sexist
Chiquitita	Not sexist	Not sexist	Sexist
If i ain't got you	Not sexist	Not sexist	Not sexist
Can't hold us	Not sexist	Not sexist	Not sexist
animals	Sexist	Not sexist	Sexist
Hey Good Lookin	Sexist	Not sexist	Sexist
Hard out here	Not sexist	Sexist	Sexist
Hotline Bling	Sexist	Not sexist	Not sexist
God made girls	Sexist	Sexist	Sexist
I like it	Not sexist	Sexist	Sexist

7

References

Brummert Lennings, H. I. and Warburton, W. A. (2011). The effect of auditory versus visual violent media exposure on aggressive behaviour: The role of song lyrics, video clips and musical tone. *Journal of Experimental Social Psychology*, 47:794–799.

Claire, M. (2017). Why the harvey weinstein allegations show that we need to acknowledge sexism in the music industry too.

Group13 (2022). Repository code and used datasets.

Isaeva, E. and Aldarova, D. (2021). Text-mining in terms of methodology and development.

Neff, S. (2014). Scholarworks at wmu sexism across musical genres: A comparison.

Neisse (2020). Song lyrics from 79 musical genres.

Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., and Donoso, T. (2021). Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Ruth, N. and Schramm, H. (2020). Effects of prosocial lyrics and musical production elements on emotions, thoughts and behavior. *Psychology of Music*, 49:030573562090253.