From Words to Actions: A Comprehensive Approach to Identifying Incel Behavior on Reddit

Ahmet Y. Demirbas
Williamsville East High School
Buffalo, New York, USA
ahmetdemirbas525@gmail.com

Jakir Hossain
University at Buffalo
Buffalo, New York, USA
mh267@buffalo.edu

Ahmet Erdem Sarıyüce
University at Buffalo
Buffalo, New York, USA
erdem@buffalo.edu

***Abstract*—The incel (involuntary celibate) community is a radicalized online subculture. Understanding its dynamics is crucial for mitigating youth radicalization and preventing online polarization. In this study, we examine Reddit communities to identify users at risk of deep engagement in incel-related subreddits. We analyze activity patterns and comments of 14,000 users and employ a two-step approach to carefully prepare a custom set of features to identify at-risk users. We first consider 7,000 incel-engaged users and use them to select 7,000 control users who are similar to the incel-engaged users in terms of non-incel activities. Recent subreddit activity patterns of those users are used to create features. We then use word2vec on the comment texts to create text-based features. We find that utilizing only the subreddit activity patterns of users achieves an accuracy of 79% while using word2vec modeling alone yields a classification accuracy of 76%. Remarkably, the two approaches have complementary strengths and integrating both approaches achieves a near-perfect classification accuracy of 99.8%. By employing a two-pronged approach, our results achieve a significant increase over previous work. By illuminating social media's role in online radicalization processes, we hope that the insights from our work can guide policymakers and platform moderators in creating safer online spaces.**

***Index Terms*—social media analysis, reddit, incel detection, toxic masculinity**

I. Introduction

Certain online communities have evolved into echo chambers, amplifying toxic ideologies and potentially leading to radicalization and social isolation [1]–[3]. One example is the "incel" (involuntary celibate) community, which has become notorious for promoting misogynistic and, at times, violent ideologies. This online subculture has experienced substantial growth in 2010s, attracting vulnerable individuals and fostering a narrative of resentment, entitlement, and perceived societal injustice [4], [5].

The incel community mainly consists of young males, typically ranging from late teens to early thirties [6]. While incels can come from various backgrounds, studies have shown a significant overrepresentation of white and Asian men within the community. Demographically, incels often report being from middle-class or lower-middle-class backgrounds, with many expressing frustration about their socioeconomic status. These individuals frequently report struggles with mental health issues, including depression, anxiety, and low self-esteem [7]. It is important to note that incel ideology is not uniform; it encompasses a spectrum of beliefs, from those who passively identify with the label to more extreme followers who advocate for violent actions.

The incel community's impact has extended beyond the digital realm, with several violent incidents linked to incel ideology. Notable examples include the 2014 Isla Vista killings [8] and the 2018 Toronto van attack [9]. Additionally, numerous self-harm incidents have been associated with incel extremism [10]. These real-world consequences highlight the need to understand the dynamics of these online communities and their potential for radicalization.

In this work, we focus on Reddit [11] to analyze, characterize, and identify incel behavior. Reddit is a popular social media platform where various subreddit communities cater to different interests and ideologies. We study user engagement patterns and comments within incel-related subreddits to understand how isolation occurs and how individuals become deeply entrenched in these communities. Based on our findings, we aim to develop a robust classifier that uses custom features to accurately identify individuals who are funneled into incel-related subreddits. Our study is guided by the following three key research questions:

- RQ1: To what extent can users' subreddit activity patterns, e.g., comment activities, classify their engagement in incel communities?
- RQ2: How can users' comment texts help classify their engagement in incel communities?
- RQ3: What key features distinguish users who become deeply engaged in incel subreddits from those who do not?

Our study employs a multi-faceted approach to identify incel behavior on Reddit, combining data collection, feature engineering, and machine learning. We analyze user activity across subreddits by collecting 500 recent comments per user via the Reddit API, amassing nearly 7 million comments from over 138,000 subreddits. We then apply word2vec modeling [12] to these comments, creating a separate 100-dimensional word embeddings for incel-engaged 7,000 users and the control group of 7,000. Consequently our feature vectors incorporate both subreddit activity levels and word embeddings, capturing both behavioral patterns and linguistic nuances (see Figure 1). We develop our classification model using multiple algorithms (Logistic Regression, SVM, Decision Trees, Random Forests, and kNN) implemented in scikit-learn.

Our analysis reveals significant insights into the effectiveness of different feature engineering approaches and machine learning models for classifying user engagement in incel communities. Our key findings are as follows:

- Using word2vec modeling alone in feature vectors yields a maximum classification accuracy of 76% with a 0.767 F1 score.
- Utilizing only the subreddit activity patterns of users in feature vectors achieves a maximum accuracy of 79% with a 0.788 F1 score.
- Remarkably, by integrating both word2vec embeddings and subreddit activity patterns into our feature vectors, we achieve a near-perfect classification accuracy of 99.8% with a 0.998 F1 score.

Our results highlight the power of combining linguistic and behavioral features for incel identifaction. These two methods have complementary strengths: Subreddit analysis provides a broader contextual picture of the user activity, whereas using word embeddings offer a deeper semantic analysis of the actual comment content. Intuitively, our two-pronged approach significantly reduces false positives and negatives by cross-referencing data and provides a robust binary classification model for detecting incel behavior on Reddit. Our results increase the accuracy bar significantly when compared against a recent work [13] that achieved a 78.8% accuracy on incel identification by using sentiment analysis.

We also find that the Random Forest model consistently outperforms other machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (kNN). It is worth noting that the performance gap between Random Forest and other models narrows significantly when we employ the combined feature approach, which suggests that our feature set is robust and independent of the classification algorithm.

Significance. The significance of our research extends beyond the specific context of incel communities. By understanding the mechanisms of online radicalization and isolation, we can gain valuable insights into similar processes occurring in other extremist or harmful online groups. Our findings have the potential to:

- Inform the development of early intervention strategies to support vulnerable individuals before they become deeply entrenched in harmful ideologies,
- Guide policymakers in crafting evidence-based regulations to address online radicalization while preserving free speech, and
- Assist social media platforms in improving their moderation strategies and community guidelines to create safer online spaces.

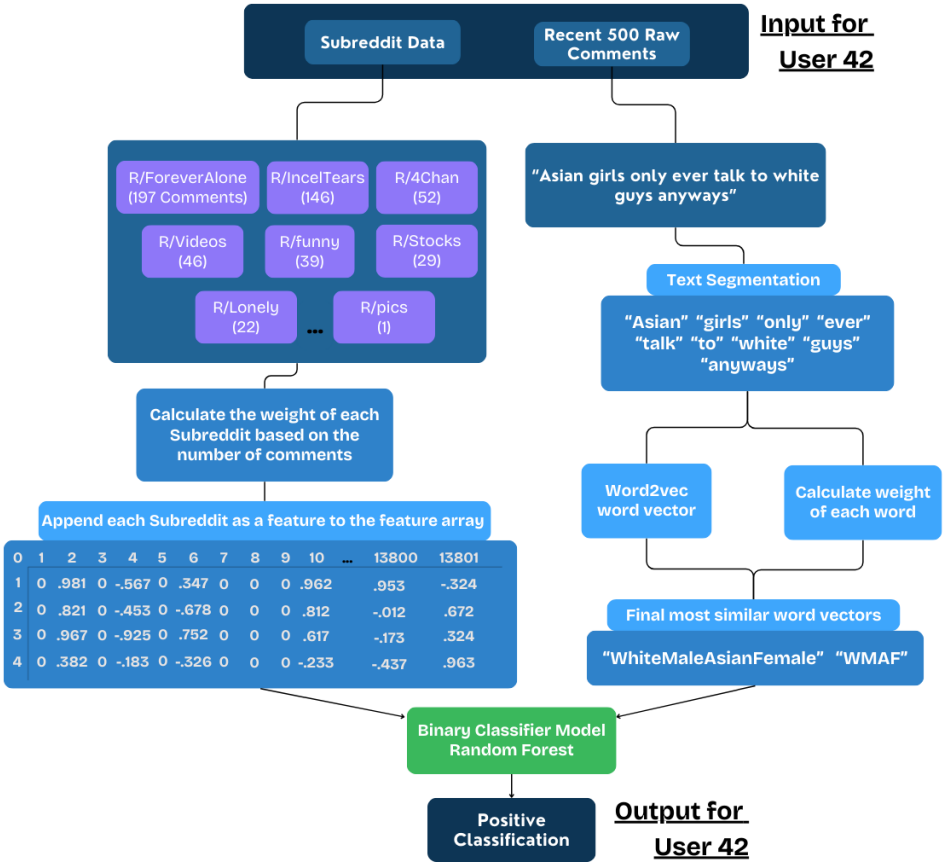

Fig. 1: Flow diagram of the inference phase of the hybrid approach model.

II. Related Work

Here we first survey notable works on online radicalization and intervention strategies, and put our work in perspective. Then we discuss specific studies that characterize and identify incel users on social media, and compare/contrast them with our work.

A. Online radicalization

Broader psychological research on the social dynamics leading to extreme belief adoption highlights a complex interplay between ideology, mental health, and online platforms in the radicalization process [14], [15]. In particular, research on online radicalization processes has revealed concerning trends in the relationship between internet usage and extremist attitudes. Exposure to radical online content is found to be associated with an increased risk of adopting extremist views and potentially engaging in political violence [16]. This risk appears to be particularly pronounced among young, educated men who are active on social media platforms [17]. Furthermore, the structure of online interactions tends to reinforce these tendencies, with users forming echo chambers that amplify their existing beliefs and limit exposure to alternative viewpoints [18].

NLP and social media analysis have emerged as powerful tools for understanding and addressing online radicalization. For instance, Mossie and Wang have developed sophisticated hate speech detection models that can identify vulnerable communities and track the spread of extremist ideologies across different platforms [19]. They identify hate speech by extracting features from posts using word n-grams and Word2Vec embeddings, and deploying deep learning models like RNNs. In another interesting work, Aljarah et al. employed BoW and TF-IDF techniques for cyber hate speech detection in Arabic Twitter data and conducted feature importance analysis to identify the most significant features [20]. Recent approaches often combine advanced NLP techniques, such as word embeddings, with large-scale data analysis of social media content. Studies have successfully applied these methods to track the evolution of user behavior in extremist communities, revealing patterns of increasing radicalization over time [21]. Such insights are crucial for developing more effective intervention strategies and for understanding the dynamics of online radicalization processes.

Intervention strategies for online radicalization have taken various forms, from peer support systems to community moderation and risk profiling. Peer-delivered interventions in online communities have shown promise in providing crisis support, particularly among vulnerable groups like veterans [22]. These strategies help mitigate feelings of isolation and provide individuals with positive social interactions that can counteract harmful narratives. However, the complexity of online risks, especially for youth, necessitates targeted prevention strategies that account for the unique manifestations of offline risks in digital spaces [23]. While community-level moderation can effectively reduce harmful activity, there is a risk that it may lead to increased toxicity in smaller, more isolated groups [24]. These findings highlight the need for nuanced, multi-faceted approaches to intervention that can adapt to the dynamic nature of online radicalization processes. We believe that our work can guide targeted intervention strategies to mitigate deep engagement in incel communities.

B. Incel discourse and identification

The incel community, in particular, has been the subject of increasing research attention due to concerns about online radicalization. Moskalenko et al. surveyed active incels and documented common mental health problems and psychological trauma of bullying or persecution among incels [7]. O'Malley et al. performed qualitative analysis of online posts and argued that the incel community is structured around five key normative orders: the sexual market, women as naturally evil, legitimizing masculinity, male oppression, and violence [6]. Papadamou et al. examined the incel presence on YouTube, revealing a concerning growth in incel-related content and the platform's potential role in steering users towards extremist material [4] . Golbeck created an archive of incel forum posts and developed techniques to track radicalization through community-specific language [25].

A qualitative analysis of incel discourse on online forums reveals the prevalence of toxic masculinity narratives and the potential for these spaces to foster harmful ideologies [26], [27]. Riberio et. al. performed a large-scale quantitative analysis of manosphere's evolution, a diverse collection of websites, blogs, and online forums promoting masculinity, misogyny, and opposition to feminism, between 2005-2020 by examining data from 51 subreddits and six online forums [28]. Through an in-depth investigation of user participation across various manosphere communities and the content of their posts, the authors discovered that newer communities like incels have gained prominence, gradually overshadowing older groups such as `Pickup Artists` and `Men's Rights Activists`. Their findings further revealed that these emerging communities exhibit heightened levels of toxicity and misogynistic rhetoric when compared to the earlier communities within the manosphere.

Recent work examined the role of online platforms in dissemination of incel ideologies using sentiment analysis, and highlighted the need for effective content moderation strategies [29]. Notably, Hajarian et al. explored identifying incel users through analysis of their comments [13]. The authors collected data from Facebook and Twitter, preprocessed and analyzed a total of 520,513 comments. Their method combines sentiment analysis and profanity checking to classify users. Using this approach, the authors achieved an accuracy of 78.8% in identifying incel users across the two platforms. While their method shows promise, it is important to note that our approach of using a combination of subreddit activity patterns and word embeddings from comments achieves a significantly higher accuracy (99.8%). Although the underlying datasets are not the same, this improvement underscores the effectiveness of our multi-faceted approach of integrating both linguistic and behavioral features in classification. In addition,

while Hajarian et al. used manual review as the ground truth of incel identification, we use a more objective ground truth by identifying the top 10,000 active users in the 9 original incel subreddits (some of which are already banned) as the positive set.

III. Methodology

In this section, we discuss the details of our data collection, feature engineering, and model development strategies.

We employ a binary classification model, with positive labels indicating users engaged in incel subreddits and negative labels for those who are not. We constructed (1) feature vectors for Reddit users based on their activities across various non-incel related subreddits, and (2) text features by using Word2Vec model on all users' comments. We validate the classifiers' generalization performance using a held-out test set. For transparency and reproducibility, the Python scripts utilized in each step of data collection and feature vector generation is available at `https://github.com/ahmetdemirbas-git/RedditIncelAnalysis`. It is important to note that throughout this study, we maintained strict privacy protocols to protect user data. No individual user's comments or account information was shared or made public at any point during or after the research. All data was anonymized and aggregated for analysis, ensuring that no personally identifiable information was exposed. This approach allowed us to conduct our research ethically while respecting the privacy of Reddit users.

A. Data collection

To define the positive user set, we first select 9 subreddits that have been identified for their incel content in the study by Ribeiro et al. [28]. These subreddits are `r/Braincels`, `Incels`, `MGTOW`, `MensRights`, `antifeminists`, `RedPill`, `HAPAS`, `ForeverAlone`, and `ProMaleCollective`. (Six of these subreddits were banned by Reddit at the end of 2017 due to their harmful content.) Then we identify the top 10,000 active users in these 9 incel subreddits by mining post and comment histories in the academic torrents subreddit dataset [30] (content spanning from May 2005 to December 2023), We exclude the banned users from the positive user set to ensure that we only consider active users whose recent subreddit activities are available for feature creation. We end up with 7,000 active users. Note that removal of the banned accounts is crucial in our analysis as we need to access to the user accounts to learn about the other (non-incel) subreddits that the users have engaged in.

To construct the negative user set, a naive way would be to select 7,000 random users from the entire Reddit population. However, this would create an easily distinguishable control group, limiting the classifier's ability to learn nuanced differences between incel and non-incel users. Instead, we aim to construct a more challenging dataset by creating a control group that has similar activity patterns to the positive user set on neutral subreddit activities. To this end, we first determine the top 50 subreddits frequented by the 7,000 positive users, then remove the 3 that overlapped with the 9 initial incel-related subreddits: `MensRights`, `HAPAS`, `ForeverAlone`. (Note that the other 6 incel subreddits were banned and do not appear in current data.) This results in 47 control mainstream subreddits that do not seem to be part of the incel community. Table I ranks (top to bottom in 3 columns) these 47 subreddits based on the total activity of positive-labeled users in each (Profanity in subreddit names is censored with asterisks). Lastly, we normalize the activity across these 47 control subreddits and determine the number of users (n) to be selected from each to get to a total of 7,000 negative users. Note that the resulting 7,000 negative users have not engaged in the 9 original incel subreddits but are similar to the 7,000 positive user set in terms of non-incel subreddit activity.

We observe that the top 47 subreddits for the negative user set largely overlap with those of the positive userset. This is expected due to our control group construction method to make it a comparable set. New additions include `AITAH`, `NoStupidQuestions`, `nba`, `mildlyinfuriating`, and `interestingasf*ck`, replacing `kotakuinaction2`, `p*ssypassdenied`, `FeMRADebates`, `FA30Plus`, and `videos`. The ranking order also differs. Interestingly, the negative user set shows greater diversity in their subreddit participation. Figure 2 shows a comparison of top subreddits in both sets. `PurplePillDebate`, is the most popular subreddit in both groups. Its subreddit description reads as: "PurplePillDebate is a neutral community to discuss sex, relationship and gender issues, specifically those pertaining to /r/TheBluePill and /r/TheRedPill." `TheRedPill` describes itself as "Discussion of sexual strategy in a culture increasingly lacking a positive identity for men.", whereas `TheBluePill` states that "/r/TheBluePill is a satire of /r/TheRedPill and the strategies discussed on that particular sub. /r/TheRedPill is a subreddit for pick up artists who discuss ways of manipulating women."

To provide more insight into the dataset, Figure 3 presents the account creation dates for the incel and non-incel accounts. The incel users tend to have older accounts due to our selection method, which utilized the academic torrents dataset spanning from 2005 to 2023. As the 6 of 9 initial incel subreddits have been banned at the end of 2017 by Reddit, we observe a significant decline in account creation dates for incels after 2018. The number of non-incel accounts shows an upward trend over time, reflecting the growing popularity of Reddit usage. Note that the incel accounts also exhibit a similar growth pattern until the 2017 crackdown. We do not use the account creation date as part of the feature sets and provide this graph here only to give more information about the dataset.

B. Feature engineering and model development

We create feature vectors for each user in the positive and negative user sets to quantify their engagement in non-incel subreddits. Using the Reddit API, we collect 500 most recent comments per user. This process yields a total of 6,951,500

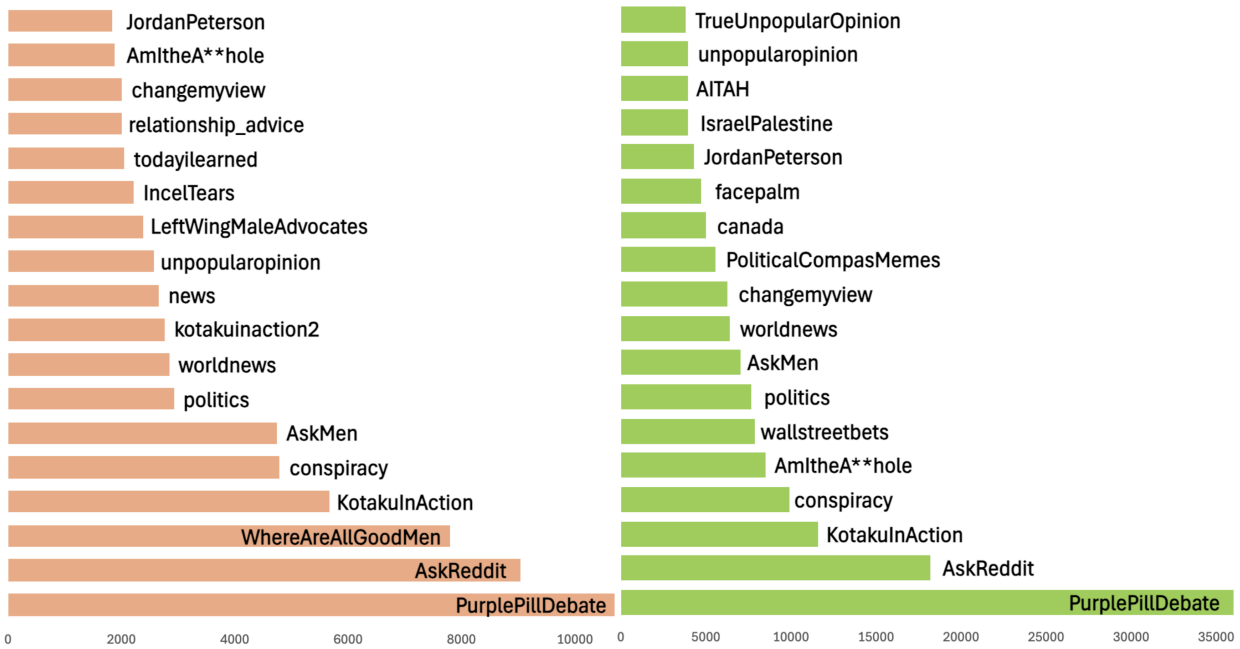

Fig. 2: Top subreddits for the positive (incel, left) and negative (non-incel, right) users. The x-axis is the total comment counts.

comments and 138,101 unique subreddits across all 14,000 users. The resulting feature set for each user includes 500 raw comments and comment counts per each subreddit, providing a comprehensive view of the user's online behavior.

To enhance the feature set, we incorporated word2vec models trained on the user comments [31]. All the comments are preprocessed by splitting a comment into a list of words with symbols removed. We use the Gensim library to train 100-dimensional word embeddings. Two separate word2vec models were trained: one on the comments of incel-engaged users and another on the control group. This allowed capturing semantic differences in language use between the two groups. The feature set was expanded to include these word embeddings of the comments of each user, in addition to the previously used subreddit activity data.

For model development, we employed five different machine learning algorithms: Logistic Regression [32], Support Vector Machines (SVM) [33], Decision Trees [34], Random Forests [35], and k-Nearest Neighbors (kNN) [36]. We implemented these models using scikit-learn, a popular Python machine learning library. Each model is tuned using grid search cross-validation to find the optimal hyperparameters. For Logistic Regression and SVM, we experimented with different regularization strengths. For Decision Trees and Random Forests, we tuned the maximum depth and number of trees. For kNN, we optimized the number of neighbors.

PurplePillDebate	conspiracy	politics
IncelTears	AmItheA**hole	aznidentity
funny	PublicFreakout	FA30plus
facepalm	gaming	Conservative
AskReddit	AskMen	worldnews
todayilearned	JordanPeterson	FeMRADebates
RedPillWomen	pics	p*ssypassdenied
videos	IsraelPalestine	collapse
WhereAreAllGoodMen	kotakuinaction2	news
relationship-advice	PoliticalCompassMemes	unitedkingdom
antiwork	AsianMasculinity	virgin
science	relationships	TrueUnpopOpinion
KotakuInAction	unpopularopinion	canada
changemyview	LeftWingMaleAdvocates	MMA
wallstreetbets	Libertarian	movies
technology	Anarcho-Capitalism	

TABLE I: 47 mainstream subreddits that are not part of the incel community but frequented by the incel users.

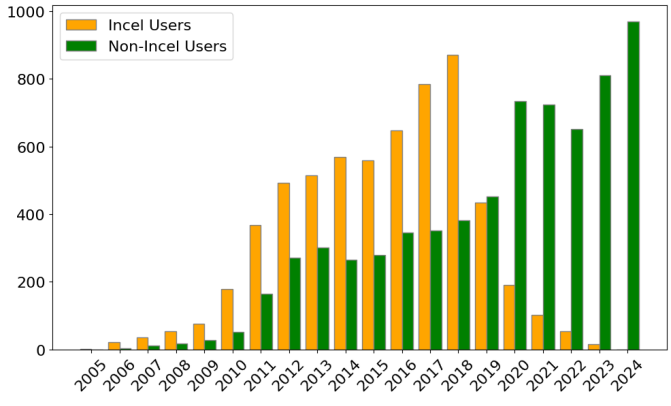

Fig. 3: Distribution of incel and non-incel users' account creation dates.

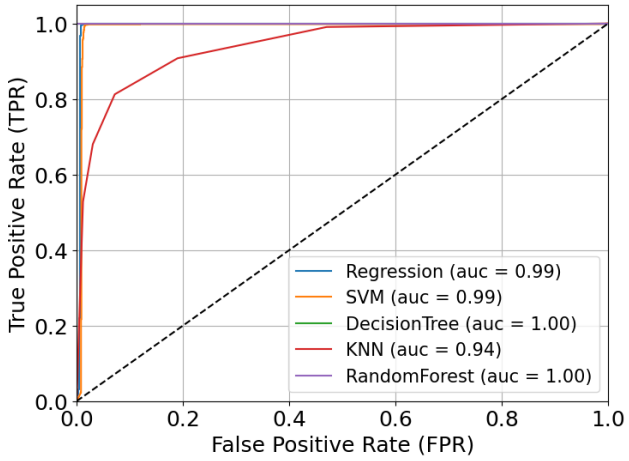

Fig. 4: ROC curve for the classifiers that use the entire set of custom features.

IV. Results

In this section, we evaluate the classification performance of the custom-features described above by using the five aforementioned classifiers. We use several metrics to evaluate model performance, including accuracy, precision, recall, and F1 score. We also generate Receiver Operating Characteristic (ROC) curves to visualize the trade-off between true positive rate and false positive rate for different classification thresholds. Each classifier is run ten times to determine its average precision and F1 score. The dataset was split 80/20 for training and testing.

We first discuss the performance of the entire set of features, subreddit activity and comment texts. Then we consider ablation studies to examine the contribution and limitation of only subreddit activity features and only comment text features. Finally, we conduct a feature importance analysis using the Random Forest model to identify which subreddits and linguistic features are the most influential in classifying users as incel or non-incel.

A. The overall performance

Our main model's performance presents the results of combining both subreddit activity data and word embedding features. By integrating these two types of information, we capture both the behavioral and linguistic aspects of user activity, resulting in a more comprehensive and accurate classification.

Table II shows the results. Using Random Forest model, we achieve an average of 99.8% accuracy. Other classifiers also perform well, resulting between 97.5% and 99.3% accuracy. This suggests that our custom features are reliable and use of any model yields significant performance.

To provide a more detailed understanding of the performance, we create the receiver operating characteristic (ROC) curve in Figure 4, which visually depicts the trade-off between true positive rate and false positive rate of the models. As the figure illustrates, all the models performed extremely well with the exception of kNN which gives an AUC of 0.94.

To corroborate the high accuracy of our results and check for overfitting, we also trained and tested the models on random subsets of the training set. Overfitting happens when the model fits the data too well - to the extent that it memorizes the training data and wouldn't generalize well to larger data. By randomly sampling our 7000 positive and negative users, we investigate overfitting. For ten trials, we randomly choose 10% of positive samples and 10% of negative samples from our initial data. We then split the new data into an 80/20 test split and ran it on all five models. Table III shows the average results of the ten random subsets. The average accuracy and F1 results are really high, and indicate that the models do not overfit the data.

While transformer-based models like BERT and GPT [37]–[39] represent the state-of-the-art in many NLP tasks, for the incel-ideology detection task our model achieves 99.8% accuracy using word2vec embeddings combined with subreddit activity features. The highly specific nature of incel terminology and discourse patterns enables word2vec to adequately capture the semantic relationships needed for this targeted classification task. Using more complex language models instead of the lightweight word2vec model would add unnecessary computational overhead without meaningful performance gains for this task. Our results demonstrate that for specialized text classification tasks with distinct vocabulary patterns, traditional word embedding techniques can still be optimal when thoughtfully combined with domain-specific features.

To illustrate word2vec's capability of capturing semantic relationships in this specialized task, Figure 5 shows the most common word associations with "women" from both the incel and non-incel lexicons as a Venn diagram. Interestingly, incel users make use of aggressive words such as "Rape", "Mutilation", "Violent", and "DomesticViolence" along with "women". Non-incel users often use "women" as a gender within the LGBT nomenclature, along with words such as "Gay" and "Trans". These distinct word associations reveal the stark contrast in how these groups view and discuss women.

ML model	Accuracy	Precision	Recall	F1 score
Regression	0.9887± .007	0.99± .01	0.99± .01	0.9887± .008
SVM	0.9932± .011	0.99± .005	0.99± .004	0.9932± .011
Decision Trees	0.9887± .009	0.99± .008	0.99± .007	0.9887± .008
Random Forest	0.9977± .006	0.99± .004	0.99± .003	0.9977± .007
kNN	0.9751± .012	0.98± .016	0.97± .012	0.9751± .013

TABLE II: Classification results using both subreddit features and word2vec.

ML model	Accuracy	Precision	Recall	F1 score
Regression	0.9732± .003	0.98± .005	0.98± .005	0.9732± .005
SVM	0.9612± .005	0.97± .01	0.95± .008	0.9612± .005
Decision Trees	0.9787± .009	0.97± .008	0.97± .007	0.9787± .008
Random Forest	0.9863± .005	0.98± .007	0.98± .002	0.9863± .005
kNN	0.9593± .015	0.96± .01	0.95± .013	0.9593± .015

TABLE III: Average results of the random subsets from ten trials.

ML model	Accuracy	Precision	Recall	F1 score
Regression	0.691 ± .021	0.69 ± .025	0.691 ± .022	0.69 ± .024
SVM	0.660 ± .031	0.680 ± .037	0.629 ± .03	0.661 ± .037
Decision Trees	0.728 ± .015	0.730 ± .012	0.728 ± .015	0.727 ± .013
Random Forest	0.789 ± .01	0.791 ± .008	0.789 ± .014	0.788 ± .01
kNN	0.664 ± .018	0.665 ± .019	0.664 ± .017	0.664 ± .019

TABLE IV: Classification results using only subreddit features.

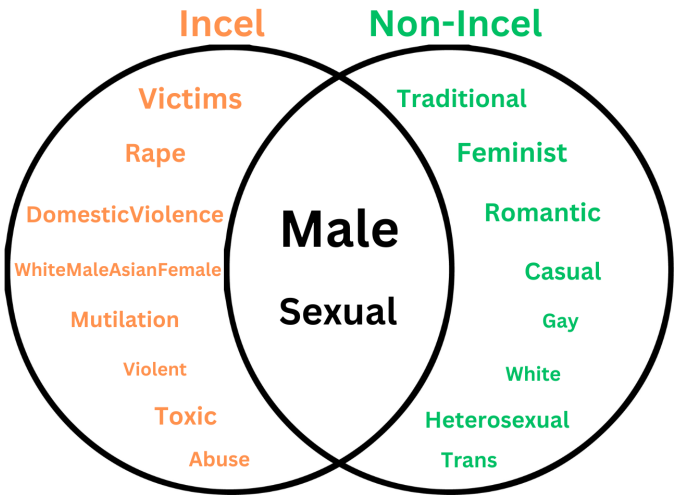

Fig. 5: Top 10 most common word associations with "women" from the incel (left) and non-incel (right) lexicon. The font size of each word relates to the strength of the word's correlation.

B. Ablation study

To isolate the impact of the features, we perform an ablation study. Below, we first evaluate model performance using only subreddit activity data in Section IV-B1, and then using only word embeddings in Section IV-B2.

1) Subreddit activity analysis: Table IV presents the results. Using only subreddit activity data, the Random Forest model achieves the highest performance with an average accuracy of 78.9% and an average F1 score of 0.788. Random Forest outperforms all the other models in every aspect by a significant margin. We believe that its ability to handle non-linear relationships and reduced sensitivity to outliers allows Random Forest to better capture the nuanced behavior of users in the incel and control group.

The subreddit activity only approach identifies users who visit communities frequented by positive set users. In this case, the false positives occur due to the users in the positive subreddits who do not necessarily engage in incel behavior. Since we choose our control group to have similar activity patterns with the positive user set, it is likely that more false positives appear when only subreddit activity is considered.

To further investigate the impact of subreddit activity, we analyze the feature importance scores for subreddits in the Random Forest classifier. We computed feature importance scores based on the decrease in model loss associated with each feature across the feature array. Feature importance for a given feature f is calculated as:

$$\text{Importance}(f) = \frac{1}{T}\sum_{t=1}^{T}\Delta\mathcal{L}_t(f)$$

where $\Delta\mathcal{L}_t(f)$ denotes the reduction in loss due to feature f in tree t, and T is the total number of trees in the forest. This score captures the cumulative influence of each feature on model predictions, providing insights into the model's reliance on specific subreddit activity and linguistic features to classify users as incel or non-incel.

Table V displays the scores which quantify each feature's contribution to the model's predictions, with higher scores

PurplePillDebate:	0.0199	RedPillWomen:	0.0045	LeftWingMaleAdv.:	0.0033
AITAH:	0.0127	mildlyinfuriating:	0.0044	videos:	0.0032
WhereAreAllGoodMen:	0.0126	unpopularopinion:	0.0043	ask:	0.0032
AskReddit:	0.0120	gifs:	0.0042	IncelTears:	0.0031
AmItheA**hole:	0.0112	changemyview:	0.0041	mildlyinteresting:	0.0031
facepalm:	0.0099	trashy:	0.0037	collapse:	0.0031
p*ssypassdenied:	0.0097	IsraelPalestine:	0.0037	Damnthatsinteresting:	0.0031
KotakuInAction:	0.0093	relationships:	0.0037	MadeMeSmile:	0.0031
wallstreetbets:	0.0081	RandomThoughts:	0.0036	FA30plus:	0.0030
worldnews:	0.0073	news:	0.0036	unitedkingdom:	0.0030
politics:	0.0067	PublicFreakout:	0.0035	TikTokCringe:	0.0030
PoliticalCompassMemes:	0.0061	AskMen:	0.0034	pics:	0.0030
NoStupidQuestions:	0.0060	IAmA:	0.0034	AdviceAnimals:	0.0029
conspiracy:	0.0058	BlackPillScience:	0.0034	interestingasf*ck:	0.0029
relationship_advice:	0.0055	dating_advice:	0.0034	MensRightsMeta:	0.0028
science:	0.0050	MMA:	0.0033	JordanPeterson:	0.0028
funny:	0.0048	TwoHotTakes:	0.0033		

TABLE V: Feature importance scores when only subreddit activities are considered.

ML model	Accuracy	Precision	Recall	F1 score
Regression	0.738 ± .016	0.737 ± .016	0.733 ± .014	0.736 ± .015
SVM	0.626 ± .026	0.628 ± .027	0.622 ± .021	0.624 ± .027
Decision Trees	0.695 ± .023	0.691 ± .02	0.694 ± .025	0.692 ± .022
Random Forest	0.767 ± .012	0.769 ± .018	0.760 ± .01	0.766 ± .011
kNN	0.655 ± .021	0.653 ± .02	0.656 ± .022	0.654 ± .02

TABLE VI: Classification results using only word2vec features.

indicating greater influence. Note that a high score does not necessarily imply importance for incel-positive classification. For instance, `AITAH`'s high score correlates more strongly with non-incel classification. Interestingly, despite Dr. Jordan Peterson's presumed association with incel users, `r/JordanPeterson` ranked low in importance for incel classification.

2) Word embedding analysis: Table VI shows the classification results when only text features are used. The word embedding-only approach yields an average accuracy of 76.7% (Random Forest) with a 0.767 F1 score. Text-based features excel at capturing general language patterns and identifying some incel-like language based on word associations. However, it faces limitations such as semantic ambiguity and lack of high-level context.

Upon further investigation, we find many specific cases where the word2vec based model incorrectly classifies users. Many of the false positive users are from incel-related neutral subreddits like r/PurplePillDebate. Those users would use specific incel terminology in the debates but they are not incels themselves. The model falsely flags these non-incels due to the raw text resembling incel language. As for the false negatives, the lack of context again flags many users with incel terminologies as non-incels because they did not necessarily use vulgar incel language. These incels are mainly on more neutral subreddits like r/askmen where they display incel ideologies without using the incel-like language. Figure 6 shows the top subreddits associated with false negatives and false positives for the word2vec model, supporting our insights into the limitations of using word embeddings alone for classification.

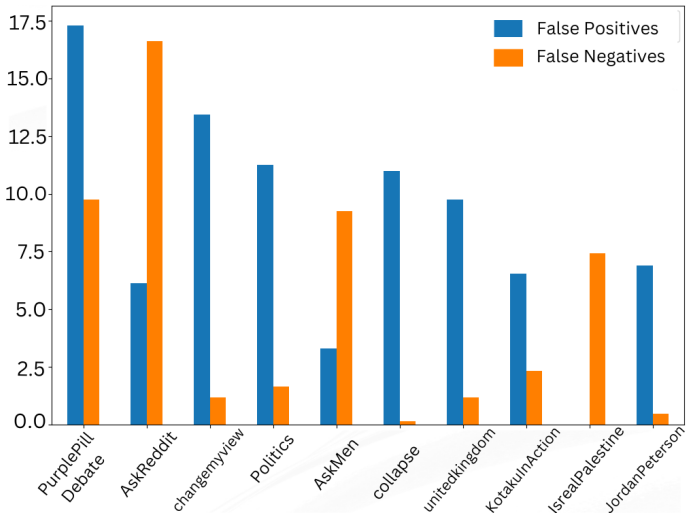

Fig. 6: Top 10 subreddits associated with the false negatives/positives reported by the classifier that uses only word2vec-based features.

V. Conclusion and discussion

The findings affirmatively answer the research questions, demonstrating that subreddit activity patterns in combination with the word2vec models effectively classify incel community engagement and reveal key features of deeply engaged users. The integration of word2vec features and raw comment data significantly enhanced the model's performance and achieved near perfect accuracy of 99.8% with a 0.998 F1 score, using the Random Forest classifier. This significant improvement can be attributed to the complementary strengths of both methods:

- Subreddit analysis provides a broader picture of user activity and where content is posted.
- Word embeddings offer a deeper semantic analysis of the actual comment content.

This two-pronged approach significantly reduces false positives and negatives by cross-referencing data. The dramatic improvement in accuracy demonstrates the power of combining linguistic and behavioral features in identifying users at risk of deep engagement with incel communities. This approach not only significantly improves classification accuracy but also highlights the complex interplay between users' language use and their patterns of community engagement in the context of online radicalization.

The strong performance of our model provides a reliable foundation for further research into incel community dynamics and for potential interventions. As online communities continue to shape social dynamics and influence individual beliefs, it is crucial to develop a nuanced understanding of how these digital spaces can foster both positive connections and harmful ideologies. Our research aims to contribute to this understanding, ultimately working towards creating a safer and more inclusive online environment for all users. The findings illuminate incel community dynamics and can provide a foundation for interventions against incel isolation to create safer online spaces.

In particular, this research can be leveraged in several ways to address the challenges posed by incel radicalization. The classification model developed in this study could be used by social media platforms and online moderators to proactively identify users at risk of deeper engagement with incel ideologies. By flagging these individuals, targeted interventions and support services could be provided to steer them away from harmful communities and ideologies [22], [23]. Additionally, policymakers and researchers could use these insights to inform the development of evidence-based policies and educational programs aimed at addressing the root causes of incel isolation and radicalization [24].

Future research directions could include longitudinal studies to track the evolution of users' comments and subreddit engagement over time. This approach would provide valuable insights into how individuals become increasingly influenced by incel communities and how their language and ideology shift as a result. By following users' online footprints, researchers could gain a deeper understanding of the radicalization process and identify potential intervention points along the way. Such longitudinal analyses could also shed light on the long-term impacts of incel engagement on mental health, social functioning, and real-world behavior.

References

[1] K. O'Hara and D. Stevens, "Echo chambers and online radicalism: Assessing the internet's complicity in violent extremism," *Policy & Internet*, vol. 7, no. 4, pp. 401–422, 2015.

[2] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini, "Modeling echo chambers and polarization dynamics in social networks," *Physical Review Letters*, vol. 124, no. 4, p. 048301, 2020.

[3] M. Wolfowicz, D. Weisburd, and B. Hasisi, "Examining the interactive effects of the filter bubble and the echo chamber on radicalization," *Journal of Experimental Criminology*, vol. 19, no. 1, pp. 119–141, 2023.

[4] K. Papadamou, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, and M. Sirivianos, "" how over is it?" understanding the incel community on youtube," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–25, 2021.

[5] A. M. Glace, T. L. Dover, and J. G. Zatkin, "Taking the black pill: An empirical analysis of the "incel"." *Psychology of Men & Masculinities*, vol. 22, no. 2, p. 288, 2021.

[6] R. L. O'Malley, K. Holt, and T. J. Holt, "An exploration of the involuntary celibate (incel) subculture online," *Journal of interpersonal violence*, vol. 37, no. 7-8, pp. NP4981–NP5008, 2022.

[7] S. Moskalenko, J. F.-G. González, N. Kates, and J. Morton, "Incel ideology, radicalization and mental health: A survey study," *The Journal of Intelligence, Conflict, and Warfare*, vol. 4, no. 3, pp. 1–29, 2022.

[8] "2014 isla vista killings," 2024. [Online]. Available: https://en.wikipedia.org/wiki/2014_Isla_Vista_killings

[9] "2018 toronto van attack," 2024. [Online]. Available: https://en.wikipedia.org/wiki/2018_Toronto_van_attack

[10] "Incel: Connection to suicide forums," 2024. [Online]. Available: https://en.wikipedia.org/wiki/Incel#Connection_to_suicide_forums

[11] "Reddit website/forum," http://www.reddit.com, accessed: 2024-08-08.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[13] M. Hajarian and Z. Khanbabaloo, "Toward stopping incel rebellion: Detecting incels in social media using sentiment analysis," in *2021 7th International Conference on Web Research (ICWR)*, 2021, pp. 169–174.

[14] G. M. Walton, G. L. Cohen, D. Cwir, and S. J. Spencer, "Mere belonging: the power of social connections." *Journal of personality and social psychology*, vol. 102, no. 3, p. 513, 2012.

[15] C. Efferson, R. Lalive, and E. Fehr, "The coevolution of cultural groups and ingroup favoritism," *Science*, vol. 321, no. 5897, pp. 1844–1849, 2008.

[16] G. Hassan, S. Brouillette-Alarie, S. Alava, D. Frau-Meigs, L. Lavoie, A. Fetiu, W. Varela, E. Borokhovski, V. Venkatesh, C. Rousseau *et al.*, "Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence," *International journal of developmental science*, vol. 12, no. 1-2, pp. 71–88, 2018.

[17] G. F. Hollewell and N. Longpré, "Radicalization in the social media era: understanding the relationship between self-radicalization and the internet," *International journal of offender therapy and comparative criminology*, vol. 66, no. 8, pp. 896–913, 2022.

[18] E. Brugnoli, M. Cinelli, W. Quattrociocchi, and A. Scala, "Recursive patterns in online echo chambers," *Scientific Reports*, vol. 9, no. 1, p. 20118, 2019.

[19] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media," *Information Processing & Management*, vol. 57, no. 3, p. 102087, 2020.

[20] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *Journal of information science*, vol. 47, no. 4, pp. 483–501, 2021.

[21] A. Kiss, "Incels: Frustrated and angry due to deprivation of intimacy: A case study of the radicalisation trajectories of an online community on a fringe social media platform," 2022.

[22] K. Perepezko, M. Bergendahl, C. Kunz, A. Labrique, M. Carras, and M. Colder Carras, "Instead, you're going to a friend: Evaluation of a community-developed, peer-delivered online crisis prevention intervention," *Psychiatric services*, pp. appi–ps, 2024.

[23] A. Alsoubai, A. Razi, Z. Agha, S. Ali, G. Stringhini, M. De Choudhury, and P. J. Wisniewski, "Profiling the offline and online risk experiences of youth to develop targeted interventions for online safety," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–37, 2024.

[24] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, and R. West, "Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–24, 2021.

[25] J. Golbeck, "A dataset for the study of online radicalization through incel forum archives," *Journal of Quantitative Description: Digital Media*, vol. 4, 2024.

[26] A. Lindsay, "Swallowing the black pill: A qualitative exploration of incel antifeminism within digital society," Ph.D. dissertation, Open Access Te Herenga Waka-Victoria University of Wellington, 2020.

[27] H. Habib, P. Srinivasan, and R. Nithyanand, "Making a radical misogynist: How online social engagement with the manosphere influences traits of radicalization," *Proceedings of the ACM on human-computer interaction*, vol. 6, no. CSCW2, pp. 1–28, 2022.

[28] M. H. Ribeiro, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, S. Long, S. Greenberg, and S. Zannettou, "The evolution of the manosphere across the web," in *Proceedings of the international AAAI conference on web and social media*, vol. 15, 2021, pp. 196–207.

[29] F. F. Fanucchi, "Examining the role of echo-chambers within online incel communities using sentiment analysis and group based trajectory modeling," Ph.D. dissertation, San Jose State University, 2023.

[30] Academic torrents, academic torrents subreddit dataset https://academictorrents.com/details/56aa49f9653ba545f48df2e33679f014d2829c10/tech&filelist=1.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[32] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

[33] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[34] L. Rokach and O. Maimon, "Decision trees," *Data mining and knowledge discovery handbook*, pp. 165–192, 2005.

[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[36] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," in *Proceedings of international conference on neural networks (ICNN'96)*, vol. 3. IEEE, 1996, pp. 1480–1483.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, 2019, p. 2.

[39] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, vol. 1, 2020.