Misogyny Detection with Applications in Modern Popular Music Lyrics

Luke Verdi
luke.verdi@berkeley.edu
UC Berkeley MIDS - W266 Final Project

Abstract

Misogynistic messages in modern culture can have a huge detrimental effect on women. As such, it is important to be able to identify and classify these messages of bias across multiple forms of communication. The present study seeks to train a misogyny identifier and classifier on Twitter and Reddit posts, with the goal of being able to be used on text outside of just a particular platform. This was tested by running the resulting generalized model on popular song lyrics from 2000-2019, identifying possible misogynistic messages in upwards of 20% of songs in each year, with a skew towards misogynistic treatment.

1 Introduction

Plato is said to have written that music "gives soul to the universe, wings to the mind, flight to the imagination, and charm and gaiety to life and to everything." While music has certainly changed immensely since his time, music is still a powerful part of popular culture and plays an important emotional role in an individual's life experiences. Thus, it is necessary to understand just what messages are conveyed by popular music. Understanding if and how popular music contains messages of bias could reveal that the imaginations of listeners are being influenced in undesirable ways, or that biased messages are being chosen by listeners. Given how unappealing either option is, this study seeks to evaluate musical lyrics for gender bias. However, in order to get to the point of running a model on song lyrics, it was first necessary to understand misogyny in other contexts, and to build off the work of others in classifying gender bias.

2 Background

When considering misogyny, a working quantifiable definition is first needed. This proves to be more difficult than simply using "hatred of, aversion to, or prejudice against women" from Webster's dictionary. While lyrics are labeled for explicit content, and work has been performed by Fell et al. (2019) to classify lyrics as explicit or not, no well recognized dataset exists which categorizes whether song lyrics are misogynistic.

Thus, two approaches are reasonable. First would be the approach by Boghrati (2020) and Faculty, which looked at Word2vec embeddings of lyrics, and assessed whether men or women in lyrics were more closely associated with positive traits, such as competence. Notably, this approach avoids having to formally classify whether a particular bit of text is actively misogynistic, and instead uses similarity to positive traits. There is also no formal training set, aside from whatever data was used to train the Word2vec embeddings.

A second approach would be to train a model on some other dataset, which would be expertly labeled for misogyny, before running that model on song lyrics. This model would have to be careful about how misogyny is labeled and ensure that the model transfers from the training dataset to the dataset of song lyrics. But it would have the advantage of being able to more explicitly state that a bit of text was misogynistic. The model could even attempt to classify misogynistic messages of different classes. As this space hasn't been explored as well, it was chosen for this study.

In selecting what dataset to train on and what model to use, several factors were taken into account. As the ability of the model to transfer is key, the model must not be over-fit to one type of text. Thus, after searching through several possible datasets, two were chosen. First was a dataset on English Reddit replies, put together by Guest et al. (2021), which determined not only misogyny but also different classes, such as treatment or derogation. It even includes sub-classes such as types of disrespectful actions or threatening behavior.

Next was a dataset of English Twitter posts label for misogyny by Anzovino et al. (2018) as part of the IBEREVAL 2018 Autonomous Misogyny Identification challenge. This dataset was graciously provided by Professor Fersini, as it is not publically available (unlike the Guest Reddit dataset, which is available to all on her git repository). It is thus worth noting that this Twitter dataset is then not included in the git repo for this project. This dataset also includes sub-classes of misogyny, but the classes chosen are different than those chosen by Guest. These discrepancies in sub-classifying misogyny are a common theme among many papers in the field, as Samghabadi et al. (2020) and Parikh et al. (2019) also use different classes.

While it is noted that Reddit replies and Twitter posts may not be completely analogous to song lyrics, the platforms are a place of open ended personal expression, which is can be similar than song lyrics. Also, as anyone who has listened to Taylor Swift knows, sometimes songs are written in response to a particular person or event, like a Reddit post.

After aligning on datasets, it was determined that a BERT model would be used for classifying and sub-classifying misogyny. BERT was chosen as Rodriguez-Sanchez et al. (2020) demonstrated that it could perform very well classifying misogyny in Tweets. Samghabadi et al. (2020) and Parikh et al. (2019) also use BERT for their classification. It is notable that Fell et al. (2019) did demonstrate possibly better results with other model architectures. But with the amount of data in the training sets being appropriate for fine-tuning a BERT model, and with the ability of BERT to offer contextualized embeddings to a variety of text types and lengths, it seems a sensible choice for this use case.

3 Methods

As mentioned, the two training datasets used had slightly different classifications for misogyny. One of the biggest differences was the fact that the Guest Reddit dataset classifies the use of a "misogynistic perjorative" as misogynistic. Meanwhile the Twitter dataset was gathered by searching for certain keywords. As such, it is likely that more words deemed to be misogynistic pejoratives would be included, but are not always classified as misogynistic, dependent on the way that they are used in context. To rectify this difference, the "misogynistic perjorative" subclass from the Reddit dataset has been re-classified as non-misogynistic. While this seems extreme, it is worth noting that the Reddit dataset allows for multiple labels. If a post contained a misogynistic perjorative but also included derogation with regards to intellectual inferiority, both would be labeled. In that case, the post would still be classified as misogynistic. This decision was also made for the purposes of generalization. A "foid" (derogatory slang term for a female) was identified as a misogynistic perjorative within the dataset, but this does not appear in any song lyrics.

With regards to sub-classifications, it proved to be very challenging to get the datasets to match up in a meaningful way which a model could learn from. Appendix A details the final subclassifications chosen. While initially more subclassifications were attempted, a model struggled to differentiate between classes like Discrediting and Stereotyping. One thing that was certainly learned from this project was that misogyny can be hard to sub-classify even as a human. As an example of this, the following Tweet was taken from the IBEREVAL dataset. As a reader, you're invited to first attempt to classify it yourself between Derailing, Discrediting, Dominance, Stereotype, Sexual Harassment and Dominance.

> *You just have power over men altogether, women are men's ultimate weakness after all... which is why women are slut...*

This is undeniably awful, and certainly misogynistic. But it's actually hard to pick between Derailing, Stereotyping and Discrediting (for the record, passive Discrediting was chosen). For this reason, only 2 sub-classes were chosen for the merged dataset: Derogation and Treatment. Once his distinction has been made, it is more clear that the above would fall under Derogation.

Once the data sets had been re-classified, some data cleaning was performed to remove Twitter usernames (preceded by an @ symbol) or any web links. This is considered safe as these characters likely aren't crucial to the classification, and wouldn't be found in song lyrics. However, any Twitter hashtags or any emojis were left in. While this wouldn't transfer to song lyrics, they could be integral to the nature of misogyny (for example, ending a Tweet with #womenarebad).

Finally, the merged training dataset was approximately evenly comprised of Twitter and Reddit data, and broken into an 80/20 train/test set. It is worth noting that approximately 23% of this dataset is labeled as Misogynistic, meaning there is a slight class imbalance. However, attempting to account for this by dropping data produced worse results.

Next, a BERT model was used to attempt to classify whether text was misogynous or not. After much experimentation, the model included a hidden layer on top of the BERT outputs, followed by a classification layer including Dropout and Layer Normalization. This model architecture is in line with Samghabadi et al. (2020) and seeks to avoid over-fitting. Also of note was that a max length of 75 was used. While this max length is shorter than the length of a song, it was necessary for the shorter nature of the text in the training set, and can be countered by breaking songs into shorter chunks when running on lyrics. A bert-based-cased tokenizer and model were used due to having around 10000 training examples (likely not enough for training a large BERT model). Early stopping was also used, along with training only limited layers, as training for longer than 3 Epochs or on all layers demonstrated harmful over-fitting between train and test.

Finally, an additional Bert Model was used to distinguish between classes of misogyny. This model was constructed in a very similar way to the first model, and was trained only on the misogynous data, labeled as Treatment or Derogation. Both of these models were trained on a GCP instance with a T4 GPU.

Once these models were properly trained, they were run on song lyrics pulled from Kaggle[1]. Songs were broken into overlapping chunks of near the maximum length the models were trained on. Repeated detection of misogynistic content in a song would result in the song being deemed misogynistic.

4 Results and discussion

4.1 Results on Twitter and Reddit Data

The BERT model detecting misogyny had the following success as defined by the test set from the Reddit and Twitter labels:

Accuracy	Precision	Recall
80.0%	57.4%	60.1%

This is by no means as impressive as values obtained by some of the contestants in the IBEREVAL competition for Subtask A (Misogyny Identification) in English, as were 2 models which eclipsed 90% accuracy. However, when considering that the AMI baseline model had 78.4%, and this study's model is not exclusively focused on Twitter data, the score is a bit more reasonable.

With this classification model it is particularly interesting to see the modes of failure. While the model certainly misses some text it should classify as misogynous and generates false positives, some misses are very understandable. For example, the following two Reddit comments were identified as misogynous by the model, but not by the experts:

> *-Women were not allowed to be hired, or even have a savings account. Cite?*
>
> *If she only post her art work people will asume a dude made it. That's why some women likes to appear next with their art.*

Due to the nature of Reddit comments being responsive, these comments themselves are not misogynistic, but are quoting or countering comments which likely are. It makes sense that a model trained without the full context would struggle to identify this. On the other hand, here is a Tweet which the model mistakenly classifies as non-misogynisitc:

> *I'll start considering women being equal to men when I'll see just one woman, just one, doing this:*

The first half of that ("I'll start considering women being equal to men") in the right context could be non-misogynistic. But as this may have been a Retweet or may link to something else, the nature of gender bias isn't immediately apparent. Given that song lyrics won't frequently have the same issues with quoting other text or referencing unseen messages, these failure modes shouldn't carry over to the song lyrics.

As for the misogyny classification model, performance was worse. Even with only two classes of misogyny in the dataset, accuracy only reached 70.1%. With a slight class imbalance in the data, this was hardly better than picking the most popular class of Derogation (65.6% of the data). Some of this is again due to the model struggling to learn (for this model there was less data to train on), but even with two classes there are again examples of perhaps ambiguous data. The below absolutely

[1] https://www.kaggle.com/jackrosener/billboard-top-100-song-lyrics-19502012

vile tweet was classified in the Twitter dataset Discredit - Active, but the last sentence could easily be classified under Threaten

> *I usually dont hate people but I actually hate that bitch. You tried to be friends with me outside of work and then when I had a crisis you decided to be professional and use that against me. Thats the worst type of person. Fucking cunt will get hers though.*

Similarly, the below Reddit comment was classified by the model as Derogation, but in the dataset is under Treatment - Disrespectful Actions - Seduction and Conquest:

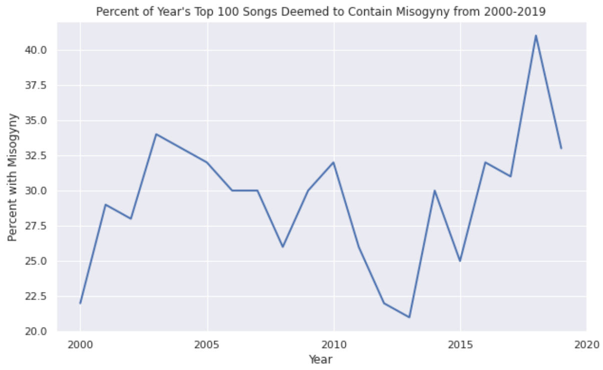
> *Cold approach is very hard. Forget people judging, worry about how the girl feels. Many might not like it or find it weird or creepy. And many will love it and be amazed by your balls to go up to her.*

Seduction and conquest is certainly appropriate, but there is also stereotyping at play, which fell under Derogation. And while this could be a problem with the re-categorization effort to get the datasets to align, several different classifications were attempted, all with similar results. A takeaway is that misogyny requires strict definitions, specifically with how to handle posts crossing over into many different categories. This definitions need to be applied by qualified individuals, preferably to data across several different domains to allow for appropriate model building.

4.2 Running the Model on Song Lyrics

While the model was initially going to be run on song lyrics going back to 1960, there is simply very little in common with the language on Twitter and Reddit the lyrics of Roy Orbison. Without being able to control for the discrepancies in language, like how a word such as “fuck” is so common in the training set, it is disingenuous to run this model on such old data without meaningful corrections. However, the word “fuck” appears multiple times in Future’s 2019 hit “First Off” along with the line: *I should pimp this bitch and make her pay me*. Depressingly, that looks incredibly similar to the Twitter and Reddit text.

In order to run the much longer songs through the model, each song was segmented into chunks, and each chunk was classified for misogyny. As a rap song with many verses could end up with over 20 chunks of text, the false positive rate of the model had to be considered. If a song with 20 sections of text happened to have one which was mistakenly categorized as misogynous, calling the whole song misogynous would be inappropriate. Thus, a threshold of 3 misogynistic sections was determined before a song would be deemed misogynous. Finally, the percentage of hit songs in a given year which were misogynous could be determined, as shown in the graph above.



While it is likely the case that the model generating false positives on certain sections is leading to an inflated percentage of songs deemed to be misogynous, it is worth noting that it is not at all hard to find misogynistic lyrics in modern songs. From the aforementioned “First Off” to “Pure Water” by Mustard Migos including lines such as *Dismantle her, I know how to handle her* or *Pimpin’ ain’t easy, make her open up and eat it* to “No Mediocre” by T.I. coming right out of the gate with *All I fuck is bad bitches, I don’t want no mediocre hoe*. All of these songs are identified as overwhelmingly misogynistic by the model (10 misogynistic sections for “Pure Water” and even more for “No Mediocre”), and each looks the part to a human eye.

Lastly, the model designed to sub-classify misogyny was also applied to the segmented song lyric dataset. While there are many issues with this model due to the difficulties in sub-classification mentioned, it was incredibly striking to see 82.5% of the sub-classifications come up as misogynistic due to Treatment. This can be compared to only 23.1% of the sub-classifications on the training set being deemed misogynistic due to Treatment by the model. Further, while the sub-classification model routinely mistakenly identified Derogation as Treatment (this happened in 20.1% of instances) it more rarely mis-identified Treatment as Derogation (only 9.2% of cases). Judging off of the sampled lyrics above, it does seem plausible that the misogyny in modern song lyrics is more skewed towards Treatment then towards Derogation, especially when compared against online posts. While both sub-classes of misogyny are harmful, it could be argued that Treatment is worse.

5 Conclusion

This project is certainly not a finished product when it comes to misogyny classification. More expertly labeled data, or better harmonization around misogyny sub-classes in datasets would allow for better model generalization and improved learning. In addition, different model architectures should be explored. However, the study is able to demonstrate that training a model on identifying and classifying misogyny across different types of text is possible. It also demonstrates that such a model can be deployed on new unseen text, and generate at least a few results which align with what a human reviewer would say.

From a less technical perspective, this project identifies that misogyny is prevalent in modern popular music, and details biased Treatment of women. Having read far too much misogynistic content in the Tweets and Reddit comments (which if it can be believed is often orders of magnitude worse than what was shared in this paper) the world needs much less of this sentiment. Hopefully by identifying misogyny across whatever platform it exists on, we can work towards removing it.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, page 57–64. Springer.

Reihane Boghrati. 2020. Quantifying 50 years of misogyny in music.

Wharton Faculty. Quantifying cultural change: An application to misogyny in music. *Journal of Consumer Research*.

Michael Fell, Elena Cabrio, Michele Corazza, and Fabien Gandon. 2019. Comparing automated methods to detect explicit content in song lyrics. *RANLP 2019 - Recent Advances in Natural Language Processing*.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework.

Francisco Rodriguez-Sanchez, Jorge Carrillo de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, page 126–131.

A Misogyny Classifications

More detail on the re-classifications made with the training data. First, the Guest Reddit dataset. Level 2 and Level 3 come from the dataset itself (within the misogynistic Level 1 classification). The New class denotes the new sub-classification.

Level 2	Level 3	New
Derogation	Intellectual Inferiority	Derogation
Derogation	Moral Inferiority	Derogation
Derogation	Other	Derogation
Derogation	Sexual or physical limitations	Derogation
Misogynistic pejorative	None	Non
Misogynistic personal attack	Gender of recipient is Female	Derogation
Misogynistic personal attack	Gender of recipient is Male	Derogation
Misogynistic personal attack	Gender of recipient is Unknown	Derogation
Treatment	Disrespectful actions Controlling	Treatment
Treatment	Disrespectful actions Manipulation	Treatment
Treatment	Disrespectful actions Other	Treatment
Treatment	Disrespectful actions Seduction and conquest	Treatment
Treatment	Threatening Physical violence	Treatment
Treatment	Threatening Privacy	Treatment
Treatment	Threatening Sexual violence	Treatment

Also, here are the re-classifications from the Twitter dataset from IBEREVAL AMI 2018:

Misogyny Category	Target	New
Derailing	Active	Derogation
Derailing	Passive	Derogation
Discredit	Active	Derogation
Discredit	Passive	Derogation
Dominance	Active	Treatment
Dominance	Passive	Treatment
Sexual Harassment	Active	Treatment
Sexual Harassment	Passive	Treatment
Stereotype	Active	Derogation
Stereotype	Passive	Derogation