HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 1 (2023)

Dataset for Identification of Queerphobia

Shivum Banerjee[1] and Hieu Nguyen[#]

[1]Hinsdale Central High School, Hinsdale, IL, USA
[#]Advisor

ABSTRACT

While social media platforms have implemented many algorithmic approaches to moderating hate speech, there is a lack of datasets on queerphobia which has impeded efforts to automatically recognize and moderate queerphobic hate speech online. Queerphobic hate speech is speech that is intended to degrade, insult, or incite violence or prejudicial action against queer people, who are those from a sexuality, gender, or romantic minority. This speech results in worsened mental and emotional outcomes for queer people and can contribute to anti-queer violence. The goal of this study is to create a dataset of queerphobic YouTube comments to further efforts to identify and moderate queerphobic hate speech. To construct this dataset, 10,000 comments were sourced from YouTube videos which represent queerness. Then, volunteers manually annotated each comment in accordance with specific guidelines. Various natural language processing (NLP) models were used to extract features from the text, and several classifiers used these features to categorize comments as queerphobic or non-queerphobic. These NLP models illustrate a baseline for performance on this data. In making this dataset, we hope to further research in the recognition of digital queerphobia and make social media platforms safer for queer people. The dataset can be found at https://github.com/ShivumB/dataset-for-identification-of-queerphobia.

Introduction

Social media such as YouTube have become important forms of communication for many people, and the anonymity of communication online allows many to express themselves freely. However, this anonymity also facilitates the proliferation of hate speech. Hate speech is a conscious and willful public statement intended to denigrate a group of people (Delgado & Stefancic, 1995), and social media uniquely allows individuals to engage in hate speech because of its lack of consequences. Typically, hate speech targets people of a religious, ethnic, gender, or sexual minority. This is detrimental to the groups it targets, as it can contribute to violence perpetrated offline and growing bigoted sentiment (Tsesis, 2002). In response, many social media platforms have implemented algorithmic approaches to moderating hate speech. Natural language processors are well-suited to the task of interpreting and quantifying language (Engonopoulos et al., 2013). These models process speech by representing words numerically and performing analyses, and it is critical to have a dataset of labeled hate speech to train supervised learning models. However, while there are many accessible datasets of hate speech as a general category, there is little data focused on queerphobic hate speech.

Queerphobia is defined as a "term used to include all forms of homophobia, lesbophobia, biphobia and transphobia" (QMUNITY, 2019, p.17), and queer people are those who belong to a romantic, sexual, or gender minority. In the United States, the queer community is particularly vulnerable to online hate speech. A 2020 report on online hate crimes linked queerphobic hate speech with worsened physical, mental, and emotional health outcomes (Hubbard, 2020) for queer victims. Therefore, a dataset specifically focused on identifying and combating queerphobia online is crucial for the safety and wellbeing of queer individuals.

Literature Review

Queerphobic hate speech is an urgent issue in the United States. The National Coalition of Anti-Violence programs notes that online platforms have become a key tool for organizing and promoting violence against queer people (Olteanu et al., 2018). Furthermore, anti-queer violence is becoming increasingly prevalent. For example, from 2017 to 2021, the number of transgender homicides more than doubled from 29 reported deaths to 56 reported deaths (Everytown Research & Policy, 2022). It is critical to moderate queerphobic hate speech to mitigate its role in the increasing violence perpetrated against queer people. In addition to contributing to anti-queer violence, digital queerphobia hurts queer people by provoking fear of physical violence and increasing emotional distress. Victims of online queerphobic hate speech "experience a wide range of negative emotional responses to their online victimization, including fear, anxiety, self-blame, and suicidal thoughts" (Hubbard, 2020, p.3). Digital queerphobia marginalizes the queer community, so it is urgent to address it by making a dataset to assist in the identification and moderation of queerphobic hate speech online.

While there are many datasets for the identification of hate speech as a general category, these datasets are limited in capturing the nuances of queerphobic hate speech. The problem of generalizability in abusive language detection datasets is complex, and while many models may perform well on benchmark datasets, their performance can degrade when tested on datasets with different characteristics (Swamy et al., 2019). This means that, for a dataset to be used in the identification of queerphobia, it must include instances of digital queerphobia. However, many datasets may contain few instances of queerphobic hate speech. To illustrate, the hate speech dataset contributed by Gibert et al. (2018) uses data from the online white supremacy forum Stormfront. Because of its nature as a white-supremacy forum, the data collected from this site is more focused on race than queerness. Therefore, this dataset may include fewer instances of queerphobic hate speech, and a hate speech detection model trained on data from the white supremacy forum might not be able to recognize queerphobia as well as a dataset created for the identification of queerphobia. For this reason, it is important to create a dataset specifically for the identification of queerphobic hate speech.

The most accessible dataset of queerphobic hate speech comes from a Tamil context. Chakravarthi et al. (2021) contributed a dataset of about 15,000 transphobic and homophobic hate comments in mixed Dravidian languages and English with the purpose of identifying homophobia and transphobia on YouTube. However, because it comes from a Tamil context, models trained on this data may not generalize to an American English context very well. It is important to understand queerphobia in a variety of linguistic contexts, and the aim of this study is to create a dataset to better understand queerphobia in an American English context.

Once this data is collected, it can be used by natural language processing (NLP) algorithms to train supervised learning models to identify queerphobic hate speech. Research has shown that NLP models can be used to effectively identify hate speech in online content. For example, in a study conducted by Davidson et al. (2017), several deep learning algorithms were trained to classify hate speech data from Twitter, the best achieving an F1 score of 0.90. The use of NLP algorithms for moderating online content has many benefits, such as reducing the burden on human moderators and enabling faster identification and removal of harmful content. However, there are also drawbacks to consider. For instance, NLP algorithms may struggle with detecting sarcasm, irony, or cultural nuances, leading to false positives or negatives (Weitzel, Prati, & Aguiar, 2016). Additionally, the use of algorithms to moderate speech raises ethical concerns, such as the risk of censoring legitimate speech or perpetuating biases against marginalized communities (Noble, 2018). To ensure that machine learning algorithms do not perpetuate biases or harm marginalized communities, it is important to use diverse and representative data when training the algorithms, and to regularly audit their performance to identify and correct any biases. Moreover, it is essential to involve stakeholders from diverse backgrounds in the development and implementation of these algorithms to ensure that they are designed and used in an ethical and socially responsible manner.

The creation of a dataset specifically designed to identify and moderate queerphobic hate speech could be critical in moderating hate speech in online content. As discussed earlier, queerphobic hate online not only damages queer individuals' emotional and mental wellbeing but can also result in violence. This dataset supplements the current gap in queerphobic data, enabling machine learning models to more accurately detect queerphobic hate speech.

Methodology

The purpose of this study is to create a dataset for the identification of queerphobia in an American English context. This will supplement the lack of data on digital queerphobia, which is critical to identifying and moderating hate speech against queer people. To create this dataset, data was collected from 20 YouTube videos and annotated by three volunteers who identify as queer or queer allies. Then, several models were fitted to this data to establish baseline performance for NLP models. This section of the paper will describe the data collection, manual annotation, data preprocessing, models, evaluation metrics, and ethics and data privacy.

Data Collection

Data was collected from the social media platform YouTube. YouTube is a video-sharing platform that allows users to upload, share, view, and comment on videos. Many viewers leave comments that relate to the content of the video they watched. For the purpose of collecting data, this is of particular interest. If a video relates to queerness in a particular way, it is possible that some comments will reflect this relationship. Thus, it may be possible to manufacture a dataset of comments that depict many facets of queerness by choosing videos that relate to queerness in different ways. This makes YouTube an excellent source for data.

When choosing videos from which to source comments, two main factors were considered. First, it is important to balance negative and nonnegative depictions of queerness. According to Hovy (2021), one source of bias in NLP modeling is data. If all the comments that relate to queerness in the dataset were negative, an NLP that trained on the data may develop a bias linking queerness to homophobia in all cases. In the inverse case, the model would not learn to recognize queerphobia because there would be no queerphobic comments. Thus, it is critical that the dataset is comprised of comments that relate to queerness in both a negative and nonnegative light. Second, it is important to include a diversity of queerness. The term "queer" is an umbrella term which embraces a diverse variety of identities (QMUNITY, 2018). In order to accurately identify queerphobia, it is therefore imperative to include videos which capture this diversity. With these factors in mind, a total of 10,000 comments were sourced from 20 YouTube videos.

After choosing different YouTube videos from which to download comments, a Google AppScript program was used to download 500 comments from each video (Banerjee, 2023). Given a video, this program provided all the comments, the corresponding usernames, the number of replies, the comments written in response, the date of commenting, and the number of likes. To protect the anonymity of commenters, superfluous information and information which could be used to identify commenters was removed.

Manual Annotation

The biggest challenge in the creation of a labelled dataset is manual annotation. Manual annotation, the process of assigning labels or tags to data by humans, is a useful method for creating labeled data for machine learning applications. Supervised learning algorithms require labelled datasets, and manual annotation is excellent for NLP modeling. However, it can also pose several problems. Manual annotation can be prone to bias and errors, as annotators may have different backgrounds, experiences, and perspectives that can influence their labeling

decisions. This can lead to biased training data and inaccurate machine learning models. Another issue is that manual annotation can be a time-consuming and expensive process, especially for large datasets or complex tasks. It may require hiring and training annotators, as well as developing quality control measures to ensure accuracy and consistency (Wissler et al., 2014).

In order to help annotators classify data consistently, the following guidelines were established:

Read the text data in its entirety to gain an overall understanding of its content and tone. Next, identify language or statements that express negative attitudes or prejudice towards queer individuals or groups. This can include derogatory or offensive slurs, stereotypes, and discriminatory language.

Consider the context in which the text was written or spoken, as this can influence the meaning and impact of language and statements. Pay attention to the historical and social context in which the language or statements were made, as well as the intended audience.

Account for the impact of the language or statements on queer individuals or groups. Language that may not be intended to be queerphobic may still be harmful or hurtful to queer individuals or groups.

Consider the intention of the author or speaker. While intention alone does not necessarily determine whether language or statements are queerphobic or not, it can be a factor to consider when deciding.

Based on your analysis, classify the text data as queerphobic or not.

Classifying text data as queerphobic or not can be complex and nuanced, so these guidelines were written to provide a useful starting point for analyzing text data for queerphobia. These specific guidelines are based on Petrillo and Baycroft's (2010) introduction to manual annotations.

To address the problem of hiring and training annotators, three volunteer English-speakers who identified as part of the LGBTQ+ community or as an ally of the LGBTQ+ community were found. These annotators were familiar with the queer community, and they were able to correctly identify examples of queerphobic comments in an initial meeting. Table 1 contains examples of comments that annotators were asked to classify. These comments were constructed to test annotators' ability to recognize differing levels of queerphobia.

Table 1. Examples of comments used to gauge annotator proficiency in recognizing queerphobia. These examples were specifically written to represent different levels of queerphobia.

Comment	Classification	Explanation
this video is shit.	Not Queerphobic	This comment does not relate to queerness in any way.
Me and my gay friends when we pull up to the bar: 4:39	Not Queerphobic	This comment references queerness in a way that is not negative.
i hate lesbians.	Queerphobic	This comment references queerness in a hateful way.
Well, I have nothing against transgenders; i just wish they wouldn't shove their lifestyle down my throat.	Queerphobic	This comment says that transgenders "shove their lifestyle" down others' throats, which is a queerphobic sentiment.
i don't support the lgbtq+ but that doesn't mean i hate them it's just that i don't approve	Queerphobic	The language used in this comment is more neutral and may suggest an apathetic stance. However, the content is queerphobic. This comment is difficult to classify, and some

		may interpret it as not queer-phobic.

However, even with guidelines and proficient annotators, some comments were difficult to classify. To deal with disagreement between annotators, a two-thirds majority of annotators resolved the classification of contested comments. The annotated data can be found at https://github.com/ShivumB/dataset-for-identification-of-queerphobia (Banerjee, 2023).

Data Preprocessing

In any data analysis project, the quality of the results is highly dependent on the quality of the data. Therefore, it is essential to preprocess the data before performing any analysis. Data preprocessing involves a set of procedures that transform the raw data into a clean, well-organized, and easy-to-analyze format (Kannan & Gurusamy, 2014).

In this study, the following preprocessing steps were taken:

- Remove HTML tags.
- Remove hyperlinks.
- Remove special characters.
- Convert to lowercase.
- Tokenize words.
- Remove stopwords.
- Lemmatize tokens.

These steps were taken to ensure that the data was ready for analysis and to reduce noise and inconsistency in the data. The first step involved removing HTML tags, which are commonly present in web-based data, to obtain only the text data. Hyperlinks were also removed as they do not contribute to the analysis and can cause noise in the data. Similarly, special characters, such as punctuation and emoticons, were removed as they do not provide any useful information for the analysis. The text data was then converted to lowercase to ensure consistency and eliminate any discrepancies caused by capitalization. Tokenization was performed to separate the text into individual words or tokens, which were then analyzed separately. Stopwords, which are commonly occurring words such as "the" and "and", were removed as they do not add significant value to the analysis. Finally, tokens were lemmatized, which involves reducing the words to their base or root form, to ensure that related words are treated as the same word and to reduce the dimensionality of the data. These preprocessing steps were performed using the Natural Language Toolkit (NLTK) library in Python (Bird, 2006).

Models

Natural Language Processing (NLP) is a field of study that focuses on the interaction between human language and computers. NLP models use statistical algorithms and machine learning techniques to analyze, understand, and generate human language. These models are adept at performing sentiment analysis, and they can be used to analyze the created dataset of queerphobic comments. In this study, we utilized several different features and classifiers to analyze the text and understand how well models can learn from this data. These models serve as a baseline for NLP performance on this data.

In NLP, the goal of feature extraction is to transform a text document into a set of numerical features that can be used as input to machine learning algorithms. Features can capture various aspects of a text document, such as its lexical, syntactic, and semantic properties. In this study, GloVe, Word2Vec, TF_IDF, and CountVectorizer were used as feature extraction techniques. GloVe and Word2Vec are pre-trained word embedding models that map each word in the corpus to a high-dimensional vector. These vectors capture the

semantic meaning of the words, allowing the model to learn accurate representations of the text (Kenyon-Dean et al., 2020). TF-IDF and CountVectorizer are traditional bag-of-words models that represent each document as a vector of term frequencies. While these models do not capture the semantic meaning of the words, they can still be effective at identifying patterns in the data (Patel et al., 2021).

A classifier is a machine learning algorithm that takes numerical input (in this case, the features from the NLP models) and outputs a classification. We experimented with several different classifiers, including decision tree, random forest, support vector machine (SVM), and gradient boosting (GB). Decision trees are a type of supervised learning algorithm that constructs a tree-like model of decisions and their possible consequences (Quinlan, 1986). Random forests are an ensemble of decision trees that generate multiple models and combine their outputs to improve accuracy (Svetnik et. al, 2003). SVM is a popular machine learning algorithm for text classification that finds the best hyperplane to separate data points into different classes (Noble, 2006). GB is another ensemble method that combines several weak classifiers to produce a stronger model (Chen & Guestrin, 2016). By using a variety of features and classifiers, we aimed to thoroughly understand and analyze the produced dataset.

Evaluation Metrics

Following the construction of different models, it is necessary to evaluate the performance of the classification model. This helps to determine the model's ability to accurately classify new data into the correct categories. To evaluate the performance of our classification algorithm, we employed various methods such as accuracy, precision, recall, and F1-score, which are defined as follows:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$\text{Precision } (P) = \frac{TP}{(TP + FP)}$$

$$\text{Recall } (R) = \frac{TP}{(TP + FN)}$$

$$F1 = \frac{2 \times P \times R}{(P + R)}$$

$$P_{weighted} = \sum_{i=1}^{L} (P \text{ of } i \ \times \text{Weight of } i)$$

$$R_{weighted} = \sum_{i=1}^{L} (R \text{ of } i \times \text{Weight of } i)$$

$$F1_{weighted} = \sum_{i=1}^{L} (F1 \text{ of } i \times \text{Weight of } i)$$

where TP stands for true positives, TN stands for true negatives, FP stands for false positives, and FN stands for false negatives. Accuracy is the proportion of correctly classified instances out of the total number of instances. The F1-score is the harmonic mean of precision and recall, which considers both false positives and false negatives. Precision is the proportion of true positive predictions out of the total number of positive predictions, while recall is the proportion of true positive predictions out of the total number of actual positive instances (Hossin & Sulaiman, 2015).

Weighted precision, recall, and F1 score compute the weighted average for each metric, with the weight proportional to the number of instances in each class. Weighted statistics were calculated to give more weight to the model's performance on the majority class of comments.

Ethics and Data Privacy

As a part of our commitment to ethical standards, we prioritize the protection of vulnerable individuals' privacy and confidentiality. To ensure this, we took the necessary steps to remove any identifying information such as user IDs, phone numbers, and addresses before sharing the data with our annotators. We recognize that data collected from social media can be particularly sensitive, especially when it pertains to marginalized communities such as the queer community. Therefore, we took great care to remove personal information to minimize any potential harm to individuals' identities. Our annotators were only given access to anonymized postings and were prohibited from contacting the author of the remark. Moreover, only researchers who agree to follow ethical criteria will be permitted to access the dataset for research purposes. We also provided our annotators with the option to opt out of the annotation process if they felt uncomfortable at any point (Gurav et al., 2019).

Results

Data Analysis

The resulting dataset was imbalanced in the number of queerphobic and non-queerphobic comments. Of the 10,000 comments labelled, 1,648 were queerphobic.

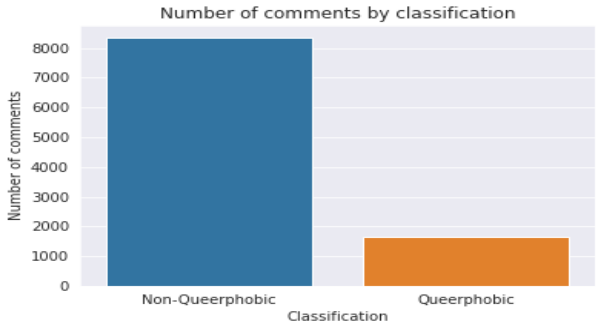

Figure 1. Number of comments by classification. 16.48% of the comments are queerphobic.

When comparing the most common words in queerphobic and non-queerphobic text, there were several differences. In the queerphobic text, the words “gender,” “children,” “tran” (the lemmatized version of “trans”), and “woman” occurred more frequently than in the non-queerphobic text. The much greater prevalence of the word “children” in the queerphobic comments as compared to the non-queerphobic comments may reflect a general tendency in transphobic rhetoric. Colliver (2021) explains that a common justification for transphobic arguments is that the advancement of trans rights threatens other communities. One specific way this rhetoric is employed is to demonize trans people as pedophiles forcing trans identity upon youth, which may explain the difference in the prevalence of the word “children.”

Another interesting difference between the most common words in the queerphobic and non-queerphobic comments is the greater presence of the word “love” in non-queerphobic comments. Halperin (2019) offers an inquiry into queer love which describes how love is central to queer activism, academia, and culture. The significance of love to queerness may have contributed to this difference.

a. Queerphobic Word Cloud

b. Non-Queerphobic Word Cloud

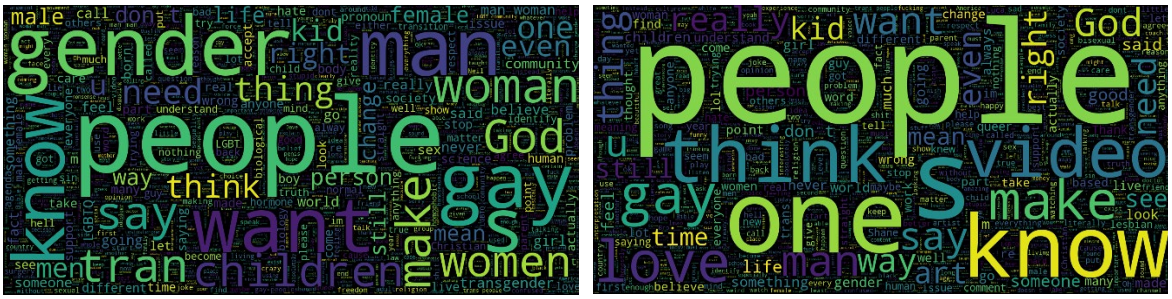

Figure 2. Word Clouds of Data. These word clouds illustrate the frequency of words in the queerphobic and non-queerphobic texts. Larger words represent higher frequencies.

On average, comments that were classified as queerphobic were wordier than non-queerphobic comments. The proportion of queerphobic comments that had 25 words or more was higher than the proportion of non-queerphobic comments that had 25 words or more.

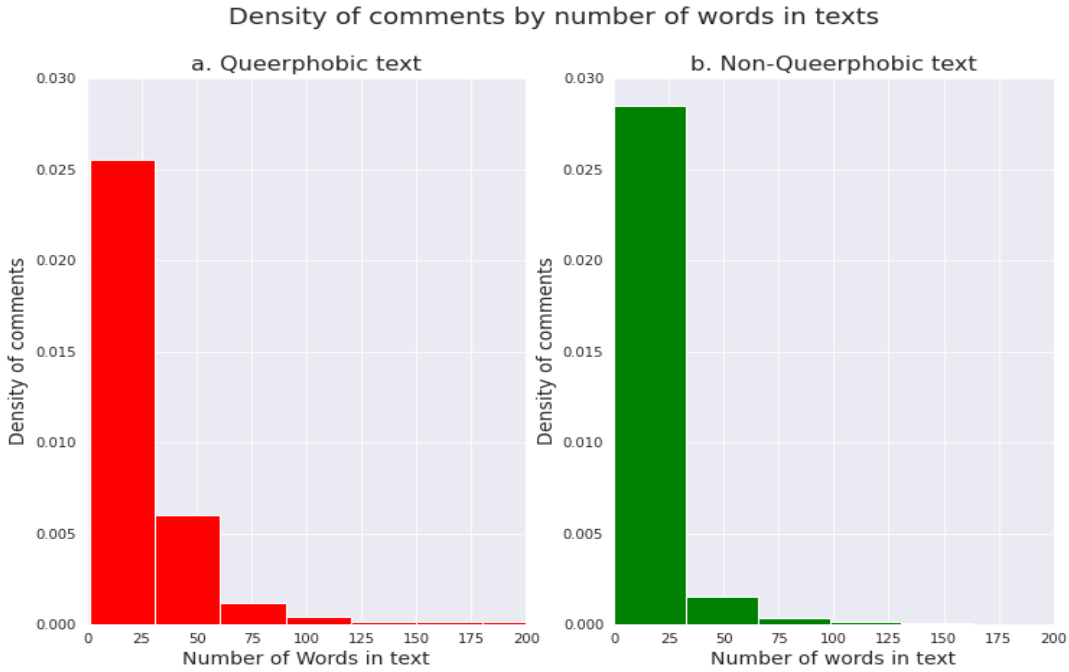

Figure 3. Density of comments by number of words in text. In comparison to the non-queerphobic comments, the queerphobic comments have a higher proportion of texts with more than 25 words.

Out of the entire corpus, some of the most common bigrams were “man woman,” “trans people,” “male female,” “lgbtq community,” and “gay people” (see Figure 4). These individual terms refer to queerness in some way. While “trans people” and “lgbtq community” are explicit references, annotators noted that the terms “man woman” and “male female” were often used in discussion about trans people. These queer-related terms are in the eight most common bigrams, which indicates that the dataset is well focused on queerness.

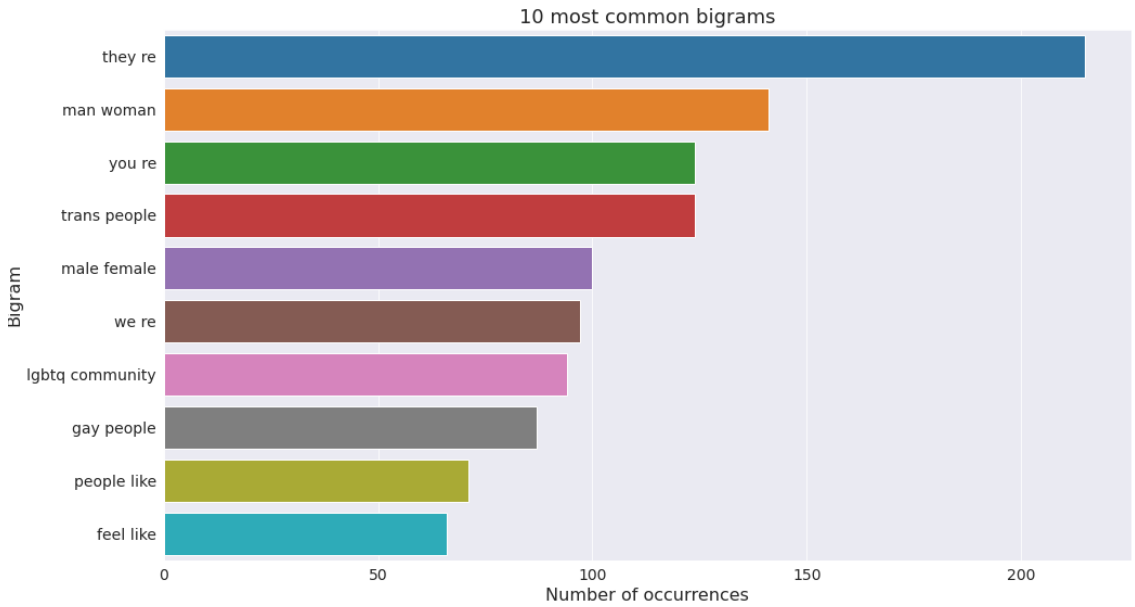

Figure 4. 10 most common bigrams in text. Of these terms, "man woman," "trans people," "male female," "lgbtq community," and "gay people" were identified as relating to queerness.

Model Analysis

To develop a baseline to understand how well models can learn from this data, 16 models were trained and tested on this data. Four classifiers were used with four feature extraction techniques.

Table 2. Analysis of data by different classifiers and feature extraction techniques. Acc, Pw, Rw, and F1w represent accuracy, weighted precision, weighted recall, and weighted F1 score. DT, RF, SVM, and GB are the decision tree, random forest, support vector machine, and gradient boosting classification models.

Classifier	Feature Extraction Technique	Acc	Pw	Rw	F1w
DT	GloVe	0.755	0.75	0.755	0.752
DT	Word2Vec	0.755	0.777	0.755	0.765
DT	TF-IDF	0.837	0.83	0.837	0.833
DT	CountVectorizer	0.842	0.84	0.842	0.841
RF	GloVe	0.829	0.786	0.829	0.764
RF	Word2Vec	0.846	0.814	0.846	0.814
RF	TF-IDF	0.848	0.822	0.848	0.8
RF	CountVectorizer	0.851	0.809	0.851	0.799
SVM	GloVe	0.827	0.683	0.827	0.748
SVM	Word2Vec	0.839	0.703	0.839	0.765
SVM	TF-IDF	0.856	0.834	0.856	0.822
SVM	CountVectorizer	0.86	0.84	0.86	0.814
GB	GloVe	0.829	0.786	0.829	0.779

GB	Word2Vec	0.841	0.812	0.841	0.817
GB	TF-IDF	0.866	0.851	0.866	0.854
GB	CountVectorizer	0.866	0.85	0.866	0.854

The feature extraction techniques that resulted in the highest evaluation metrics were TF-IDF and CountVectorizer. GloVe and Word2Vec, which performed worse than these, are both word embedding algorithms that represent words as dense vectors in high-dimensional space, where related words are located closer. These algorithms are more attuned to semantics. CountVectorizer and TF_IDF are much simpler algorithms that use word frequency to extract features from text. They don't account for semantics, and they treat each word individually. The superior performance of CountVectorizer and TF_IDF over GloVe and Word2Vec may indicate that the semantic relationships between words are not very important for the task of recognizing queerphobia, or that the dataset is not large enough for the word embeddings to be trained effectively. It could also mean that simpler techniques like CountVectorizer and TF-IDF are more effective at capturing the most important features for the task, such as the frequency of certain words or phrases.

Discussion

As discussed in the literature review, queerphobic hate speech marginalizes the queer community by threatening physical violence and provoking mental and emotional distress. It is therefore critical to recognize and moderate queerphobic hate speech. However, there is little data on digital queerphobia. This study aims to create a dataset for the identification of queerphobia to supplement insufficient data and further research on the moderation of queerphobic hate speech.

The dataset, which can be found at https://github.com/ShivumB/dataset-for-identification-of-queerphobia (Banerjee, 2023), is comprised of 10,000 comments sourced from 20 YouTube videos that represent queerness in a variety of ways. Three volunteers who identified as queer or queer allies manually annotated the data according to written guidelines. This data was preprocessed and used to train multiple NLP models, establishing baseline performance for NLP models. Accuracy, weighted precision, weighted recall, and weighted F1 score were used to evaluate the different models. As part of our commitment to ethical standards, we protected the privacy of commenters' whose information was collected, and we emphasized to annotators that they had the right to opt out of annotation at any point.

From the analyses performed on the data, several patterns were illustrated in the queerphobic and non-queerphobic data. While the queerphobic data had an emphasis on the word "children", which was often used to denigrate trans people as sexual predators, the non-queerphobic data had an emphasis on the word "love," which is an important part of queer activism and culture. Surprisingly, both queerphobic and non-queerphobic comments referenced the word "God" in roughly equal measure. While one might assume that religious references would oppose queerness in this context, many comments that used the word "God" were Bible verses that were not queerphobic in and of themselves.

The analyses performed on the models illustrated that the simpler feature extraction algorithms CountVectorizer and TF-IDF outperformed the more complex word embedding algorithms GloVe and Word2Vec. This may be because the semantic relationship between words is irrelevant to the classification of queerphobic comments, that CountVectorizer and TF-IDF captured important features that GloVe and Word2Vec missed, or that classifiers that used GloVe and Word2Vec did not have enough data to be trained effectively.

This dataset is valuable because it provides manually labelled annotations that are crucial for supervised learning NLP models. These models can be used to help moderate queerphobic hate speech, and in making this dataset, we hope to make social media a safer place for queer people.

Limitations & Future Research

Several limitations should be taken into consideration when interpreting the results of this study. First, the dataset used in this study was imbalanced, with a much smaller number of instances of the queerphobic class compared to the non-queerphobic class. To account for this imbalance, the data may have to be preprocessed to undersample the majority class. Future studies can aim to collect a more balanced dataset by choosing sources which are more likely to have a high proportion of queerphobic comments.

Second, comments were classified into the binary of queerphobic or non-queerphobic. There are many different types of queerphobia, and these may manifest differently. Reducing the complex task of recognizing the precise type of queerphobia to recognizing queerphobia makes annotation more time efficient, but this loses important information. For future research, it would be important to consider a more complex classification system.

Third, no analysis was performed on annotator reliability, which could potentially affect the accuracy of the annotations. We attempted to mitigate these limitations by selecting highly qualified annotators, providing clear instructions and training, and using well-established methods for data processing and analysis. However, future studies should perform analysis on annotator reliability.

Finally, the NLP methods used in this study were relatively simple. In future studies, it would be important to investigate more complicated NLP algorithms, such as Google's BERT word embedding model.

Overall, while the results of this study provide valuable information for the identification of queerphobic hate speech online, the limitations described above should be considered when interpreting the findings, and future research should explore these limitations.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Banerjee, S. (2023). Dataset for identification of queerphobia. *Github*. https://github.com/ShivumB/Dataset-for-Identification-of-Queerphobia

Bird, S. (2006). NLTK: The natural language toolkit. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, *1*(1), 213-217. https://aclanthology.org/P04-3031/

Chakravarthi, B., Priyadharshini, R., Ponnusamy, R., Kumaresan, P., Sampath, K., Thenmozhi, D., Thangasamy, S., Nallathambi, R., & McCrae, J. (2021). Dataset for Identification of Homophobia and Transphobia in Multilingual YouTube Comments. *ArXiv*. https://doi.org/10.48550/arXiv.2109.00227

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, *22*(1), 785-794. https://dl.acm.org/doi/10.1145/2939672.2939785

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International AAAI Conference on Web and Social Media*, *11*(1), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955

Delgado, R., & Stefancic, J. (1995). Ten Arguments against Hate-Speech Regulation: How Valid? *Northern Kentucky Law Review*. https://scholarship.law.ua.edu/facarticles/564

Engonopoulos, N., Villaba, M., Titov, I., & Koller, A. (2013). Predicting the resolution of referring expressions from user behavior. *Proceeding of the 2013 conference on empirical methods in natural language processing*, *1*(1), 1354-1359. Association for Computation Linguistics. https://aclanthology.org/D13-1134
Everytown. (2022). *Hate, violence, and stigma against the LGBTQ+ community.* Everytown Research & Policy. https://everytownresearch.org/report/remembering-and-honoring-pulse/
Gilbert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2*(1), 11-20. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5102
Gurav, V., Parkar, M., & Kharwar, P. (2019). Accessible and Ethical Data Annotation with the Application of Gamification. *International conference on recent developments in science, engineering, and technology, 1230*(1), 68-78. REDSET. https://doi.org/10.1007/978-981-15-5830-6_6
Halperin, David M. (2019). Queer Love. *Critical Inquiry, 45*(2), 396-419. https://doi.org/10.1086/700993
Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, e12432. https://doi.org/10.1111/lnc3.12432
Hubbard, L. (2020) Online Hate Crime Report: Challenging online homophobia, biphobia and transphobia. London: Galop, the LGBT+ anti-violence charity. https://galop.org.uk/resource/online-hate-crime-report-2020/
Kannan, S., & Gurusamy, V. (2014). Preprocessing Techniques for Text Mining. https://www.researchgate.net/publication/273127322_Preprocessing_Techniques_for_Text_Mining
Kenyon-Dean, K., Newell, E., & Cheung, J. C. (2020). Deconstructing word embedding algorithms. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *1*(1), 8479–8484. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2011.07013
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* New York University Press
Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology, 24*(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565
Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *Proceedings of the 11th International AAAI conference on web and social media, 11*(1), 221-230. International AAAI Conference on Web and Social Media. https://ojs.aaai.org/index.php/ICWSM/issue/view/271
Patel, A., & Meehan, K. (2021). Fake News Detection on Reddit Utilising CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine. *2021 32nd Irish Signals and Systems Conference (ISSC)*, *32*(1), 1-6. https://doi.org/10.1109/ISSC52156.2021.9467842.
Petrillo, M., & Baycroft, J. (2010). Introduction to manual annotation. *Fairview Research.* https://gate.ac.uk/teamware/man-ann-intro.pdf
QMUNITY. (2019). *Queer terminology from A to Q.* https://qmunity.ca/wp-content/uploads/2019/06/Queer-Glossary_2019_02.pdf
Swamy, S. D., Jamatia, A., Gämback, B. (2019). Studying generalizability across abusive language detection datasets. *Proceedings of the conference on Computational Natural Language Learning (CoNLL), 23*(1), 940-950. Association for Computational Linguistics. https://doi.org/ 10.18653/v1/K19-1088
Tsesis, A. (2002). *Destructive messages: How hate speech paves the way for harmful social movements.* New York University Press
Weitzel, L., Prati, R. C., & Aguiar, R. F. (2016). The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques? *Springer International Publishing Switzerland.* https://doi.org/10.1007/978-3-319-30319-2_3

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences, 43*(6), 1947-1958. https://doi.org/10.1021/ci034160g

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*(1), 81-106. https://doi.org/10.1007/BF00116251

Hossin, M., & Sulaimanm, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process, 5*(2), 1-11. https://doi.org/10.5121/ijdkp.2015.5201

Wissler, L., Almashraee, M., & Monett, D. (2014). The gold standard in corpus annotation. *5th IEEE Germany Student Conference, 1*(1), 1-4. https://doi.org/10.13140/2.1.4316.3523