The International Young Researchers' Conference
DOI: 10.34614/iyrc2021f21

FreeFeed: Combating Degenerate Feedback Loops With Linguistic Inferences From Human Interactions On Social Media

Aashika Jagadeesh
Fair Lawn High School, United States of America
aashika.jagadeesh@gmail.com

Abstract

Social network recommendation systems are frequently linked to encouraging polarization and widening ideological division, but this effect has rarely been examined in detail. Pernicious feedback loops are often created when these systems are trained with data that originates from users already exposed to algorithmic recommendations. This study analyzes the influence that feedback loops have on user mental health and assesses the effect of a Bayesian choice model (FreeFeed) in its ability to prevent the harmful reinforcement of views. At first, the Twitter API was filtered off of 4 factors: drugs, relationships, academics, and physical appearance. After 120,000 tweets were collected and preprocessed, the tweets were used to train/test a generalized logistic regression model and a multi-layer perceptron neural network. The models were compared on values such as the F1 score (max 0.963), AUC(max 0.990), and accuracy (max 93.7%). The algorithm was then implemented into an online simulation and tested on a set of social media users (n = 102) in New Jersey to identify both the impact of the revised model and the recommendation system model on self-esteem. Over the course of 3 weeks, participants completed a survey before and after use, in which responses were scored on the Rosenberg Self-esteem Scale. Significant statistical difference was determined between the revised model and the recommendation system model in the online simulations, which proves that policy makers and platform users should take these effects into consideration when they govern the use of feed algorithms.

Keywords: Machine Learning, Polarization, Human-Computer Interaction, Social Behavioral Science, Ethical AI

Introduction

Echo chambers and degenerate feedback loops function as metaphors that encapsulate the public fear that recommendation systems can manipulate user opinion by limiting the information users can consume online (Jiang et al., 2019). A primary concern is that recommendation systems combine with the tendency to communicate with like-minded individuals to create an environment that primarily presents opinion-reinforcing content to users. Recently, advances have been made in understanding the direct influence that recommendation systems have on the dissemination of fake news. According to Cohen (2018), users may believe misinformation as a result of algorithms that tailor cultural artifacts customized to the user in the form of a social distribution system. It was also noted how these algorithms sophisticate an understanding of social network analysis through their "invisibility" in the public eye. In this sense, users often fail to

understand how fake news integrates its way into the user feed. Further work was done on researching the correlation between echo chambers and political homophily, where researchers suggested that homophily was more apparent in the network of reciprocated followers than in the nonreciprocated network (Colleoni et al., 2014). In addition, studies conducted by the European Union Institute for Security Studies (EUISS) have demonstrated the relationship between echo chambers in social networks and the spread of vaccine misinformation online (Raemdonck, 2020).

Along with political bias and the spread of misinformation, various bodies of work have attributed mental health disorders to the extensive use of social media networks. For instance, in the Department of Psychology in Toronto (Hogue et al., 2018), 118 female undergraduate students were tasked with liking an image of an attractive peer and completing a visual analogue scale measure of state body image. It was concluded that young women who interacted with posts of an attractive peer experienced an increase in negative body image shortly afterwards. Further work done by researchers at the University of Cologne (Appel et al., 2015) helped establish a relationship between Facebook use, social comparison, envy, and depression. It was found that social comparisons and envy were common experiences amongst users on social media apps.

Although the notion of echo chambers is well-accepted, its direct influence on mental health and user consumption is not often well-understood. We seek to characterize the influence of feedback loops in the context of the recommendation system and study the consequences of algorithmic confounding on user self-worth. As these systems have been linked to altering user opinions and decisions, it is well within our ethical responsibilities to understand the system's implications for individual health and self-esteem. Issues of fairness and transparency in the creation of suggestion softwares must be considered. It is our hope that researchers will use our work to assess the impact of recommendation systems across all users and develop efficient models with reduced degeneracy.

The main contributions in this paper are as follows:

- This study proposes FreeFeed, a Bayesian *choice model*, which accounts for opinion-reinforcement and limited exposure to alternative views. The model is fair by preventing negative bias towards unpresented alternatives. In this sense, marginal probabilities of selecting specific options that were never presented to the user are independent of other choices.
- A simulation of FreeFeed was created by collecting 120,000 tweets from the Twitter API and training an NLP algorithm. A generalized logistic regression and a multi-layer perceptron were tried. The NLP algorithm was then implemented into an interactive website, along with a control algorithm that recreated the effects of feedback loops in recommendation systems.
- The website was tested on a set of social media users (n=102) from ages 15 to 25. Participants were split into three cohorts: a control exposed to an algorithm that resembled the effects of a traditional recommendation system, a group that used FreeFeed for 15 min per day, and a group that used FreeFeed for 30 minutes per day. The psychological study was conducted over the course of 3 weeks, and participant progress was regularly monitored.

Description of Choice Model

A discrete choice model determines the probability of a user selecting an option among K alternatives. It follows from the belief that utility of choice is a function of the properties of the choices and the characteristics of the individual making the choice (Columbia Public Health, n. d. b.). In a given scenario, we can suppose that K is

large, but users are limited in the options they can choose from. Sets of choices are made from selected subsets of alternatives, where C is an indication of all non-empty subsets of K (Train, 2002). In this study, we examine the probability of selecting an item K from a presentation C. Our proposed model, FreeFeed, conforms to the independence of irrelevant views. That is, sets of choices {Kt}t=1 Tare subject to chance variation and the probability of selecting a specific option is independent of other items in a given presentation {Ct}t=1 T.

This type of decision making problem is often regarded as a bandit problem (Lattimore et al., 2019). With a bandit framework, the Bayesian model serves two primary purposes: observations are used to better inform inferences of user preference and posterior samples induce effective presentation mechanisms, where the current best alternatives are explored. The posterior frames k with high priority either when k was rarely presented or frequently selected. This thus follows Thompson Sampling (Thompson, 1993), ensuring that underrepresented options will appear later on. For instance, in a subset {3,4}, if 3 does not appear in a single iteration, it is guaranteed a second presentation with Thompson Sampling.

Data Collection for Choice Model Simulation:

Both ground truth and test data were collected through the Twitter API (Twitter API, 2012). To communicate with the Twitter API and collect real time tweets, the Tweepy Python library (Tweepy Twitter API, 2020) was used. A python script was created with the Tweepy Twitter API and ran over the course of a month. Data was filtered on parameters that prevented the collection of retweets, tweets from automated accounts, companies, etc. Tweets were also filtered off of 4 factors: relationships, academics, drugs, and physical appearance (Table 1). Training data was filtered in this fashion to focus model function on identifying the relationship between degeneracy and user mental health.

High-Stress Factor	Presence in Tweet
Relationships	**Viewpoint #1:** tweets mention girlfriend, boyfriend, family, friends, parties, etc. **Viewpoint #2:** tweets mention being single, alone, or independent
Academics	**Viewpoint #1:** tweets mention college applications, assessments, GPA, professors, teachers, clubs, sports **Viewpoint #2:** tweets promoting calmness, meditation, self-confidence
Drugs	**Viewpoint #1:** tweets mention being drunk, commercializing excessive drug use, informal references of drugs **Viewpoint #2**: tweets mentioning downsides of excessive drug uses, discouraging substance abuse, providing support for drug addicts
Body Image/Physical Appearance	**Viewpoint #1:** tweets promoting weight loss, dieting, commercializing cosmetics and weight loss products, promoting modeling agencies, etc. **Viewpoint #2:** tweets promoting body positivity, natural beauty, etc.

TABLE 1: Tweet Filtering Criterion

However, to truly simulate the nature of the Bayesian choice model, the model was trained to distinguish between alternate perspectives. Tweets were then classified into high-stress factors and viewpoints (if applicable), yielding a dataset of ~800,00 tweets. Furthermore, a second round of validation was completed to remove misclassified and irrelevant content, which resulted in a dataset of ~120,000 tweets. Factors and perspectives were kept primarily balanced.

After data was collected and fully filtered, preprocessing was conducted. Unicode characters and stop words (ex. the, or, and) were removed and replaced with an empty string to ensure that there was no impact on model training. Lemmatization was then performed on the dataset to reduce words in the tweets to their basal/dictionary form. The Stemming technique was also used to reduce words to their word stem by inflection.

After the training matrix was fully formatted, a Generalized Logistic Regression Model from

the Scikit learn library was used to develop the first model (Scikit Learn, 2020). A Multilayer Perceptron Model (neural-network based approach) was then tried. Next, the models were created and scored on the data from a test matrix.

Website Design Implementation:

The trained models were implemented into a website created from Angular and a Node.js backend (Figure 1). The site was made to resemble the appearance of a user feed, with the option to view, like, and reload posts. A user's interaction history was stored in the Firebase API and used to determine which posts would appear next. In this way, each user's feed was unique to a user's individual preferences. Unlabeled tweet and image sets were created as a post bank, which could be retrieved from when the user pressed reload. All tweets and associated image sets were collected from Twitter Streaming API and filtered off of relevance to a factor in the filtering criterion (Table 1).

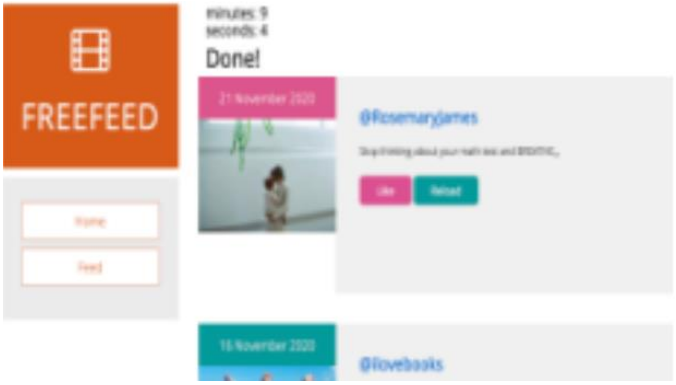

FIGURE 1: Website for model simulation

The unlabeled datasets served as input, in which the model classified the factor and viewpoint that was promoted. To be properly classified as demonstrating a specific factor or viewpoint, the model would need to pass a minimum confidence threshold of 0.65.

If the score passed the minimum threshold, the post would be labeled and used to determine which posts to present after reload is pressed. When the user interacts with labeled post, the post and its corresponding classification are stored in the Firebase API. Depending on the frequency of user interactions with identically labeled posts, the model would begin to alter the feed and introduce posts demonstrating alternate viewpoints.

Psychological Study:

Social networks were used to recruit participants, as this would ensure that all test subjects were social media users. The sample was constituted of random social media users in Fair Lawn, New Jersey. Parental approval was required for test subjects that were less than 18 years of age.

A self-esteem survey was provided to each of the participants prior to using the website, which entailed a total of 10 questions (shown in the Appendix). The questions included statements such as "I certainly feel useless at times" and "I take a positive attitude towards myself." Participants responded with how much they agreed with each of the statements by bubbling in the following: Strongly Agree, Agree, Disagree, Strongly Disagree. The responses were scored on the Rosenberg Self-esteem Scale of 30 points (Center of Disease Control, 2005). Each choice was worth a certain number of points from a scale of 0 to 3. After completing all 10 questions, a participant could have a minimum score of zero and a maximum of thirty. Scores between 15 and 25 are considered to be average, while a score exceeding 25 suggests an extremely high self-esteem; a score below 15 suggests an extremely low self-esteem (SRLab, 2014).

Participants were randomly assigned to 3 cohorts: a control that used a feed that resembled the effects of a traditional recommendation system for 15 minutes per day (Group 1), a group that used FreeFeed for 15 minutes per day (Group 2), and a group that used FreeFeed for 30 minutes per day (Group 3). Group 1 had 35 participants, Group 2 had 34 participants, while Group 3 had 33 participants. With these divisions, the cohorts were kept primarily balanced. The website had two settings: a sample feed with an

algorithm that mimics the function of recommendation systems and a sample feed that deploys the FreeFeed algorithm.

Results were primarily based off of self-esteem scores, which were compared before and after use to determine the direct effect of the choice model. Score differences were then averaged across the span of three weeks to serve as viable data points. Demographic data was also analyzed to determine basic characteristics of the sample population. Model accuracies were computed to determine how effectively the models simulated a particular choice model. Furthermore, $\chi 2$ tests were conducted to identify stochastic variables when users interacted with the feedback loop. All analyses were performed at a significance level of 5%, and tests were two-sided. The 95% confidence intervals (CI) of odds ratios (OR) were calculated as well.

Results

Demographics

105 participants initially were collected in total, with respondents consisting of males (n = 43) and females (n = 62). Three participants did not use the application for a long enough period of time before reporting their self-esteem score or fully interact with the FreeFeed interface for 3 weeks. Only 102 respondents were considered in the final data to reduce error and bias rates.

The breakdown of ethnicities was as follows: 42% of participants identified as Asian; 42% identified as Caucasian; 9% as African-American; 12% as Hispanic/Latin-American. All subjects were between ages 15 and 25.

Self-Esteem Scores:

Male self-esteem scores averaged to be greater than that of females [19.08 vs. 16.75; t = 1.720; p = 0.086] (Table 2). The average score difference for participants in Group 2 was 3.89, while the average score difference for Group 3 was 4.69. Group 1, on the other hand, exhibited results that varied significantly from Group 2 and 3 with a score difference of -2.24 (Figure 2). Both the Bayesian choice model [t = 2.027; p = 0.025] and the simulated recommendation system [t = 2.331; p = 0.012] instigated significant changes in user self-esteem (Table 3). There were a few outliers in the data, where participants exhibited no change in scores after extensive use. Some participants experienced changes that allowed for a 10-point increase, which resulted in significant fluctuations. In this sense, model training was not inclusive enough to have an equal impact on all individuals involved in the study.

Scaled Participant Self-Esteem Scores

Mean Score	18.64
Male Mean Self-Esteem Score	19.08
Female Mean Self-Esteem Score	16.75
Range	25*
Median	17

**5-30*

TABLE 2: Participants from all cohorts

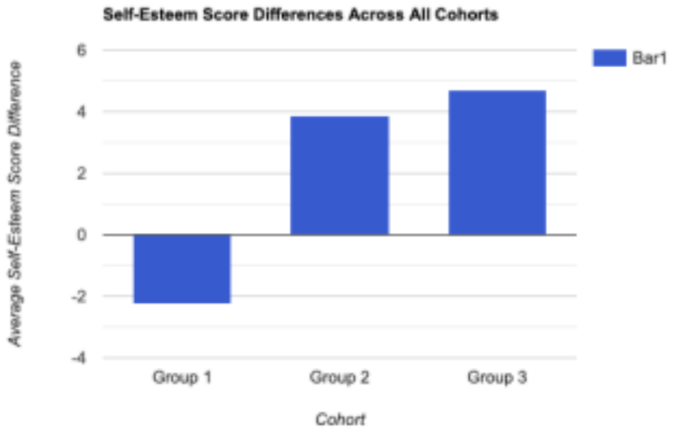

FIGURE 2

Users from group 3 exhibited more magnified increases in self-esteem scores, which indicates a positive relationship between the time spent interacting with the choice model simulation self-esteem. By the contrary, the relationship depicted by the control cohort follows a different trend: there was an average decline in self-esteem scores (Table 3).

Effect of Bayesian Choice Model (15 min) & Simulated Recommendation System (15 min) on User Self-Esteem

	α	σ	Mean difference	t-statistic	*p*
Bayesian Choice Model (FreeFeed)	0.05	7.182	3.89	2.027	0.025
	α	**σ**	**Mean difference**	**t-statistic**	***p***
Recommendation System	0.05	2.321	-2.24	2.331	**0.012***

*$p<0.05$

TABLE 3

A negative relationship can thus be seen between the usage of the recommendation system and the self-esteem scores of individuals.

One of the primary purposes of this was investigation was to determine if a discrete Bayesian choice model could have a significant influence on self-worth when compared to traditional recommendation systems. The results of univariate logistics demonstrate that there is significant statistical difference between both the choice model and recommendation system simulation in their effect on self-esteem [OR = 2.03; 95% CI OR = 1.48–2.82, $p < 0.05$] (Table 4).

Statistical Difference Between Bayesian Choice Model and Simulated Recommendation System

	Value
α	0.05
σ	0.896
z	1.65
$\overline{X}$	1.479

TABLE 4

Model Performance:

After models were fully trained, metrics were collected and analyzed. Some discussions follow. Loss was measured through Sparse Categorical Cross Entropy, and the loss was minimized after 200 epochs. The loss value of the Multi-Layer Perceptron Neural Network was 0.183, while the loss value of the Logistic Regression Model was 0.223. Using the formatted test matrix, the final accuracy of the Neural Network was calculated to be 93.7%, and the final accuracy of the Logistic Regression model was 88% (Figure 3).

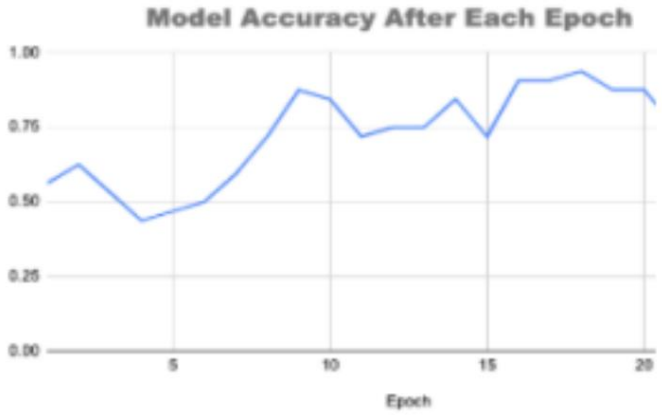

FIGURE 3: Model accuracy growth curve for the Logistic Regression Model for every 20 epochs

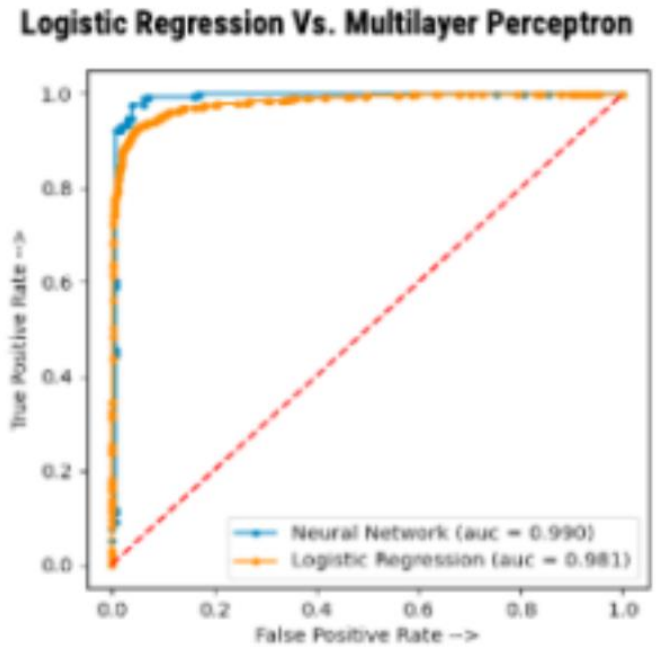

FIGURE 4

As an example, Figure 4 compares the ROC curves of the Neural Network and Regression Model. The curves are above the random decision line, which exists at, i.e., the (0,0) to (1,1) diagonal, thus indicating that they are good decision tests. Even at a True Positive Rate of

70%, almost all False Positives can be eliminated for both models. In this sense, the models are capable of making a rather accurate decision about classifications regarding linguistic structure. The AUC (area under the curve) was also measured, and the Multi-Layer Perceptron ultimately achieved the maximum AUC of 0.990. Taking into account the false and true positive rate, the F1-score was calculated. It was calculated separately for both views (Table 1), the macro average, and the weighted average (Table 5 & 6).

Logistic Regression	Precision	Recall	F1	Support
Viewpoint #1	0.89	0.93	0.91	998
Viewpoint #2	0.97	0.95	0.96	1,162
Macro Average	0.93	0.94	0.93	2160
Weighted Average	0.95	0.94	0.95	2160

TABLE 5: The Logistic Regression Model's Precision, Recall, F1-Score, and Support values are provided above. Viewpoints #1 and #2 refer to the classifications provided by Table 1.

Multilayer Perceptron	Precision	Recall	F1	Support
Viewpoint #1	0.96	0.93	0.94	998
Viewpoint #2	0.97	0.98	0.98	1,162
Macro Average	0.97	0.96	0.96	2160
Weighted Average	0.97	0.97	0.97	2160

TABLE 6: The Multi-Layer Perceptron Model's Precision, Recall, F1-Score, and Support values are provided above. Viewpoints #1 and #2 refer to the classifications provided by Table 1.

A Wilcoxon signed rank test (non-parametric) was used to quantify observed data and identify behavior that would result in significant statistical difference between the two models. Using the accuracy values, it was ultimately determined that there was significant statistical difference between the Logistic Regression Model and Neural Network [$p=0.013$; $p<0.05$]. A post-hoc analysis was then conducted, in which effect size was calculated to be 0.14. The classifications of effect sizes are traditionally small ($d = 0.1$), medium ($d = 0.5$), and large (d 0.8) (Sullivan, 2012). Under these guidelines, the effect size is small but not trivial.

Upon measuring the performance of the model through various mediums, it was ultimately determined that the Multi-Layer Perceptron was more effective for the problem at hand.

Chi-Square Test:

Tweets were separated into unigrams and bigrams, which were all stored into a vocabulary. A Chi-Square Test was completed on the tweets and their corresponding factors to determine stochastic variables, thus removing features that were irrelevant to classification.

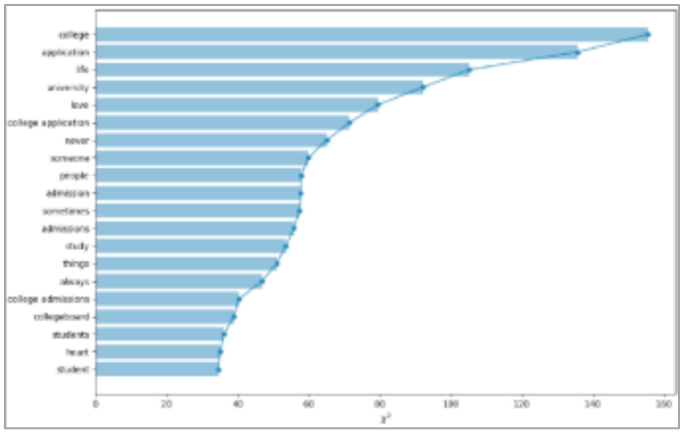

FIGURE 5: Academics

In the Figure 5, the most predictive word was "college." The intercept chi-square value was 49.2. In Figure 6, the most predictive word was "truth." The intercept chi-square value was 246.2. In Figure 7, the most predictive word was "body", while the intercept chi-square value was 38.3. In the Figure 8, the most predictive word was

"friendzone" with an intercept chi-square value of 27.3.

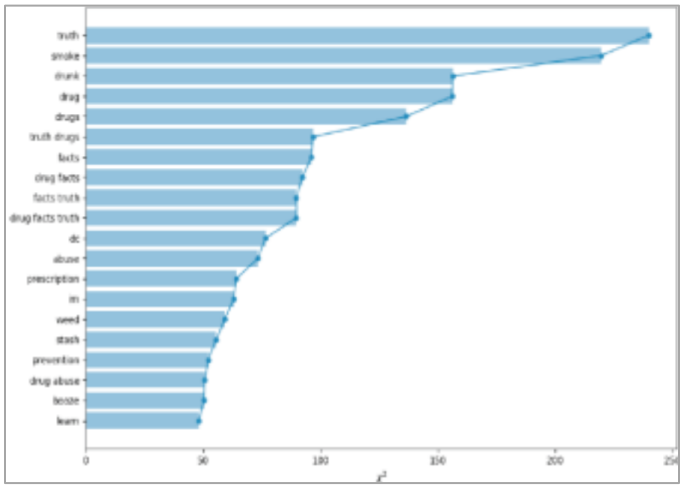

FIGURE 6: Drugs

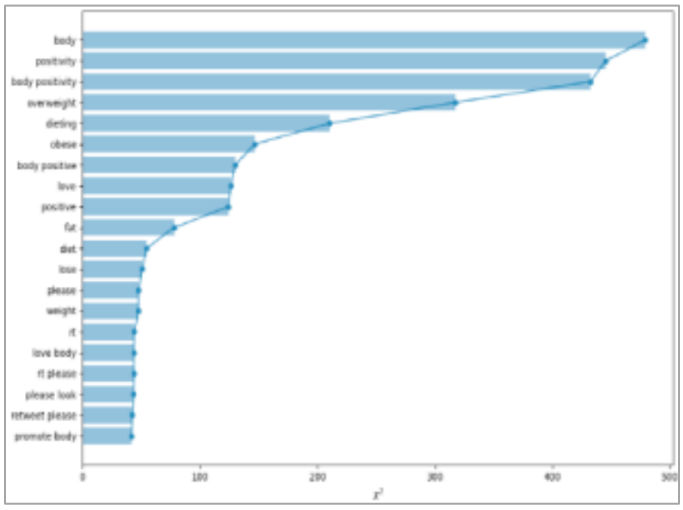

FIGURE 7: Body Image/Appearance

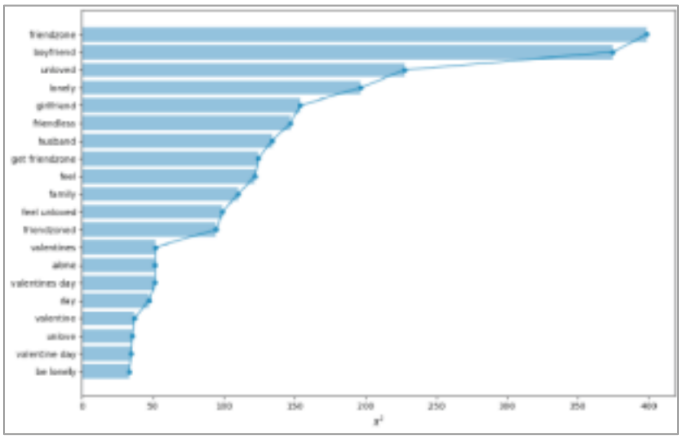

FIGURE 8: Relationships

Discussion:

In this paper, we researched degenerate preference systems, where users select options amongst a limited set of views. We propose a choice model that is aware of this bias and accounts for it by conforming to the independence of irrelevant views. The model was simulated through NLP techniques and assessed on participants for its influence on self-esteem.

A positive correlation was noted between the time exposed to the FreeFeed simulation and self-esteem scores. On the other hand, when users interacted with a model that mimicked recommendation systems, self-esteem scores decreased. After conducting statistical analysis, there was determined to be a significant statistical difference between the two simulations.

Conclusively, this work integrates elements from choice modelling, Bayesian inference, and self-esteem analysis to propose a novel idea that combats degenerate feedback loops in personalization systems.

Conclusion:

The FreeFeed model proved to have substantial benefit to user self-esteem, which indicates the potential for Bayesian choice models in the market. The harmful side-effects of self-reinforcing feedback loops necessitate the creation of models that investigate alternatives and learn to make the most optimal presentations. Our work can be further expanded to include a large sample size, further validating our findings and increasing the accuracy of our models. The model could also be assessed for its influence on body image, suicide ideation, and depression. Nonetheless, our findings may lead to the development of more complex model architectures, as well as a deeper understanding of the influence that recommendation systems have on users.

References

Jiang, R., Chiappa, S., Lattimore, T., György, A., Kohli, P. (2019). Degenerate Feedback Loops in Recommendation Systems. https://dl.acm.org/doi/abs/10.1145/3306618.3314288.

Cohen, J. N. (2019). Exploring Echo-Systems: How Algorithms Shape Immersive Media Environments. https://eric.ed.gov/?id=EJ1198674.

Colleoni, E., Rozza, A., Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. https://academic.oup.com/joc/article-abstract/64/2/317/4085994.

Raemdonck, N. V. (2020). The Echo Chamber of Anti-Vaccination Conspiracies: Mechanisms of Radicalization on Facebook and Reddit. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3510196.

Hogue, Jacqueline V., and Mills, Jennifer S. (2018). The Effects of Active Social Media Engagement with Peers on Body Image in Young Women. http://www.sciencedirect.com/science/article/pii.

Appel, Helmut, et al. (2015). The Interplay between Facebook Use, Social Comparison, Envy, and Depression. http://www.sciencedirect.com/science/article/pii.

Columbia Public Health. (n. d. b.). Discrete Choice Model and Analysis. https://www.publichealth.columbia.edu/research/population-health-methods/discrete-choice-model-and-analysis.

Train, Kenneth. (2002). Discrete Choice Models With Simulation. https://eml.berkeley.edu/books/train1201.pdf

Lattimore, Tor, and Szepesvári, Csaba. (2019). Bandit Algorithms. https://tor-lattimore.com/downloads/book/book.pdf.

Thompson, William R. (1993). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. 25(3/4):285–294

Twitter API. (2012). Twitter API Documentation. https://developer.twitter.com/en/docs/twitter-api.

Tweepy Twitter API. (2020). API Reference. https://docs.tweepy.org/en/stable/api.html.

Scikit Learn. (2020). Sklearn.linear_model.LogisticRegression. https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Center of Disease Control. (2005). Measuring Violence-Related Attitudes, Behaviors, and Influences Among Youths: A Compendium of Assessment Tools https://www.cdc.gov/violenceprevention/pdf/yv_compendium.pdf

SRLab. (2014). Rosenberg Self-Esteem Scale. https://www.sralab.org/rehabilitation-measures/rosenberg-self-esteem-scale

Sullivan, Gail M., et al. (2012). Using Effect Size—or Why the P Value Is Not Enough https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/