Application of Neural Topic Models for Exploration of Themes in Online Communities : Final Report

G075 (s2004127, s2450915, s2444768)

Abstract

In recent years, online male support groups, which are part of the so-called *manosphere*, have been put in the spotlight because of their rise in misogynistic language, calls to violence, and growing popularity. However, the unstructured nature of online communities means that it is challenging to interpret. Neural Topic Models have received little attention and provide a potential solution over classical methods like LDA. An obstacle in the way of expansion of adoption of Topic Modelling in Social Sciences is evaluation of the model outputs, for which human review remains a top choice. The development of rigorous methodology for human evaluation, however, has received little attention. In our project we explore the use of neural topic models and compare their performance to LDA using topic coherence and topic diversity as quantitative metrics. Furthermore, we then treat the themes found by social scientists as gold standard in our evaluation, and present two methods to evaluate results of our topic model: topic and gender keyword association tests and document topic comparison.

1. Introduction

To the digital humanities, online forums and social media platforms such as Facebook, Reddit and community forums are a gold mine of opinions and written accounts of individual's experiences. In recent years, online male support groups, which are part of the so-called *manosphere*, have been put in the spotlight because of their rise in misogynistic language, calls to violence, and growing popularity. While having all this data (much of it freely available) is a golden opportunity to extract narratives *en masse*, the unstructured nature of textual data means that it is challenging to interpret and analyze without putting lots of manual labor into reading and annotating the posts.

Topic modelling is a technique of text mining developed for discovery of themes or topics present in a corpus of texts. Importantly, it is a method for unsupervised classification of documents, which means that its goal is to extract these narratives without the need for annotating documents. The assumption is as follows: documents are assumed to belong to a series of topics and topics have a probability distribution over a fixed vocabulary. The challenge lies in evaluation of topic modelling output: quantitative measures might not capture everything and human evaluation is ill-defined, inherently biased and often reduced to eyeballing.

Theme analysis in the social sciences is the process by which themes are discovered in a corpus by social scientists. One obstacle is that the researchers are often limited to a small subsample of all documents and can be biased to topics researchers expect to find. To tackle the labor challenge, social scientists have used traditional topic modelling methods like Latent Dirichlet Allocation (LDA). However, due to its poor performance on modelling short texts, LDA is ill-equipped for social media data.

The goal of our report, we attempt to offer a solution to human evaluation by proposing a method to compare neural topic models against findings from content analysis done by social scientists, which we treat as the gold standard. We take as case study posts from Incel and manosphere forums and employ two experiments to validate the results of our neural topic model to the gold standard topics extracted by humans. Our research question is the following: what methodology can we use to evaluate topic models against the output of manual thematic analysis? To achieve this, we:

1. explore the use of neural topic models, comparing their performance to LDA for reference, using topic coherence and topic diversity as quantitative metrics to pick which is the most promising.
2. treat the themes found by (Vallerga & Zurbriggen, 2022) as gold standard in our evaluation, and present two methods to evaluate results of our topic model: topic keyword and gender association tests, as well as comparing documents representative of topics in human evaluation to those given by a neural topic model.

2. Related work

2.1. Topic modelling

2.1.1. LDA

Among digital humanities and social computing scholars, text mining to extract themes has often been limited to variations of Latent Dirichlet Allocation (Guzman)(Jelodar & Frank, 2021), as few have taken advantage of recent advances in deep learning. While LDA has been shown to be reliable (Hoyle et al., 2022), it struggles on short text

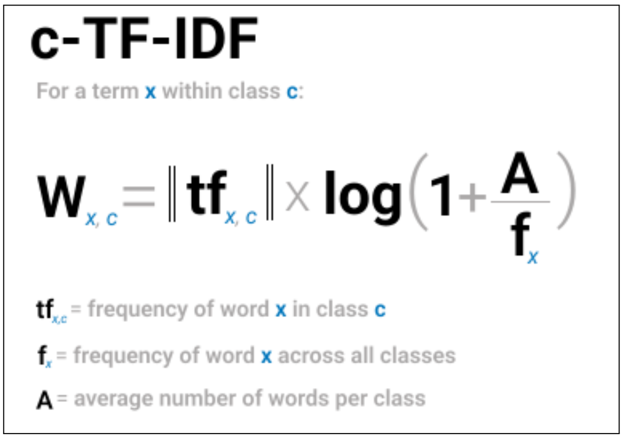

Figure 1. Calculation of c-TF-IDF

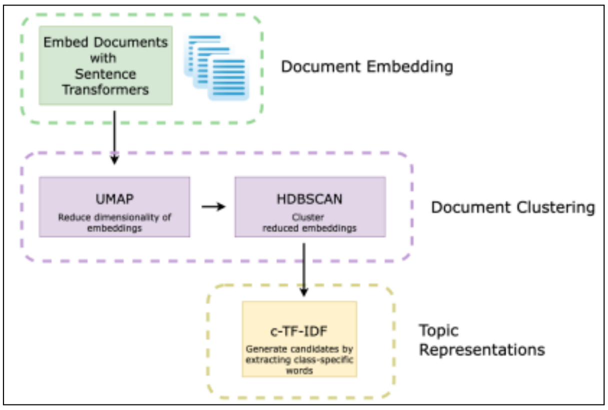

Figure 2. BERTopc architecture

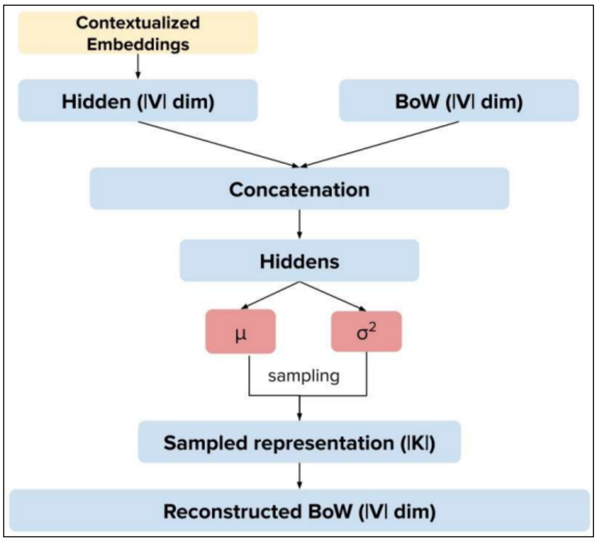

Figure 3. CTM architecture

(Hong & Davison, 2010) (text under 50 tokens), making it incompatible for many posts from forums and subreddits. Furthermore, LDA is unable to capture the semantic relationship between words in a corpus. Most neural topic models (NTMs) leverage variational autoencoder architectures, which are able to capture words and semantic similarity of sentences using embeddings. In our paper we explore three such neural topic models: BERTopic (Grootendorst, 2022), Top2Vec (Angelov, 2020) and Combined Topic Model (Bianchi et al., 2021), all of which make use of sentence embedding variational autoencoder models as a way to enrich the traditional bag-of-word co-occurrence measures.

2.1.2. Neural Topic Models

BERTopic (Grootendorst, 2022) combines a sequence of modular components (see Figure 2). In a nutshell, the model starts by creating an embedding for every document in the corpus using the sentence transformer SBERT (Reimers & Gurevych, 2019) (Document Embedding step). Then the algorithm performs dimensionality reduction with UMAP (McInnes et al., 2020) to avoid the curse of dimensionality (Verleysen & François, 2005) before clustering the resulting document embeddings using HDBSCAN (Malzer & Baum, 2020), an extension of DBSCAN that leverages hierarchical clustering (Document Clustering step). Finally, it generates a bag-of-words for each cluster by combining all documents in a cluster together and calculating each words' c-TF-IDF1 weight (Topic Representations step). For its final output, it keeps the top *n* words per topic where *n* is a user-defined hyperparameter.

Here *c* refers to the cluster (topic), *t* to the term, *A* is the average number of words per cluster, and f_t the frequency of *t* in a topic the inverse class frequency term (ie. How much information a word brings to a class). c-TF-IDF thus gives the importance score for words within a cluster allowing for multiple words to belong to different topics with different weights.

Top2Vec(Angelov, 2020) : is very similar to BERTopic except that it jointly embeds the document and word vectors in one embedding. BERTopic does not make that assumption: it separates the embedding and the topic creation stage. Dimensionality reduction and clustering is done on the embedding space containing both the document and word embeddings using procedures identical to BERTopic. Finally, it calculates the centroid of each cluster and returns the top *n*-closest words per topic where *n* is again a hyperparameter.

Contextualized Topic Model (CTM) (Bianchi et al., 2021) combines the neural topic model ProdLDA (Srivastava & Sutton, 2017) with SBERT (Reimers & Gurevych, 2019) embedded representations. ProdLDA uses an encoder-decoder model on a BoW document representation. The resulting representation is concatenated with the SBERT document representation projected through a hidden layer (see Figure 3).

2.2. Evaluation methods

The most popular evaluation methods are topic coherence, topic diversity, and human evaluation(Zhao et al., 2021). Experiments show that topic coherence is inline with human evaluation of topic interpretability (Lau et al., 2014) (see 1 for the equation). Topic diversity measures how diverse the discovered topics are. The assumption is that a good topic is one that doesn't share many of its most frequent content words with other topics. Finally, according to (Hoyle et al., 2022), human evaluation remains the best method to evaluate a topic model. The definition of what this human evaluation consists of, however, has received little attention. (Gillings & Hardie, 2022) argue that closely reading the top representative documents of a topic is a better technique than merely reviewing the top n most frequent tokens per topic. In our project we explore another route by comparing topic modelling results directly against human content analysis. Importantly, our method assumes that thematic analysis has already been done and due to the small sample-size and bias, a quantitative confirmation metric can be beneficial.

2.3. The Incel community

In recent years, misogynistic online communities have been brought into the spotlight. In 2014, Elliot Rodger killed six people, including two women, and wounded one outside a sorority house. Before the shooting, he posted a video detailing plan for his attack and complaining of being rejected by women and being envious of sexually active men. Rodgers frequented online communities of men who called themselves "incels", short for *involuntary celibates*. Incels believe they are unable to find romantic or sexual partners due to the circumstances that are outside of their control. While starting out as an online support group, these online communities have been characterized by resentment and hatred towards women, along with the belief that they are entitled to sex from them. They are predominantly white, heterosexual males.

2.4. Studies of the Incel community

The community has been the object of some studies recently. For our purpose, these can be split in two types: qualitative and quantitative research.

2.4.1. Qualitative research

Qualitative research tends to focus on content analysis. It involves familiarizing oneself with the corpus, creating codes, searching for themes, defining, and naming themes as coders discover and discuss the material amongst one another. Here coding refers to thematic encoding, related terms in computer science research would be tags or labels. The goal of these encoding and content analyses as a whole is to answer a research question about the studied community (Bengtsson, 2016). (Zdjelar, 2020) encodes posts with the goal of understanding how members view community using 200 threads combining 3400 posts. (O'Malley et al., 2020) randomly collect posts using web searches to identify their "norms, values, and beliefs".

Finally, (Vallerga & Zurbriggen, 2022) aimed to study the different forms of masculinities in the Incel community. They built their corpus with 227 posts from the incel forum *incels.is* and subreddits *r/Incel* and *r/TheRedPill*. Due to constraints on the number of posts they could read, they randomly selected the week June 4th 2018 to June 10th 2018 from the timespan between July 2017 to June 2018. To further reduce their corpus size, they randomly selected 27 posts from *r/TheRedPill* and collected 200 posts from posts in both *r/Incel* and the forum *incels.is*. Their research questions were:

RQ 1 How do member of the Incel community view men?

RQ 2 How do they view women?

The coding was done by three students with no prior knowledge of the community. The authors acknowledge that they are only reading an infinitesimal number share of posts. Though they argue posts are "repetitive" and thus their sample size is large enough, no evidence is presented to support these claims.

The authors identify 11 themes which they present in a hierarchical order in Figure 4. Themes are organized in three tiers: scientific justification, attitudes and beliefs, and actions, behaviours and solutions. In tier 1, the authors argue Incels view themselves as rational and scientific, advocating for

1. Gender Essentialism: Incels assume people of the same gender share an underlying essence
2. Evolutionary Psychology:the gender essentialist view is justified by innate biological differences as the cause for the differences in behavior.

In Tier 2, the authors present the different motivations and male archetypes underpinned by the gendered differences. They put forward six themes:

3. Motivations of women - deceptive: women are inherently manipulative to exploit men, dishonest, and make fabricated rape claims
4. Motivations of women—Promiscuous: women have an evolutionary need to be promiscuous in order to find the most evolutionary fit male mate
5. Motivations of women—Trading sex for power: women seek power from men through their sexuality which they trade for male held power such as money, fame and physical strength. This, posters found frustrating
6. Typologies of men—Alphas/Chads: they refer to sexually successfull men which have desireable traits both physically and in status. Posters differed on whether a man can become an Alpha.
7. Typologies of men—Betas: they are weak men who rely on trading money, power or emotional intimacy for sex. Incels expressed vitriol for betas for stooping so low as to get manipulated by women and for being fragile and lost.

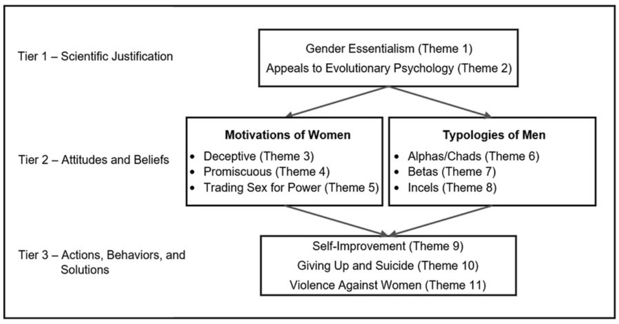

Figure 4. Themes presented by (Vallerga & Zurbriggen, 2022) in their thematic analysis

8. Typologies of men—Incels: posters saw themselves as being biologically, genetically, physically, mentally and immutably undesirable or worthy of sex. They see no hope for themselves and express anger and suicidal tendencies

Finally, in Tier 3, posters argue for different courses of action:

9. Self-improvement: involves becoming less emotionally expressive and more physically attractive, usually by becoming more muscular. This was done by sharing progress and assessing their attractiveness.
10. Giving up and suicide: posters believed their lack of romantic success, mental and physical characteristics was permanent. There is no way to make their lives better and they contemplated adopting a nihilistic point of view (blackpill) and suicide.
11. Violence against women: expressed their frustration through descriptions of extreme physical and sexual violence, mostly against women but also sexually successfull men.

Importantly, for each theme, the authors give us detailed descriptions of the themes along with representative documents. For our project, we use the themes found by (Vallerga & Zurbriggen, 2022) as the baseline against which to compare the quality of the topics found by our model, and present two novel probing methods to do so.

2.4.2. Quantative research

Quantitative analysis has focused on: characterizing the evolution of popularity levels, growth, and inter-forum migration of users (Horta Ribeiro et al., 2021), explore the use of misogynistic language (Farrell et al., 2019), along with applying topic modelling to discover themes (Jelodar & Frank, 2021). Importantly, approaches using topic modelling have been restricted to LDA without considering novel neural approaches. Furthermore, interpretation of results was done by eyeballing without a gold standard thematic evaluation found by social scientists. Our approach is novel in 2 ways: **First**, we explore the use of neural topic models, comparing their performance to LDA for refernce, **Secondly**, we treat the themes found by (Vallerga & Zurbriggen, 2022) as gold standard in our evaluation. **Finally**, we present two methods to compare results of our topic model, to those of from (Vallerga & Zurbriggen, 2022).

3. Dataset and task

3.1. Task

The task of topic modelling can be described as a grouping (clustering) task where a collection of documents is grouped based on the differences in distribution of tokens (words) inside these documents. What differentiates topic modelling from the classic classification task is the absence of pre-defined rules for each of the categories (groups). The model's output is then used to interpret the topics discussed in the set of documents, based on the top words associated to each of the clusters.

3.2. Dataset

In this paper we attempted to mimic the data used by the (Vallerga & Zurbriggen, 2022) whose data was not made available. They took posts from a combination of *incels.is*, *r/Incels* and *r/TheRedPill* between July 2017 and June 2018 and limited themselves to half a random week of posts for *r/RedPill* and a single day of posts for *r/Incels* and *incels.is*. We assume that the themes found by (Vallerga & Zurbriggen, 2022) are representative of the posts of that year and use as data a subset of data scraped by (Horta Ribeiro et al., 2021).

The dataset contains posts from 6 forums and 51 subreddits. We first narrow our dataset to only include posts from incels.is, *r/Incels*, and *r/TheRedPill*. Then, we process the data by removing posts with the keywords '**[deleted]**' or '**[removed]**' as they simply indicate the post has been removed and do not provide meaningful topics relevant to the incel community. Then, we remove websites, trailing whitespaces and newlines. Preprocessing is kept to a minimum as CTM, Top2Vec and BERTopic use transformer-based sentence embeddings to create document embeddings (see Section 4 for more details about the models). Removing stop words, lemmatization or any heavy pre-process would only create meaningless sentences whose embeddings lose sensitive to the entirety of the semantic content of the sentence. Due to computational constraints, we were unable to run a hyperparameter search and compare all models using all posts in our range. Thus, we create two datasets: $\mathcal{D}_{small}$ for comparing models and $\mathcal{D}_{large}$ for a hyperparameter search on our most promising model, and final comparison to (Vallerga & Zurbriggen, 2022).

For $\mathcal{D}_{large}$ we keep posts posted between the 1st of July 2017 and the 30th of June 2018. We exclude posts outside of that year even if the thread was created in that time span. The sentence embeddings quality degrades for posts with more than 256 tokens long so we exclude those. To filter out meaningless spam posts we remove posts with less than

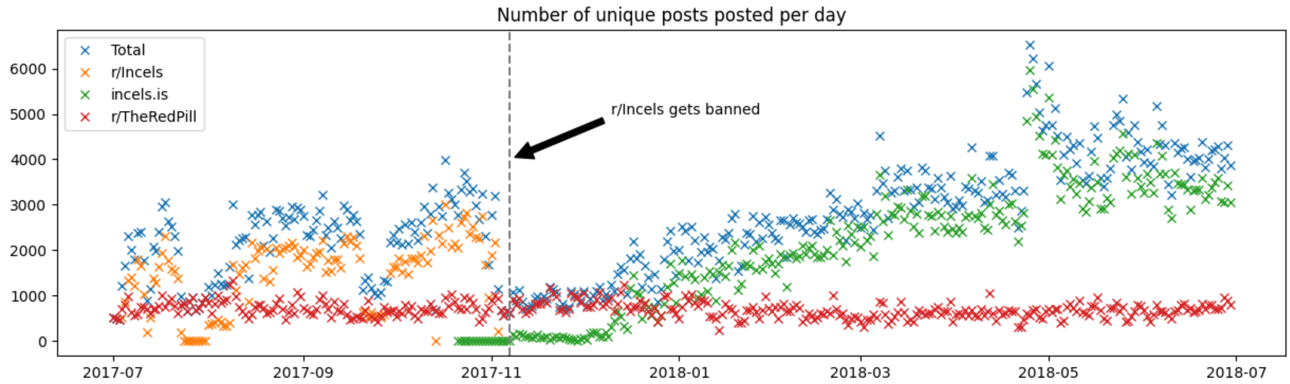

Figure 5. Number of posts per day

10 tokens. Finally, to keep posts that are representative of the community we filter out unpopular posts. While our dataset doesn't include the number of views or likes, we take number of answer as a proxy for popularity. The rationelle is that posts with high engagement represent posts on important topics. We filter out posts with less than 20 comments. Our final dataset has 1,212,626 unique posts. Note that there are no posts from incels.is prior to the 7th of November 2017 as it was created hours after the ban of *r/Incels*. Similarly, there are no posts from *r/Incels* after that date because it was banned.

For $\mathcal{D}_{small}$ and we simply limit our span to posts posted within the first two months, between 1st of July 2017 and the 31st of August 2017. In total, $|\mathcal{D}_{small}| = 164,909$ unique posts. Furthermore, $\mathcal{D}_{small} \subset \mathcal{D}_{large}$.

3.3. Evaluation Metrics

On both $\mathcal{D}_{small}$ and $\mathcal{D}_{large}$ we will measure our topic model's performance using *topic coherence* and *topic diversity*. Both are standard measures of topic model performance in the litterature (Zhao et al., 2021).

Topic coherence determines how well a topic is supported by the reference corpus on which a topic model was trained on. While many are used, a standard metric used in the litterature is Normalized Pointwise Mutual Information (NPMI) implemented:

$$NPMI(w', w*) = \left(\frac{log(\frac{P(w',w*)+\epsilon}{p(w')p(w*)})}{log(-P(w', w*) + \epsilon)}\right) \tag{1}$$

w' and $w*$ are both words among the top n words in a topic T, and $P(w', w*)$ the probability of token w' and token $w*$ occuring in the same document using ϵ smoothing. The NPMI score is calculated over all pairs of words in each topic and averaged to produce the final coherence measure. I is normalized between -1 and 1, where a positive NPMI value indicates a pairs of topic words co-occur frequently in the same document. The idea is that topics formed by tokens which occur frequently in the same document are topic coherent with the corpus they've been trained on.

Topic diversity is simply the percentage of unique words in the top n words for each topic.

To compare the topic found by our neural model to those found by (Vallerga & Zurbriggen, 2022), we employ two methods: thematic keywords-gendered term association test and document topic validation. As both are novel methods, we present those in the methodology section, Section 4.

4. Methodology

One the goals of this paper is to enrich the toolkit available for evaluation of topic models. According to the researchers in the field, human evaluation is still one of the top methods of evaluation (Guzman), used actively in current research, and our aspiration is to start the journey of bridging the gap between qualitative human evaluation and quantitative metrics for topic modelling output. In our paper, we present a methodology to compare the findings of a topic model to the themes found by human content analysis.

4.1. Initial Model Comparison

Due to computational constraints, we start by selecting which topic model is the most promising by comparing their coherence and topic diversity scores. We used 4 different topic modelling algorithms to choose from: LDA, BERTopic, CTM and top2vec. LDA was included due to its widespread use in the field of computational social science and the latter three models are Neural Topic Models, which are of main interest to us. Each of the models was trained on the same dataset $\mathcal{D}_{small}$. Furthermore, for each model, the only hyperparameter that we test is the Number of Topics. We do this because it's 1) the only one they have in common 2) given that we were unable to find their optimal hyperparameter values due to computational power issues. Ideally we would have found the best hyperparameter tuned to each individual model. However, given that we are only using this step to determine which is the most *promising*, we believe our initial comparison to provide a good insights. The results from these experiments are presented in table 1.

Since the coherence score is most similar to human topic evaluations(Lau et al., 2014)(Zhao et al., 2021), we decide to give more weight to topic coherence over topic diversity when evaluating each model's performance.

4.2. Best Model Hyperparameter Search

We perform a hyperparameter search for the best model of from Section4.1, BERTopic. This time, however, it is done a much larger, and thus different dataset, $\mathcal{D}_{large}$. According to (Grootendorst, 2022), the main hyperparameters that influence clustering quality are: mininum topic size and the final number of topics. As in step 1, we perform our hyperparameter search using both topic coherence and topic diversity. Results can be found in Table 3. Our best scoring model is used in our final experiments presented in Section 4.3.

4.3. Topic-to-Human Evaluation

In our final experiment, we introduce the results of two human-to-topic-model evaluation metrics. The driving idea is given that the best means to evaluate a topic model is by having humans confirm whether or not the resulting topics are good, we attempt to provide two quantitative measures to calculate the similarity of a topic model's output to themes found by humans. Human content analysis studies a corpus and extracts themes in order to answer a research question. Thus, our topic model will be judged by its ability to provide topics which agree with those found by human analysis when processed through the critical lens of a research question. (Vallerga & Zurbriggen, 2022) give us both detailed descriptions of each theme and examples of representative documents with the aim of explicating 1) How Incels view men, and 2) How Incels view women.

4.3.1. Topic Keyword and Gender Association

In this experiment we test whether our neural topics make the same gender-topic associations that Vallerga and Zurbriggen find. We first start by generating a list $X = x_1, ..., x_n$ and $Y = y_1, ..., y_n$ of masculine and feminine gendered terms respectively. Then, for each theme $T_i = t_1, ..., t_{11}$, where, for example, t_{10} is Theme 4: Women as deceivers, we generate a list of phrases and keywords that best described the theme given by Vallerga and Zubriggen.

Since BERTopic came out as our most promising model, we leverage its ability to capture the semantic meaning of a list of words with SBERT. We create an embedding for the list of keywords for a given theme (T_i) and then compute the cosine similarity between these keywords and all topics to find the top 10 closest clusters (topics). Each cluster is defined as a distribution over words $W_{i,j} = w_1, ..., w_{25}$ where words are weighted by their c-TF-IDF value. Here i refers to the i^{th} theme in (Vallerga & Zurbriggen, 2022) and j the j^{th} topic found by a topic model. We calculate the average weight of gendered terms in $W_{i,j}$ for each of the two gendered term lists along with the average weight of words characterizing $W_{i,j}$ for comparison. Because certain topics are more associated to our keyword embeddings, we also calculate the *weighted* average for each list of gendered terms. Here, the weight is equal to $1 + cosine\,sim(\phi(t_i), x_i)$ where ϕ is the SBERT embedding model. Finally, because certain topics have many highly weighted top words we also use the average ranking of each gendered term belonging to X or Y that appears in $W_{i,j}$.

For example, we can define:

$$\begin{aligned} X &= \{\text{man}, \text{men}, \text{male}\} \\ t_4 &= \{\texttt{promiscuity}, \texttt{deception}\} \end{aligned} \tag{2}$$

running `BERTopic.get_topics`(t_4) we get ten topics $W_{4,j}$ each defined by 25 words with weights $c_1 * w_1, ..., c_{25} * w_{25}$. Here 4 is in reference to theme T4. We find all mentions of the terms $\{\text{man}, \text{men}, \text{male}\}$ in each list of topics $W_{4,j}$. Then we can take the average over their weights c_i and the weighted average where the weight is equal to the cosine similarity between t_4 and $w_{i,j}$.

4.3.2. Representative-Predicted Document Comparison

In this experiment, we simply take every representative document $d \in D$ that (Vallerga & Zurbriggen, 2022) come up with as representative of each topic, find their embedding using the SBERT embedding layer, calculate the closest topic by cosine similarity, and evaluating whether the top returned topic agrees with Vallerga and Zurbriggen's topic.

5. Experiments

5.1. Model Selection

The first part of experimentation involved testing four models of interest to select the algorithm that will be trained on $\mathcal{D}_{large}$ data set. The results, presented in 1, capture the performance of each of the models and the effect of changing the Number of Topics parameter. BERTopic leads the way for all the values of Number of Topics parameter that we tested. Additionally, we can see that not only is BERTopic leading, but it is also far ahead of the rest of the models, as none of the other algorithms were able to escape the zone with negative coherence values.

We take this opportunity to remind the reader that coherence is the metric that correlates closely with human evaluation, so the importance of diversity scores was downplayed in the choice of the model for the deep hyperparameter search in the next step of our experimentation. Also, a peculiar, but, perhaps, unsurprising finding is that diversity seems to correlate negatively with coherence. Intuitively, this can be explained by the fact that if the topics are not coherent, the top words that "characterize" such topics will likely be more unique, as they will likely be more random.

5.2. BERTopic Hyperparameter Search

Having selected the topic modelling algorithm, we progress into the phase of deep hyperparameter search for our final BERTopic model. NTMs rarely have a large set of hyperparameters that can be used for experimentation, and BERTopic is not an exception to this rule. We trained the

	Number of Topics									
	10		20		30		40		50	
BERTopic	**-0.005**	40.0	**-0.015**	37.5	**0.015**	43.0	**0.021**	46.2	**0.026**	52.0
CTM	-0.096	86.0	-0.052	81.0	-0.031	70.3	-0.017	61.0	-0.022	56.7
top2vec	-0.160	96.0	-0.232	89.0	-0.206	91.0	-0.261	82.0	-0.251	84.0
LDA	-0.038	53.0	-0.047	47.0	-0.027	44.3	-0.024	45.2	-0.043	56.6

Table 1. Coherence (left) and diversity (right, expressed as percentage) scores for the four models tested in the phase one of model selection, along with impact of the Number of Topics hyperparameter. For each column, the model with the best coherence score is emboldened; the best coherence score overall can be found in the last column.

models by varying 2 hyperparameters: number of topics (like in part one) and the minimum topic size (which refers to the minimum number of documents that need to be classified into a cluster for it to be considered a topic). Both parameters seek to limit the number of topics the model can extract from the training set of the documents. The results presented in Table 2 shed light on the impact of the hyperparameters on performance of BERTopic – the range of coherence values is stretching from -0.0011 (MTS=100, NoT=400) all the way to 0.0472 (MTS=500, NoT=400). This wide span of evaluation metric values confirms appropriately chosen hyperparameter values for experimentation and strengthens the support for the model that was chosen for testing our novel topic modelling evaluation methods. Interestingly, the trend with negative correlation between the coherence and diversity scores continues here as well. The model parameters that yield the worst coherence score also provide the highest diversity score. And once again, our final choice was heavily skewed towards the model scoring the highest on coherence.

5.3. Human Evaluation

For the subsequent experiments, we pick the hyperparameters which performed the best according to topic coherence in our hyperparameter search in Section 5.2:

- min_topic_size = 500
- num_topics = 400

5.3.1. Keyword - Gender Association Test

The goal of this experiment was to determine whether BERTopic was able to find the same thematic gendered associations found by (Vallerga & Zurbriggen, 2022) in support of finding out what Incels thought of men and women.

For this experiment, we generated a list of keywords based off of each theme found in the human evaluation along with keywords relating to gendered terms. These can be found in the Appendix 6. Then, using the BERTopic model we selected in Section 5.2, we found the single embedding for each group of keyword and their 10 most similar topics described as a list of 25 tokens according to the procedure detailed in 4.3 and calculate their related metrics. Topics strongly related to specific gendered terms should have a high average score, weighted average, and rank compared to the rest of the words in the topic. Furthermore, this difference should also be observed for themes associated to one gender over another.

We would expect topics relating to keyword descriptions of T1 & T2 should give highly, equally weighted, gender terms. The keyword descriptions for T2, T3, T4, & T11 should return topics with highly weighted terms referring to women. Finally, the keywords for T6, T7, T8, T9 & T10 should return topics with highly weighted male terms.

All results are in Table 3.

We find that T1 is heavily associated to male keywords who average higher weights than other words. Female keywords, on the other hand, score lower. For keywords describing T2, no gendered keywords appeared in any of the top 25 words for each of the 10 topics. We also find that for T2, T3, T4, & T11, topics are not all heavily associated to women. In fact, only T3 is slightly associated to women and barely above the word average. The rest are heavily associated to men. Finally, T6, T7, T8, T9, & T10 is also mixed. T7 and T9 are more heavily associated with women, though they under, or equal to the average word weight in the topics. T8 is the theme whose keywords returned topics that were the most highly associated to men, and gender as whole.

It seems that most topics returned by the keywords either had an inconclusive to slight association to women when compared to men. However, those which were associated to men tended to exhibit very strong associations. We hypothesize that this is because Incels speak about either effects of women on men, or the male condition alone. Thus, whenever women gendered words are used, they are used in conjunction with male words. However, the opposite isn't necessarily true.

One strong limitation is that while topic models can capture the co-occurrence of words, they are unable to describe their relation. However, this relation is fundamentally what's the most important. For example, for T3, topics had similar association levels to each gender. If we were to run a POS-tagger, we might find out that men are often objects of verbs with women as subjects. In that case, while both co-occurred an equal amount and both appeared in sentences together, we would be able to validate that the topic was associated to "women doing to men". This would agree perfectly with the descriptions of themes found by (Vallerga & Zurbriggen, 2022).

		Number of Topics											
		50		75		100		150		200		400	
MTS	100	0.0156	69.4	0.0081	68.6	0.0058	72.8	0.0014	78.6	-0.0002	79.6	-0.0011	80.4
	150	0.0024	65.8	0.0103	69.3	0.0136	72.2	0.0169	73.9	0.0181	75.6	0.0202	76.9
	200	**0.0168**	62.8	**0.0199**	66.7	0.0236	68.7	0.0245	71.8	0.0265	73.0	0.0334	74.1
	300	0.0139	62.6	0.0197	63.7	0.0216	65.2	0.0292	70.2	0.0370	70.5	0.0443	71.2
	500	0.0124	57.8	0.0167	57.9	0.0226	61.2	0.0359	65.8	**0.0437**	66.6	**0.0472**	67.8
	700	0.0036	49.4	0.0150	56.4	**0.0250**	60.0	**0.0377**	61.5	0.0396	60.1	0.0396	61.8

Table 2. Coherence (left) and diversity (right, expressed as percentage) scores for BERTopic models with various values for Minimum Topic Size (MTS) and Number of Topics. For each column, the MTS value that yields the best coherence score is **emboldened**.

Themes	Average Words	Average Std Words	Average Topics	Std Topics	Average Score		Weighted Average		Average Rank		Std Rank		Total	
T1	0.009	0.012	0.330	0.047	0.014	0.006	0.019	0.009	7.3	9.38	4.58	7.29	**7**	**6**
T2	0.010	0.012	0.300	0.05	0	0	0	0	0	0	0	0	0	0
T3	0.01	0.013	0.317	0.039	0.006	0.007	0.007	0.009	12.5	10.38	6.02	**8.09**	3	3
T4	0.012	0.012	0.401	0.027	0.012	0.005	0.017	0.007	12	12.71	6.68	6.47	3	3
T5	0.010	0.012	0.389	0.028	0.014	0.006	0.019	0.009	8.55	11.17	5.77	7.61	**7**	**6**
T6	0.012	0.009	0.485	0.037	0.01	0.005	0.015	0.008	**13.33**	**12.88**	6.82	4.81	4	5
T7	**0.018**	**0.017**	0.512	0.087	0.004	0.006	0.008	0.010	9.5	12.1	5.5	6.41	1	4
T8	0.010	0.013	0.355	0.070	**0.019**	0.007	**0.025**	0.010	9.67	7.13	**7.74**	5.62	3	3
T9	0.008	0.009	0.408	0.035	0.005	**0.008**	0.007	**0.011**	10	9	1	6.27	1	4
T10	0.009	0.0106	0.420	**0.096**	0.011	0.006	0.017	0.009	3	14	0	6.68	1	2
T11	0.010	0.011	**0.536**	0.039	0.015	0.006	0.023	0.009	7	12.78	5.64	7.90	5	4

Table 3. Results from the Keyword Gender Association Test. In columns where there are two numbers, the leftmost number refers to calculations done with male keywords and rightmost women keyword. Average Words is the average weight of all words over all topics, Average Std Words the average standard deviation of each topic, Average Topics the average cosine similarity between the keywords and the topics, Average Score the average weight of gendered terms appearing in the top 25 terms per topic, Weighted Average the weighted average, Average Rank the average rank of appearing gendered terms in each topic, Std Rank the standard deviation, and Total the total number of gendered terms that appear

5.3.2. Document Topic Comparison

The aim of this experiment was to determine whether representative documents of themes found by human analysis were related to similar topics found by BERTopic. Our dataset contains all posts mentioned by (Vallerga & Zurbriggen, 2022). Thus, no extra computation was need to get the document embeddings and top topic.

We report the results of the experiment detailed in Section 4.3.2. Due to space constraints we are unable to include all documents that (Vallerga & Zurbriggen, 2022) found representative in this section. Thus, they are relegated to Appendix 6. However, we report that overall the model did not perform well. Themes 1, 3, 5, 6, and 11's representative document never aligned with the representative documents returns by BERTopic. Furthermore, in total 8/27 documents were clustered as belonging to BERTopic's topic -1 for outliers, or words, with no strong topic association. This indicates that not only had BERTopic not match with (Vallerga & Zurbriggen, 2022)'s theme, it was deemed as to not belonging to any specific topic at all.

For example: "Your male genes are programed to want women just for sex cause your male superiority doesn't need anything better from them." was associated to T2 in (Vallerga & Zurbriggen, 2022) but to cluster -1 in BERTopic. The authors argue the document exemplifies appeals to evolutionary psychology.

One could argue that given that (Vallerga & Zurbriggen, 2022) put T2 at the top of he hierarchy, maybe the theme is too abstract for a topic model to pick up on. Similarly to T2, T1 predicted both documents belonged to the cluster -1. This included the post: "Neurologically, the famale brain hasn't changed much since the caveman era. They still desire whatever what was the best genetic fit for them back in the day." which shares clear the same appeals to essentialist evolutionary psychology as the previous document.

This issue isn't limited to descriptions of themes high in the hierarchy, Take the document: "Talk to the iron. Lift. Can't say it enough. Read any of the posts in TRP and you'll see this again and again. FUCKING LIFT. Get yourself in shape and sorted out before you try and get girls", representative of theme T9 according to(Vallerga & Zurbriggen, 2022). BERTopic did associate to a cluster. However, its cluster presented keywords relating to the gym: {'lifting', 'lift', 'squat', 'bench', 'strength','weight', 'deadlift', 'training'}. The topic keywords are certainly related to the document, however, its relation is much more literal than the human interpretation which sees it as a case of motivation, instead of a simple cluster of words relating to lifting weights.

However, some themes aligned very well. Themes T3 and T4's representative documents aligned well with those of BERTopic. For the post "There is no bigger liar than the

Theme	Number of matched documents
T1 - Gender Essentialism	0/2
T2 - Evolutionary psychology	0/2
T3 - Woman as deceptive	2/2
T4 Promiscuous	2/3
T5 Trading sex for power	0/3
T6 Alpha/Chad	0/2
T7 Betas	1/2
T8 Incels	1/2
T9 Self-improvement	1/3
T10 Giving up/suicide	1/4
T11 Violence Against Women	0/3

Table 4. Document Comparison Scores for each topic

female human. All the rising false rape charges essentially prove this. All the false hope they give toward sub8 men also proves this fact.", its top topic contained words like {'lying', 'allegations', 'consent', 'false'}. This topic aligns well with T3, which relates to Incel's beliefs that women make false sexual-assault allegations.

6. Conclusions

In conclusion, we can first validate that some neural topics models are able to perform better than LDA on social media data and achieve good scores on standard metrics. Furthermore, using sentence embeddings not only helps models perform well, but also offers flexibility and a wider range of analysis methods.

Our goal in this project was to present a method to evaluate the results of neural topic models against a human analysis. In the Keyword-Gender Association Test from Section 5.3.1, we were unable to consistently replicate the gendered association found by (Vallerga & Zurbriggen, 2022) in their analysis. We identified the lack of information about the relationship between gendered terms and other terms as a potential path to explore. Finally, in the Document-Topic Comparison, BERTopic was unable to find the same topics for representative texts of themes in (Vallerga & Zurbriggen, 2022).

Due to computational and time constrainsts, we were unable to properly evaluate all topic models nor perform the comparisons done in Section 5.3. We hope this can be done in the future.

Note, we originally wanted to use Neural Topic Models to study the incel community to perform sociological research. However, our project slowly shifted to a methods paper. We acknowledge that studying the Incel community is disturbing. If our aim from the start was to perform a methods paper we would not have chosen such a toxic community as case study.

References

Angelov, Dimo. Top2vec: Distributed representations of topics, 2020.

Bengtsson, Mariette. How to plan and perform a qualitative study using content analysis. *NursingPlus Open*, 2:8–14, 2016. ISSN 2352-9008. doi: https://doi.org/10.1016/j.npls.2016.01.001. URL https://www.sciencedirect.com/science/article/pii/S2352900816000029.

Bianchi, Federico, Terragni, Silvia, and Hovy, Dirk. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 759–766, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.96. URL https://aclanthology.org/2021.acl-short.96.

Farrell, Tracie, Fernandez, Miriam, Novotny, Jakub, and Alani, Harith. Exploring misogyny across the manosphere in reddit. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 2019. doi: https://doi.org/10.1145/3292522.3326045.

Gillings, Mathew and Hardie, Andrew. The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities*, 12 2022. ISSN 2055-7671. doi: 10.1093/llc/fqac075. URL https://doi.org/10.1093/llc/fqac075. fqac075.

Grootendorst, Maarten. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

Guzman, Allyssa. Libguides: Digital humanities tools and resources: Home. URL https://guides.lib.utexas.edu/c.php?g=815192&p=5817590.

Hong, Liangjie and Davison, Brian D. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pp. 80–88, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450302173. doi: 10.1145/1964858.1964870. URL https://doi.org/10.1145/1964858.1964870.

Horta Ribeiro, Manoel, Blackburn, Jeremy, Bradlyn, Barry, De Cristofaro, Emiliano, Stringhini, Gianluca, Long, Summer, Greenberg, Stephanie, and Zannettou, Savvas. The evolution of the manosphere across the web. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):196–207, May 2021. doi: 10.1609/icwsm.v15i1.18053. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18053.

Hoyle, Alexander, Goel, Pranav, Sarkar, Rupak, and Resnik, Philip. Are neural topic models broken?, 2022.

Jelodar, Hamed and Frank, Richard. Semantic knowledge discovery and discussion mining of incel online community: Topic modeling. *CoRR*, abs/2104.09586, 2021. URL https://arxiv.org/abs/2104.09586.

Lau, Jey Han, Newman, David, and Baldwin, Timothy. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1056. URL https://aclanthology.org/E14-1056.

Malzer, Claudia and Baum, Marcus. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, sep 2020. doi: 10.1109/mfi49285.2020.9235263. URL https://doi.org/10.1109%2Fmfi49285.2020.9235263.

McInnes, Leland, Healy, John, and Melville, James. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

O'Malley, Roberta Liggett, Holt, Karen, and Holt, Thomas J. An exploration of the involuntary celibate (incel) subculture online. *Journal of Interpersonal Violence*, 37(7-8):1–28, Sep 2020. doi: https://doi.org/10.1177/0886260520959625.

Reimers, Nils and Gurevych, Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Srivastava, Akash and Sutton, Charles. Autoencoding variational inference for topic models, 2017.

Vallerga, Michael and Zurbriggen, Eileen L. Hegemonic masculinities in the "manosphere": A thematic analysis of beliefs about men and women on the red pill and incel. *Analyses of Social Issues and Public Policy*, 22(2), Apr 2022. doi: https://doi.org/10.1111/asap.12308.

Verleysen, Michel and François, Damien. The curse of dimensionality in data mining and time series prediction. In *Proceedings of the 8th International Conference on Artificial Neural Networks: Computational Intelligence and Bioinspired Systems*, IWANN'05, pp. 758–770, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3540262083. doi: 10.1007/11494669_93. URL https://doi.org/10.1007/11494669_93.

Zdjelar, V. *Alone Together: Exploring Community on an Incel Forum*. Theses (School of Criminology). Simon Fraser University, 2020. URL https://books.google.co.uk/books?id=3w15zgEACAAJ.

Zhao, He, Phung, Dinh, Huynh, Viet, Jin, Yuan, Du, Lan, and Buntine, Wray. Topic modelling meets deep neural networks: A survey, 2021.

Theme	Keywords
masculine	'masculinity', 'masculine', 'man', 'men', 'males', 'guy', 'guys', 'dudes', 'boy', 'boys', 'he', 'him'
feminine	'femininity', 'feminine', 'woman', 'women', 'females', 'female', 'she', 'girls', 'girl', 'foids'
incels	'incels', 'truecels', 'us', 'ourselves', 'ourself', 'myself', 'me', 'I'
T11	'violence', 'violence against', 'to kill all', 'to hurt all', 'kill them', 'hurt them', 'revenge', 'rape', 'sexual assault'
T10	'giving up', 'to commit suicide', 'getting blackpilled', 'blackpilled','commit suicide', 'depression', 'depressed', 'rot in bed', 'rotting in bed', 'rope myself', 'I want to die.', 'kill myself', 'death', 'to die'
T9	'I self-improvement', 'I become more physically attractive', 'to better myself', 'body modification myself', 'invest in positive outlets', 'getter better', 'better version of myself'
T8	'genetic disadvantage', 'incel', 'Incel', 'incels', 'undesireable', 'unattractive', 'ugly', 'discrimination', 'living as an incel'
T7	'Beta', 'beta', 'betas', 'Betas', 'weak', 'emotional', 'trade financial security for sex', 'financial security in exchange for a relationship that includes sexual intimacy'
T6	'sexually successful', 'alpha', 'chad', 'Alphas', 'Chads', 'Alpha', 'Chad', 'successful', 'socially desireable traits'
T5	'inherently seek power through their sexuality', 'only power is their sexuality', 'trading sex for power', 'seek power in echange for sex', 'exchange power for sex'
T4	'promiscuous', 'primary motivation is promiscuity', 'need to find the most (evolutionarily) fit mate', 'are biologically designed to want sex', 'promiscuityas an opportunity for sexual domination'
T3	'inherently deceptive', 'deceitfulness is a means of manipulating', 'will falsely indicate sexual interest', 'because they are deceiving, no accusation of rape is believable'
T2	'biological essentialism', 'immutable biological differences as the cause of differences in behavior', 'gendered behaviors as genetic, biological, or having evolved', 'evolutionary psychology explains gendered behaviors', 'are naturally predisposed'
T1	'Gender essentialism', 'people of the same gender share a deep, underlying essence', 'completely distinct attributes', 'scientific', 'neurologically'

Table 5. Table outlining the Themes and the corresponding extracted keywords

Theme	Post
T1 Gender Essentialism	"Women appear to be master manipulators because their biology allows them to hold multiple contradictory beliefs at the same time - that she loves Billy and wants to get fucked by Chad. In fact not only does their biology allow for this, but it is actually the lynchpin of female sexual strategy. The fact that these beliefs are sincerely held is what makes them so effective. Humans are evolved to have a sixth sense for when we are being lied to by someone we know well. Female cognitive dissonance evolved to defeat the human brain's lie detecting ability."
	"Neurologically, the famale brain hasn't changed much since the caveman era. They still desire whatever what was the best genetic fit for them back in the day."
T2 - Evolutionary psychology	"Your male genes are programed to want women just for sex cause your male superiority doesn't need anything better from them."
	"We are, indeed, no longer in hunter-gatherer times. However, much of what was at play then still applies today. This includes women of course desiring bigger and more physically intimidating men, among many other things. It's all evolutionary behavior bro."
T3 - Woman as deceptive	"Yeah man all women are going to lie about that. It's in their nature to deceive sexually, whether it's lying about their past exploits and N counts [number of sexual partners], or tricking people into pregnancy or cheating and using sex in a relationship to manipulate behaviors."
	"There is no bigger liar than the female human. All the rising false rape charges essentially prove this. All the false hope they give toward sub8 men also proves this fact."
T4 Promiscuous	"Women can't handle freedom to a point where if you leave them to their own devices they will fuck around like crazy. And science has proved it many times over but people just ignore it."
	"That's the current state of females fellas. They are nothing but whores and parasites seeking to get fucked by as many Chads as possible."
	"yup.. genetics are the only thing that matter. femoids can give typical blue pilled response but once my aunt saw the hunter eyes and superior frame of the chad she immediately got wet like a dog in heat…. femoids can't fight biological ticks"
T5 Trading sex for power	"Our gynocentric society is full of propaganda that says men should chase women, that women are the prize, that the man who buys flowers and stays by her side (sometimes chasing for years) … eventually wins. That is a load of horse shit. The man of value, instead, brings wisdom, strength, mental fortitude, leadership, wealth, and excitement to the table. Women (girls) crave this. It is built into their evolutionary psychology and biology. It is so hard-wired into them … not even all the movies, TV shows, media propaganda, and fiction books can overcome this instinct. Women may say one thing (that they should be attracted to nice guy soy boys who buy them dinner and drinks), but we at the RedPill know not to listen to what women say, but watch what they do."
	"Would be funny to see a billionaire incel here, though if you even have over a million there is no way you can be incel."
	"And the science (not exactly science just common sense again) is: Chad can do whatever he wants, and he will be perceived as all-good, subhuman can do whatever he wants too, he will be perceived as all-bad."
T6 Alpha/Chad	"Because they are the submissive inferior and thus you can do what you want with them and they take it. They're the doormats, you're the superior … Keep your guard up and don't ever make a bitch feel cute or like anything ever again unless she EARNS it. No meaningful attention until she earns it by being a dope ass girl whose always there when u want her etc. She gets a treat when she does right by you. Not for being a breathing vagina owner."

	"But also do it with the knowledge that she's not yours, it's just your turn. Would you have felt as bad post-breakup if your mindset throughout had been that the relationship would end sooner rather than later, and to just enjoy the time you do have together while you can?"
T7 Betas	"Sorry kid, you are Beta as fuck. Its laughable how big your ego is. You are the definition of a girlfriend. Chicks talk shit about guys they manipulate and don't bang you … You realised you would never be able to get this girl, so you sabotaged her yourself. You're pathetic."
	"Are you so desperate that you are now at the point in which you are 100% willing to beta provide? Also are you wealthy or earning enough to do this? in case you finally have a chance to beta bux a hot women. If you beta bux and marry the bitch then you deserve inceldom for life and eternal punishment for being a cuck. If you beta bux for that pussy for a little while and then dump the bitch on the streets then that is ultimate revenge for your life thus far."
T8 Incels	"I was born. I was socially inept. I was a trouble child. I played lots of video games and spent a lot of time online. I struggled with my peers. My family distanced themselves from me. My mental health worsened. I now play lots of video games and spend too much time online watching anime and posting on forums/image boards. I am ugly."
	"Living as an incel is basically dying. Not being able to put your penis inside a non escort will literally fuck your life up."
T9 Self-improvement	"Invest your personal time into positive outlets. Lift, have a social network, have hobbies, make money outside work. Do the things you need to for ranking up in SMV."
	"In a nutshell mate, you've got a LOT to learn about Life and TRP's a great place for you to begin, (So, first read ALL the sidebar material AND the sites like The Rational Male) and it will take some time (possibly years) for you to 'internalize' TRP's teachings/advice."
	"Talk to the iron. Lift. Can't say it enough. Read any of the posts in TRP and you'll see this again and again. FUCKING LIFT. Get yourself in shape and sorted out before you try and get girls."
T10 Giving up/suicide	"The bright side of the blackpill is knowing it's over anyway so I can relax and enjoy watching the Brazil x Austria soccer game on tv with my family."
	"Before I'd be like "fuck games I need to approach 50 women today or I'm FUCKED" - "YOU ONLY APPROACHED 4, FUCKER!!! PIECE OF SHIT!!!"
	"i feel so burned out with life at just 16, im so tired already, what the fuck, i know this isnt normal i dont know what to do, social isolation is mentally jarring and i can feel myself deteriorating, i have no friends, hardly any family (even fewer of which that care), no money and no life, i literally spent all day just rotting in bed and thinking. I'm gonna need some next level cope for this one boys, help me out, im being genuine here, i dont have the balls to rope just yet, though death sounds rather enticing right now."
	"i need to fucking die already. i live for nothing, and base pleasures do nothing for me. my fear of death is nothing compared to the fear of waking up in the morning."
	"The best way to solve depression: suicide."
	"The best cope is rope. Jk"
T11 femoids and stop words	"If all of us grab a weap/knifu we can make the day of the incelindependenceday". Another advocated violence for the sake of "revenge": "I wouldn't rest until every incel gets his revenge. Go ER or fuck with the femoids for the sake of good ol' times."
	"thats what would ultimately happen. Im not encouraging it. Im just saying based on how more and more men are becoming incel due to sexual distribution getting more unequal. The only way some men would be able to obtain sex would be via rape."
	"If I somehow manage to get laid, she'd be the only one on drugs there."