An Exploration into the Utility of Sexually Explicit Comments Data Using Natural Language Processing

Lemon Lin Reimer

Columbia University

Quantitative Methods in the Social Sciences

Abstract

The current research endeavors to utilize Sexually Explicit Media (SEM) comments to produce ecologically valid information regarding SEM consumer activities, as well as explore the quality of the information held within comments data. These goals will be achieved using archival comments data scraped from the SEM-hosting pornographic website xhamster with a specific emphasis on ‘roleplaying’ behaviors. Roleplay, in this context, is defined as any form of ideation related to the content of the video. These comments will be analyzed through the theoretical lens of parasocial interaction, where an intimate relationship is formed between a viewer and an actor where the actor cannot reciprocate (Horton & Wohl, 1956). Specifically, this research will look at two possible manifestations of parasocial interaction: participatory and observational. Both participatory and observational roleplay are determined by camera angle, where the consumer is placed in a third-person observational perspective or a first-person participatory perspective. The research question at hand is “Does archival comments data meaningfully contribute information regarding the content of the video or the behaviors of consumers?” The hypotheses are as follows: 1) Archival comments data is predictive of the camera angle (observational/3rd person perspective or participatory/1st person perspective), 2) Archival comments data contains themes of expressing sexual behaviors, and 3) Archival comments data contains themes of expressing sexual preferences.

*Keywords:* sexually explicit media, natural language processing, classification

An Exploration into the Utility of Sexually Explicit Comments Data Using Natural Language Processing

It is no secret that Sexually Explicit Media (SEM), such as online pornography, is a pervasive media type that is highly accessible and impactful to human behavior. Pornography is consumed at high rates in the United States, with one study on early adolescent exposure reporting 66% of males and 39% of females in their sample experiencing SEM that year (J. D. Brown & L’Engle, 2009). However, most empirical research into SEM consumption use self-report measures, interviews, or laboratory experiments as their methodology (Weitzer, 2011). Low external validity, coupled with a high prevalence of third-person effects (Hald et al., 2015; Lee & Tamborini, 2005; Nelson, Leickly, Yang, Pereira, & Simoni, 2014; Reid, Byrne, Brundidge, Shoham, & Marlow, 2007), make current findings suspect regarding the generalizability of the sexual behaviors expressed by participants. Researchers in sexology have also underprivileged the growing archival data produced by SEM consumption online such as user comments on pornographic websites. This study endeavors to utilize this unused data to produce ecologically valid information regarding SEM consumer activities, as well as determine the efficacy of SEM comments data as a resource.

*Presence, Immersion, and Parasocial Relationships in Visual Media*

Parasocial interaction (PSI) is a social relationship first introduced by Horton and Wohl (1956), who define it as a non-reciprocated relationship between a person and a media persona, real or fictional. This form of social interaction is seen in a wide variety of medias such as hip hop music videos, Star Wars, and sports (W. Brown, Basil, & Bocarnea, 2003; Hall, 2019; Shin, Song, Kim, & Biocca, 2019) as well as across developmental stages (Hoffner, 1996; Hoffner & Buchanan, 2005; Rubin & McHugh, 1987). However, very few studies have looked at pornographic media through a parasocial lens.

Generally, PSI is framed as a dyadic relationship akin to a one-sided friendship, where the participant and the media persona are both separate individuals. However, studies have also been done on wishful identification, where the participant attempts to become the persona through emulation or by maintaining an emotional and cognitive connection with the persona outside of the media setting (Bui, 2017). While PSI does not always involve emulation, wishful identification is defined by emulation of the media persona. Generally, wishful identification is seen as less common, while PSI is a very common social activity (Bui, 2017; Horton & Wohl, 1956). These two phenomena are not perfectly distinct, and they share many attributes. For one, they both tend to follow the similarity-attraction principle, where people generally find others with similar attributes to be attractive because they are familiar (Bui, 2017). This is true about both choices of media persona for PSI and wishful identification, where men are more biased towards male media personae and women are gender ambivalent (Bui, 2017; Hoffner, 1996; Hoffner & Buchanan, 2005). In a sample of 155 children aged 7 to 12, 91.1% of the boys and 52.6% of girls chose a favorite television character of the same gender (Hoffner, 1996). Similarly, the study conducted by Bui found that PSI, wishful identification, and gender of participant predicted the gender of the media persona chosen as a favorite celebrity where higher PSI, higher wishful identification, and being a female participant all increased the chance of the chosen media persona being female (2017). In general, Bui also found the same trend as Hoffner, where men had a high preference for male celebrities but women only chose female celebrities slightly more often than male ones. In multiple studies, wishful identification and PSI have been highly correlated where higher PSI leads to higher wishful identification and vice versa (W. Brown et al., 2003; W. J. Brown, 2009). Some studies treat these constructs as distinct (Bui, 2017) while others treat them as two behaviors held under the umbrella term of ‘media involvement’ (W. Brown et al., 2003). The current research falls under the latter interpretation, where parasocial interaction is the overarching concept and wishful identification is a subset behavior.

The underlying theories behind PSI and why it occurs fall into two major camps: uses and gratifications theory and uncertainty reduction theory (Rubin & McHugh, 1987). Uses and gratifications theory is comprised of three assumptions: “(a) people are goal directed in their behavior, (b) they are active media users, and (c) they are aware of their needs and select media to gratify these needs,” (Rubin & McHugh, 1987, p. 280). Uncertainty reduction theory is similar, in that it assumes that people use passive and active strategies in order to reduce uncertainty about a media persona. Based on the uses and gratifications framework, it is possible to experience PSI fleetingly, as an aspect of immersion. However, uncertainty reduction theory insinuates some level of repetition or prolonged exposure as an aspect of developing a sense of consistency within the media persona.

In Horton and Wohl’s original explanation of the PSI phenomenon, they highlight the autonomy of the participant, the importance of camera angles, and the consistency of the actor as foundational to PSI (1956). In the way of participant autonomy, the assumption is that the participant is ultimately knowledgeable and in control of what they consume. If the media persona becomes unfavorable or if the PSI is no longer fulfilling, the participant can withdraw from the interaction at any time. The only way in which this relationship is controlled by the persona is that the media persona crafts their role to be interactive to an extent, and so the type of interaction is guided by the content of the media itself. Through Horton and Wohl’s original interpretation, there is support for both theories: the interaction is strengthened by minimizing uncertainty and being guided by the persona, but the participant is also goal-driven in their admission into the PSI and can withdraw if the interaction is unfavorable. In the case of SEM, there are few chances for uncertainty reduction since the actors are generally anonymous and interchangeable. For this reason, it is believed that the PSI occurring between SEM actors and consumers is closer to the uses and gratification theory than the uncertainty reduction theory.

Roleplay, defined by the researchers as media interaction using ideation or hypotheticals involving the participant and one or more SEM actors, is a parasocial behavior. Overall, the experience of roleplay in the SEM context is to publish a narrative online that includes the consumer as a part of the fantasy. However, where the consumers place themselves is variable and is believed to fall into two possibilities. One option is that consumers use a similar process to wishful identification, where they replace an actor or strongly identify with an actor. Examples of this behavior would be phrases like “It should be me in this scenario” or “I want to be this actor; they seem to be having an amazing time.” For this research, the behavior described will be called ‘participatory’ roleplay. The other option is for participants to insert themselves into the scenario without replacing or identifying with any actors. Examples of this behavior would be “I wish I was there with them, they’re both so beautiful” or “I love watching them, it would be even better if I was in the room.” For this research, the behavior described will be called ‘observational’ roleplay. As a disclaimer, both observational and participatory roleplay are capable of being passive or active, where observational roleplay can involve direct interactions with one or more actors and participatory roleplay can involve voyeur elements where the actors are not directly interacted with. The terminology of participatory and observational roleplay exists to separate the traditional parasocial dyad (observational) from the wishful identification (participatory). These two forms of roleplay comments constitute the dependent variable of interest in the current study.

Camera angle plays a major role in the immersive quality of SEM, and so it is possible that more immersive camera angles could increase the likelihood of a parasocial interaction. At this point in time, only two studies have researched the impact of immersion in SEM using virtual reality (VR) pornography compared to traditional, two-dimensional pornography: Elsey, van Andel, Kater, Reints, and Spiering (2019) and Simon and Greitemeyer (2019). Notably, both of these studies were published in the past year. Each compares virtual reality (VR) pornography to traditional, two-dimensional pornography. They both define immersion as being the technical restraints of viewership such as the quality of the video, the camera angle (first person perspective being more immersive than third person perspective) and the scope of what is visually accessible information available to the consumer. Presence, on the other hand, is defined as the subjective experience of being placed in the virtual environment. Elsey and colleagues found significant differences between males and females in their study, where men were more aroused by VR pornography than two-dimensional pornography but there was not a significant difference in arousal for women. Camera angle, however, caused a significant shift in arousal for both men and women, where a first-person participatory perspective was more arousing than a third-person observational perspective. Therefore, in the current study, camera angle is considered as a potential indicator to whether or not the video is simulating a participatory or observational experience.

*The Third Person Effect and Measurement of Sexual Behaviors*

While the majority of men who have sex with men (MSM) recognize the impact of unsafe sex act depictions, they rarely report being personally negatively affected (Hald et al., 2015; Nelson et al., 2014). This is known as the third-person effect, where “people perceive media’s impact on the attitudes and behavior of others to be greater than it is on themselves” (Lee & Tamborini, 2005, p. 294). The third person effect was first theorized by W. P. Davison in 1983, and has be fundamental to media effects research ever since, with particular influence on the study of socially undesirable media that is perceived as having a negative impact on consumers such as Sexually Explicit Media (SEM) (ibid). In application, the third-person effect has been observed in both MSM and heterosexual mixed-gender samples.

In the case of MSM, the negative outcome of interest is often condomless anal sex. Ninety-three percent of Hald and colleagues’ (2015) sample of MSM reported personally experiencing positive effects from SEM exposure such as gaining a better understanding of their sexual orientation and preferences as well as increasing their enjoyment of sex. However, only 7% reported personally experiencing any negative effects due to SEM exposure. Nelson and colleagues (2014) also found that their interviews with MSM showed an overwhelmingly positive personal experience with SEM including the educational and sexuality-affirming effects. These interviews also gave evidence to strong third-person effects, where the MSM interviewed showed concern for how SEM negatively impacts other MSM but not themselves. The participants were concerned that SEM set expectations for risky sex behaviors in other MSM and normalizes sex without condoms in the MSM community, but did not report being personally swayed to change their sexual behaviors or perform riskier sex acts due to SEM.

The same effect is found in heterosexual men and women, but with different concerns. In the study conducted by Lee and Tamborini (2005), the negative impact was on moral values concerning sex and attitudes towards the opposite sex. They found that the majority of participants (45.7%) perceived greater harm to others than to themselves. In the study conducted by Reid and colleagues (2007), the third person effect focused on male normative and female normative responses to pornography. Participants were asked to rate how ‘aroused and excited’ (male normative) or ‘repulsed and offended’ (female normative) by SEM relative to average men and women. The researchers found both first person and third person effects. For first person effects, men perceived themselves as more aroused and excited than the average woman and women perceived themselves as more repulsed and offended than the average man. For third person effects, men perceived women as more repulsed and offended than themselves and women perceived men as more excited and aroused than themselves.

These findings on first and third person effects were largely produced through qualitative data such as interviews, anonymous surveys, and self-report measures. While some researchers would argue that qualitative research is the most effective method for understanding the effects of SEM (Attwood, 2005), I argue that self-report designs rely heavily on the assumption of authenticity regarding a still sensitive and personal subject that is conventionally kept private in the United States. While surely the findings of the researchers cited here are valuable insights into SEM consumption, the high prevalence of third person effects and the social incentive to report what is perceived as normative sexual behavior may skew research towards a spuriously positive result that totes the many psychological, social, and educational benefits of pornography. Even with the promise of anonymity, these results may also under-report the frequency at which pornography is consumed, the commonality of kink and paraphilia, and the extent to which pornography can be personally harmful or unrealistic.

While quantitative research does exist, they are often in laboratory settings and the SEM stimulus is received in an ecologically uncommon format. For example, the study conducted by Landry and colleagues (2016) involved placing sensors on participants genitals and using distraction-reducing goggles during SEM exposure. The SEM was also preceded by nature documentary clips so as to set baseline genital measurements. This scenario is highly uncommon to the private and intimate experience of autoeroticism experienced by regular consumers and may limit the extent to which participants can become sexually aroused.

At this time, there are not many studies in the social sciences that use online archival consumer data from pornographic websites that host SEM. A few studies by Markey and Markey have utilized keyword searches to find longitudinal or event-based changes in the frequency of pornographic searches (P. M. Markey & Markey, 2013; P. Markey & Markey, 2011), and PornHub itself investigates and publishes metrics of SEM video popularity such as keywords and view counts (“Pornhub Insights,” n.d.). However, no study to date has utilized the comments posted by consumers on online video pornography. It would greatly benefit the field of sexology to investigate this ecologically valid and public information using a mixture of quantitative, qualitative, and exploratory analysis. In the case of the pornographic website xhamster, comments often display the self-reported gender and sexuality of the consumer which gives opportunity for gendered comparisons as well as specialization into the commentary of MSM online. This data also gives an opportunity to delve into the scantly researched topic of sexual roleplay behaviors with others and with non-reciprocating actors such as those depicted in SEM videos.

**Method**

*Data*

The unit of analysis is individual comments posted on a selection of sexually explicit videos hosted on xHamster. Consumers may be sampled from multiple times, dependent on if they posted more than one comment within the sample of videos. Replies are excluded from analysis, since the main subject of interest is consumers’ interactions with the videos rather than with each other. The final sample consists of the top video results from a selection of search queries: Cuckold, VR, POV, and Hidden Cam. VR and POV were chosen to represent a participatory/first person perspective while Hidden Cam and Cuckold were chosen to represent a observational/third person perspective. The number of comments will be variable between each video, and it is possible that one or more of the videos will have no comments. Videos with no comments will be excluded from analysis. Comments that are corrupted past legibility will also be excluded from analysis in order to help maximize the accuracy of the predictive algorithm.

*Procedure*

First, the webscraper is launched in order to collect the initial data. The webscraper is a python program using the selenium package and the Google Chrome webdriver to launch a headless browser. This package is commonly used to automate interfacing with a website such as clicking buttons and typing into search bars. The package is not particularly efficient, but it targets xpaths and cleanly pulls html attributes and text held in multiple containers. It also supports cross-pagination, which is particularly useful for this application. After the driver is launched in headless mode, the program inputs each search query into the search bar and pulls links to the top videos into a list. Next, a function iterates through the list and scrapes the raw comments from each link alongside relevant metadata like location (by country), gender, sexuality, date posted in UTC, and other relevant information. Finally, the data is cleaned and organized into a dataframe for export.

Next, the thematic extraction. For this part, the gensim package in python was used with a focus on LDA modelling. Five topical extractions were produced: Two for first-person perspective videos (one lemmatized and the other clean text), two for third-person perspective videos (one lemmatized and the other clean text), and one without any subsetting and lemmatized.

Finally, the modelling. The dependent variable was the search query in order to assess if the content of the comments was context-driven and specific to the genre of the video. The independent variable was the content of the comment itself, which had been cleaned and vectorized two ways: using a count vectorization and using a tf-idf vectorization. In both cases, both uni-grams and bi-grams were used during the vectorization. After vectorization, the dataset went up to to 11,677 dimensions, which needed to be reduced. Using a simple Principal Components Analysis (PCA), the number of dimensions was reduced to around 125 which is significantly lower. After the PCA, the models were constructed using the random forest classifier from the scikit-learn package in python. The model was optimized through a grid search that tuned the hyperparameters of maximum depth and number of estimators. Finally, both the tf-idf and count models were tested against a subset of data excluded from the model and analyzed for accuracy. This process was repeated for the dependent variable of camera angle, where the model predicted if the video was participatory or observational.

**Results**

To start, the number of comments were relatively evenly distributed across the four genres of VR, POV, cuckold, and hidden cam, with an average of 548 comments per genre. Their popularity was also relatively even. VR videos received an average 84 upvotes, POV videos averaged 221 upvotes, cuckold videos averaged 416 upvotes, and hidden cam videos received an average 302 votes. This makes cuckold videos the most popular and VR videos the least popular. A graph generated by matplotlib in python depicts the average number of upvotes and downvotes by genre in the Tables and Figures section of this paper.

The first hypothesis was that archival comments data is predictive of the camera angle (observational/3rd person perspective or participatory/1st person perspective). In order to test this hypothesis, two classification algorithms were trained: one to predict point of view (observational or participatory) and another to generally predict the theme/genre of the video (cuckold, POV, VR, or hidden cam). Both of these models had poor accuracy, regardless of the vectorization method. When tf-idf vectorized and predicting genre, the model had an overall accuracy of 0.38 (38%). The general standard is 80% accuracy, so this is extremely low. Precision, which is the ratio of correct predictions for a genre to the actual number of comments from that genre. The ‘VR’ and ‘POV’ genres had the highest precisions at 0.46 and 0.41 respectively. The ‘cuckold’ and ‘hidden cam’ genres had the lowest precisions at 0.40 and 0.34, respectively. Since there are four possible classifications, a precision of 0.25 would be equivalent to chance, so this model is only marginally better than chance at correctly predicting the genre of the video that the comment was posted on. Using the count vectorized version of the model, the precision and accuracy worsened with an overall accuracy of 0.37 (37%) and a highest precision with ‘cuckold’ at 0.41.

The model predicting camera angle did not fare much better, despite simplifying the classification from four possible labels to two (observational or participatory). In this case, the count vectorized model outperformed the tf-idf vectorized version. The accuracy of the tf-idf model was 0.63 (63%), the precision for predicting participatory videos (POV, VR) was 0.54, and the precision for predicting observational videos (hidden cam, cuckold) was 0.65. Since there are only two possible classifications, a precision of 0.50 would be equivalent to chance, so this model is only marginally better than chance at correctly predicting the camera angle based on the content of the comments. The count model was marginally better at prediction than the tf-idf model, but not by a trivial amount. The accuracy of the model was 0.64 (64%), the precision for predicting participatory videos was 0.59, and the precision for predicting observational videos was 0.65.

The second and third hypotheses were that archival comments data contains themes of expressing sexual behaviors and that archival comments data contains themes of expressing sexual preferences. This was tested using five thematic extraction models. All models extracted a total of five topics comprised of ten words each and the results can be found in Tables and Figures section of the paper. The first model looked for overall themes in the data using lemmatized comments with stop words removed. This is likely the most fruitful of the models, with the most predictive topic having a probability of 0.312. The two models using cleaned comments with stop words kept was the least fruitful, since the stop words took up valuable space in the topics. However, it is worth including because of the nature of roleplay, where usually unimportant pronouns like ‘her’ and ‘him’ give important information about the actor of interest. Both the lemmatized and clean comment models did not significantly differ by camera angle.

**Conclusion**

This study attempted to introduce a novel methodology and data source to sexology research that has the potential to increase ecological validity and minimize subjectivity regarding a highly emotional and taboo subject in the United States. Previous research has relied on three major methodologies: experimental designs in laboratory settings, self-report surveys, and content analysis using subjective codebooks. Each of these approaches have benefits and detractions. Laboratory experiments allow for causal relationships and a greater degree of control so that just the main variables can be isolated. However, this line of research loses touch with lived experience and is hard to generalize into ecological contexts (Elsey et al., 2019; Landry et al., 2016; Simon & Greitemeyer, 2019). Self-report surveys are easy access and can be anonymous, allowing for quick and effective data collection with greater ecological validity than the laboratory design. However, self-report measures are susceptible to third-person effects and the desire to be behaviorally normative, which might moderate results to be excessively positive or underreport first person effects (Lee & Tamborini, 2005; Reid et al., 2007). Finally, content analysis allows for valuable insights into the media itself and the prevalence of various depictions such as risky sex behaviors (Attwood, 2005; Downing, Schrimshaw, Antebi, & Siegel, 2014; White, Dunham, Rowley, Reisner, & Mimiaga, 2015). The current research offered a novel pathway to observational research: the analysis of archival comments data on pornographic websites. It’s possible to handle comments data through a content analytic approach, but for the reasons stated here, this study also experimented with the implementation of natural language processing (NLP) for sexology.

This research was useful to clarify two research questions. First, do comments data have anything to say about sexual preferences or behaviors in consumers? The hypothesis was ‘yes.’ In order to answer this question, the data went through a thematic extraction process. The results of this process were mixed. On one hand, words that are theoretically associated with preference such as like, love, hot, and awesome appeared regularly in various themes, though they didn’t really follow any pattern. Based on video popularity, observational videos have a higher average number of upvotes than participatory videos so it would be plausible that these words would be more associated with the former than the latter. However, this is not the case. These words seem to appear as a theme equally for both camera angles. Comments on SEM videos may be naturally positively biased, making the comments section a space for kudos instead of contention. Negative valence words in general are rare to appear in these themes, with only ‘devalue’ being explicitly negative. Other words like expletives are likely to be positive in this context rather than negative. Other words related to physicality such as weight, race/ethnicity, sex, and genitalia were all also popular keywords, further suggesting that there is expression of preference in comments. Unsurprisingly for the medium, there are no words that would suggest that commenters are showing preferences towards personality traits. Overall, there is evidence towards the theory that comments are indicative of sexual preferences.

Sexual behaviors are a little more unclear. Many words in this context have multiple meanings, such as ‘fuck’ being both a descriptor and an action. One clear sexual behavior is commenting on ejaculation, which occurs in two themes. More ambiguously, keywords like would, want, see, watch, please, and feel are expressive of some kind of fantasy or ideation in relation to one or more of the actors. These keywords occur most frequently in the lemmatized observational model, which is somewhat contrary to previous studies. Elsey and colleagues (2019), as well as Simon and Greitemeyer (2019), gave evidence to the theory that first-person perspective videos are more immersive and therefore participatory in nature. According to these results, observational videos that use a third-person perspective generate more roleplaying behaviors than participatory videos. With this new information in mind, it is possible that videos with less immersive quality require that the consumer generate their own immersion by roleplaying with the actors in a non-reciprocal manner. By describing their fantasies in the comments section, they are raising their immersion and converting the video from an observational experience to a participatory one. These results are interesting, but are also incomplete. Further experimentation into this topic should consider mixing n-grams with topical extraction, so that the vague keywords of want and like can be paired with actual preferences and behaviors. Want what? Like what? By expanding these key terms and giving more context, these topical extractions may have more to say on the value of comments data for observational study.

The other research question at hand is, are the contents of comments dependent on the camera angle of the video? Sticking a little longer with the topical analysis, there was a frequent extraction of the keyword VR in these topics, especially when subset to only include participatory videos. This is indicative that at least one of the four genres (VR, POV, cuckold, and hidden cam) has context-relevant comments. However, to really investigate this question, some predictive modelling is necessary. The hypothesis was that comments are predictive of both genre and camera angle, but a simple linear model would not be effective to answer this question. In traditional regression analysis, a value like comments data is a factor with far too many levels and functionally no repetition, where every comment is unique. Due to the complex nature of the prediction data, this research has utilized a classification algorithm for the task. The scikit-learn random forest classifier in python uses unsupervised regression modeling ‘behind-the-scenes’ in order to model the unique characteristics of comments in different genres or camera angles. Then, the algorithm applies this model to a novel set of pre-labelled data and attempts to classify each comment into the range of possible categories. This functionally expresses whether or not the complex information introduced to the model through comments data is an effective predictor of the content of the video itself, whether that be defined as camera angle or genre.

When predicting genre, the model was promising but ultimately ineffectual. Some genres performed better than others, with VR gaining 46% precision. This was foreshadowed by the thematic extraction, where VR was mentioned often in participatory videos and not observational ones, making the keyword a strong delineator between the two camera angles. In this dataset, it seems that the VR keyword occurs more often in VR videos than in any other genre of SEM video, which is understandable. However, the other genres have far less context-dependent comments. This may be due to the blanket positivity of the comments section, which homogenizes the dataset across categories. There may also be an issue of excessively short comments that are comprised of only one or two words that are descriptive but generic, leading to an oversaturation of context independent keywords.

In the case of predicting camera angle, the data became even less context dependent. The accuracy and precision of the classifications was only marginally different from chance after dimension reduction, hyperparameter tuning, and manipulation of vectorization and n-gram size, which is a bad sign for the utility of comments data in this context. While the conventions of qualitative analysis would expect that comments data would be interlinked with the content consumed, this is seemingly not the case. Potentially a different model structure such as naïve Bayes or neural networks would be more effective at identifying and applying any unique characteristics of the comments data towards the prediction of content. It is also possible, based solely on the thematic extraction, that this model structure would be better suited to predict the gender of the actors displayed which would make for an interesting sequel to this exploration. Ultimately, though there are hints of some context-relevant information inside comments data, this procedure did not extract it in a meaningfully predictive manner.

*Limitations and Future Research*

There are some glaring limitations to this research that are worth mentioning so that future research can be prepared to handle these obstacles. The most persistent issue was with cleaning the data. The internet is without borders and comments come from a wide range of locations. The data was overcrowded with so many languages that it was difficult to English or frankly any singular language. ASCII formatting, regular expressions, and comparison to an English corpus were all ineffective to varying degrees. By the end of this study, the data was still littered with Spanish and German comments, though the approaches listed were able to discard a fair amount of other languages such as simplified Chinese and Sanskrit. These non-English words were also common enough to appear in the thematic extraction which is concerning.

Furthermore, text-speak and misspelling were extremely common in English comments, which further sullied the data. Condensed phrases like ‘ur’ and exaggerated words like ‘aaaaand’ made their way into the thematic extraction as well, further taking away space from more theoretically relevant keywords. It is possible that this issue of ‘dirty data’ alone could be responsible for the inaccuracy of the classifier, which is a mark against the ease of utility for this data. Compared to more straightforward data such as survey responses, this data was significantly more difficult to clean. A more minor concern regarding the dataset is that comments can be short. One-word comments like ‘hot’ or ‘nice’ are not particularly enlightening and certainly not context-specific. Future studies may want to consider excluding comments that fall under a specific character or word length in order to prioritize comments that contain more information.

With these pitfalls of the study aside, this exploration into comments data is exciting and opens a new frontier for sexology research. In the digital age, observational designs for a highly private behavior are now possible using the breadcrumbs of interfacing with SEM on the internet. Admittedly, this data seems mostly concerned with autoeroticism, so researchers invested in intercourse are not likely to benefit from this new resource. However, many researchers interested in arousal, sexual preference, SEM exposure, pornographic content analysis, and autoeroticism could stand to benefit from archival data. This research only used one theoretical framework (parasocial interaction) in order to speculate on why comments are posted and what content is likely to occur in comments. Other theories, hypothesized behaviors, or SEM genres could change the outcome of this study entirely. Even just switching out the current keywords with new ones would introduce new information to the literature, since this data is so unexplored.

It doesn’t end with comments, either. Machine learning is a new frontier of its own that is making waves in the social sciences, sexology included. There are countless utilities for classifiers, machine vision, speech recognition, facial recognition, and natural language processing in the realm of sex research. For example, studies in content analysis could use facial recognition and machine vision to make arguments regarding objectification, which is often operationalized as the inclusion or exclusion of a face (Landry et al., 2016; Shim, Kwon, & Cheng, 2015). The purpose of this study isn’t to make an exhaustive list of future uses of these new resources, but this will hopefully inspire future endeavors using a new and exciting methodology for sexologists.

References

Attwood, F. (2005). What do people do with porn? Qualitative research into the consumption, use, and experience of pornography and other sexually explicit media. *Sexuality & Culture: An Interdisciplinary Quarterly*, *9*(2), 65–86. https://doi.org/10.1007/s12119-005-1008-7

Brown, J. D., & L’Engle, K. L. (2009). X-rated: Sexual attitudes and behaviors associated with US early adolescents’ exposure to sexually explicit media. *Communication Research*, *36*(1), 129–151. https://doi.org/10.1177/0093650208326465

Brown, W., Basil, M., & Bocarnea, M. (2003). The Influence of Famous Athletes on Health Beliefs and Practices: Mark McGwire, Child Abuse Prevention, and Androstenedione. *Journal of Health Communication*, *8*, 41–57. https://doi.org/10.1080/10810730305733

Brown, W. J. (2009). Mediated Influence of Pope John Paul II. *Journal of Communication & Religion*, *32*(1).

Bui, N. H. (2017). Exploring similarity characteristics, identification, and parasocial interactions in choice of celebrities. *Psychology of Popular Media Culture*, *6*(1), 21–31. https://doi.org/10.1037/ppm0000082

Downing, M. J. Jr., Schrimshaw, E. W., Antebi, N., & Siegel, K. (2014). Sexually explicit media on the Internet: A content analysis of sexual behaviors, risk, and media characteristics in gay male adult videos. *Archives of Sexual Behavior*, *43*(4), 811–821. https://doi.org/10.1007/s10508-013-0121-1

Elsey, J. W. B., van Andel, K., Kater, R. B., Reints, I. M., & Spiering, M. (2019). The impact of virtual reality versus 2D pornography on sexual arousal and presence. *Computers in Human Behavior*, *97*, 35–43. https://doi.org/10.1016/j.chb.2019.02.031

Hald, G. M., Træen, B., Noor, S. W., Iantaffi, A., Galos, D., & Rosser, B. R. S. (2015). Does sexually explicit media (SEM) affect me? Assessing first-person effects of SEM consumption among Norwegian men who have sex with men. *Psychology & Sexuality*, *6*(1), 59–74. https://doi.org/10.1080/19419899.2014.984516

Hall, A. E. (2019). Identification and parasocial relationships with characters from Star Wars: The Force Awakens. *Psychology of Popular Media Culture*, *8*(1), 88–98. https://doi.org/10.1037/ppm0000160

Hoffner, C. (1996). Children’s wishful identification and parasocial interaction with favorite television characters. *Journal of Broadcasting & Electronic Media*, *40*(3), 389–402. https://doi.org/10.1080/08838159609364360

Hoffner, C., & Buchanan, M. (2005). Young adults’ wishful identification with television characters: The role of perceived similarity and character attributes. *Media Psychology*, *7*(4), 325–351.

Horton, D., & Wohl, R. (1956). Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance. *Interpersonal and Biological Processes*, *19*(3), 215–229. https://doi.org/10.1080/00332747.1956.11023049

Landry, S., Goncalves, M. K., & Kukkonen, T. M. (2016). Assessing differences in physiologic subjective response toward male and female orientated sexually explicit videos in heterosexual individuals. *Canadian Journal of Human Sexuality*, *25*(3), 208–215. https://doi.org/10.3138/cjhs.253-A4

Lee, B., & Tamborini, R. (2005). Third-Person Effect and Internet Pornography: The Influence of Collectivism and Internet Self-Efficacy. *Journal of Communication*, *55*(2), 292–310. https://doi.org/10.1111/j.1460-2466.2005.tb02673.x

Markey, P. M., & Markey, C. N. (2013). Seasonal variation in internet keyword searches: A proxy assessment of sex mating behaviors. *Archives of Sexual Behavior*, *42*(4), 515–521. https://doi.org/10.1007/s10508-012-9996-5

Markey, P., & Markey, C. (2011). Pornography-seeking behaviors following midterm political elections in the United States: A replication of the challenge hypothesis. *Computers in Human Behavior*, *27*(3), 1262–1264. https://doi.org/10.1016/j.chb.2011.01.007

Nelson, K. M., Leickly, E., Yang, J. P., Pereira, A., & Simoni, J. M. (2014). The influence of sexually explicit online media on sex: Do men who have sex with men believe they “do what they see”? *AIDS Care*, *26*(7), 931–934. https://doi.org/10.1080/09540121.2013.871219

Pornhub Insights: Digging deep into the data. (n.d.). Retrieved November 11, 2019, from https://www.pornhub.com/insights/

Reid, S. A., Byrne, S., Brundidge, J. S., Shoham, M. D., & Marlow, M. L. (2007). A Critical Test of Self-Enhancement, Exposure, and Self-Categorization Explanations for First- and Third-Person Perceptions. *Human Communication Research*, *33*(2), 143–162. https://doi.org/10.1111/j.1468-2958.2007.00294.x

Rubin, R. B., & McHugh, M. P. (1987). Development of parasocial interaction relationships. *Journal of Broadcasting & Electronic Media*, *31*(3), 279–292. https://doi.org/10.1080/08838158709386664

Shim, J. W., Kwon, M., & Cheng, H.-I. (2015). Analysis of representation of sexuality on women’s and men’s pornographic websites. *Social Behavior and Personality: An International Journal*, *43*(1), 53–62. https://doi.org/10.2224/sbp.2015.43.1.53

Shin, M., Song, S. W., Kim, S. J., & Biocca, F. (2019). The effects of 3D sound in a 360-degree live concert video on social presence, parasocial interaction, enjoyment, and intent of financial supportive action. *International Journal of Human-Computer Studies*, *126*, 81–93. https://doi.org/10.1016/j.ijhcs.2019.02.001

Simon, S. C., & Greitemeyer, T. (2019). The impact of immersion on the perception of pornography: A virtual reality study. *Computers in Human Behavior*, *93*, 141–148. https://doi.org/10.1016/j.chb.2018.12.018

Weitzer, R. (2011). Pornography’s effects: The need for solid evidence. *Violence Against Women*, *17*(5), 666–675. https://doi.org/10.1177/1077801211407478

White, J. M., Dunham, E., Rowley, B., Reisner, S. L., & Mimiaga, M. J. (2015). Sexually explicit racialised media targeting men who have sex with men online: A content analysis of high-risk behaviour depicted in online advertisements. *Culture, Health & Sexuality*, *17*(8), 1021–1034. https://doi.org/10.1080/13691058.2015.1027738

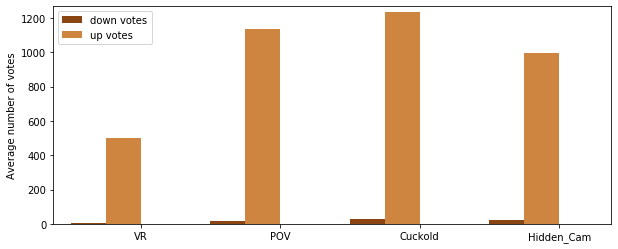
Tables and Figures

Table 1: *Thematic extraction models*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Topic | Keywords | Probability |
| Model 1 – Overall with lemmatization | 0 | Hot, nice, love, video, vr, awesome, fuck, sexy, please, get | 0.312 |
| 1 | Hot, good, girl, ur, wow, fuck, get, boob, one, dick | 0.19 |
| 2 | Sexy, come, perfect, aaaaaaaand, vr, xhamster, de, big, tit, yummy | 0.148 |
| 3 | Great, name, shit, beautiful, lady, damn, video, fat, ich, cock | 0.146 |
| 4 | Like, use, wife, love, vid, fuck, cum, make, vr, great | 0.16 |
| Model 2 – Participatory | 0 | Awesome, milf, sunny, que, du, je, omg, h, cute | 0.138 |
| 1 | Hot, I, her, she, so, sexy, love, is, great | 0.288 |
| 2 | Geile, a, tits, pussy, ist, die, und, auch, and | 0.109 |
| 3 | My, I, nice, to, and, a, the, me, in, ur | 0.182 |
| 4 | This, is, the, of, to, for, they, a, with | 0.191 |
| Model 3 – Observational | 0 | Please, her, la, my, me, more, de, elle, une | 0.13 |
| 1 | The, to, I, nice, a, and, her, of, is | 0.255 |
| 2 | Great, very, ich, und, nice, a, video, geil, die | 0.191 |
| 3 | She, I, a, it, and, to, the, that, her | 0.235 |
| 4 | Hot, the, I, love, is, her, this, to, a | 0.264 |
| Model 4 – Participatory with lemmatization | 0 | Girl, beautiful, good, sunny, sexy, sie, ist, suck, nicht, fuck | 0.108 |
| 1 | Great, sex, man, woman, men, like, devalue, love, make, feel | 0.166 |
| 2 | Love, awesome, tit, fuck, cock, amaze, vr, video, big, perfect | 0.242 |
| 3 | Hot, name, sexy, fuck, pussy, please, geile, die, wow, und | 0.206 |
| 4 | Nice, ur, like, get, hot, dick, cock, black, yoni, video | 0.119 |
| Model 5 – Observational with lemmatization | 0 | Girl, please, awesome, get, yes, perfect, cock, see, video, tit | 0.12 |
| 1 | Ich, fake, und, look, fuck, da, die, geil, video, auch | 0.104 |
| 2 | Hot, love, great, sexy, fuck, would, video, cum, as, pussy | 0.237 |
| 3 | Nice, fuck, wife, like, get, love, amaze, cock, good, watch | 0.207 |
| 4 | Like, great, lol, girl, mmm, r, good, know, look | 0.089 |

*Note: probabilities are cumulative across the discrete probabilities of each keyword.*

Figure 1: *Bar plot of upvotes and downvotes per genre*

**