Social Media Analytics
January 2023

r/seduction: community and content understanding of the subreddit

Bui Xuan Julia Lan[1], Pasinato Alessio[1], Salvaterra Michele[1]

Abstract

This project aims to analyze the subreddit r/seduction, which is a forum for discussing and sharing tips on how to attract and pick up romantic partners. The subreddit has faced controversy in the past for promoting toxic and manipulative behavior, as well as objectifying and devaluing women. The main objective of the project is to gain a better understanding of the current community and detect whether it is negative or not, to extract the interactions between users, detect communities, interests and the most discussed topics. Data of posts and comments were collected for the month of October 2022 and analyzed using sentiment analysis, network modeling and topic modeling. The results show that the subreddit has a positive sentiment overall, and the most discussed topics are related to appearance, approaches to women, relationships, and sexual relations. The project suggests that further research could be done by analyzing more data collected over a larger period of time, in order to gain a more reliable and relevant understanding of the r/seduction community as it is today.

Keywords

Graph theory — Social Media Analytics — Text Mining — Sentiment Analysis — Reddit

[1] *Master's Degree in Data Science, University of Milan - Bicocca*

Contents

1. Introduction

Reddit is a social media platform that allows users to submit content, such as text posts or links, and vote submissions up or down to determine their position on the site's pages. Users can also discuss and comment on submissions. Reddit has a wide range of communities, called "subreddits," which focus on specific topics.

Reddit users, also known as "redditors," can create their own subreddits or join and participate in existing ones. Moderators, who are elected by the subreddit community, are responsible for managing the subreddit and enforcing its rules.

One subreddit of interest is r/seduction [1], which is a forum for discussing and sharing tips on how to attract and pick up romantic partners. It has a specific set of guidelines and rules, and participation is restricted to those who are over the age of 18. The subreddit has faced controversy in the past for promoting toxic and manipulative behaviour. However, it remains a popular destination for those interested in the topic of seduction.

The main objective of this project consists in analyzing the subreddit, in order to detect whether it is negative or not, to extract the interactions between users, detect communities, their interests and the most discussed topics.

2. Research questions

The idea that give rise to the project consists in a better understanding of this seduction community, reported as toxic and negative, promoting a manipulative behaviour, as it also emerges from some posts [1] [2]. They criticize the r/seduction community, emphasizing that it is based on users that are no longer giving advice on how to attract romantic partners based on a polite and educated behaviour, but they provide a guide based on misogyny and toxic masculinity. The tactics and strategies promoted are often criticized for objectifying and devaluing women, and for promoting a view of women as mere objects to be conquered. Additionally, the subreddit has been accused of promoting false and harmful stereotypes about women and relationships.

[1] Just unsubbed from r/seduction
[2] Toxic view women

Nevertheless, it is important to note that this is only a perception of the subreddit and some people may have a different view on it.

Based on these beliefs, the research questions that guided the development of the project are the following:

- Analysis of the subreddit's contents, discussed topics and their relative sentiment
- Identification of most influential users
- Analysis of the subreddit structure and users interactions
- Identification of communities and their relative characteristics

3. Data collection

The Pushshift.io API Wrapper (PSAW) library [2] was implemented in order to collect data. This method allows to specify the desired time period for the data to be crawled and it does not have the constraint of the maximum number of post to be retrieved. Reddit's API was not used, as it is limited to collect at maximum 1000 posts and this would have resulted in a problem if the number of data to retrieve were higher. The period of interest was set to be October 2022. A total of 1349 posts were collected, but those who consisted only in a title and a removed body text were discarded, since they were posts removed by the user, reducing the number of posts to 756. On the other hand, the total number of comments collected is 14489. For both of the datasets, only the features considered to be important for the analysis are kept.

For the posts dataset the attributes are:

- **title**: title of the post
- **text**: body of the post
- **author**: author of the post
- **id**: post identifier
- **created_utc**: time the submission was created, represented in Unix Time
- **URL**: post's URL
- **upvote_ratio**: percentage of upvotes from all votes on the submission
- **num_awards**: number of awards the post has received

For the comments dataset the attributes are:

- **author**: author of the comment
- **body**: comment's content
- **comment_id**: id of the comment
- **created_utc**: time the comment was created, represented in Unix Time
- **URL**: comment's URL
- **n_upvotes**: upvotes of the comment
- **parent_id**: the ID of the parent comment (prefixed with t1_). If it is a top-level comment, this returns the submission ID instead (prefixed with t3_)

4. Exploratory analysis

The first analysis consisted in the frequency distribution of the publication of posts, considering the period starting from October 1st to October 31st, as showed in the graph below.

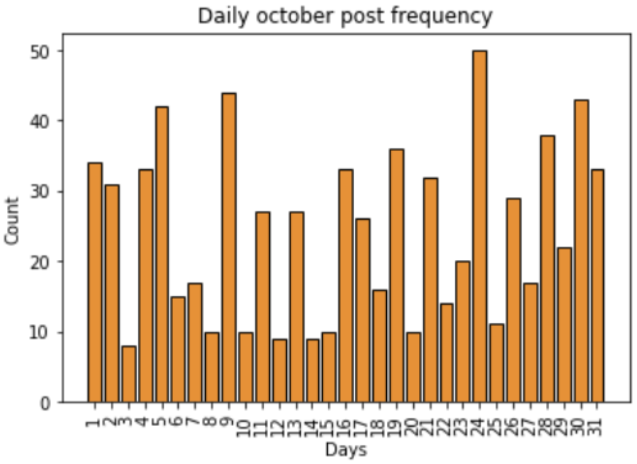

Figure 1. Post of october.

The maximum number of posts published corresponds to 50, dated October 23rd. On the other hand, the minimum number of posts published corresponds to 8, dated October 3rd. On average, 24 posts per day were published, with a standard deviation of 12,4.

In order to understand whether there is a user that stood out in publishing posts, the frequency distribution of published posts per author was also analysed. It emerged that there are 521 users that did publish at least one post, the average number of published posts per author is 1.5, with a minimum of 1 and a maximum of 21 posts, published by "EntertainerMaximum79."

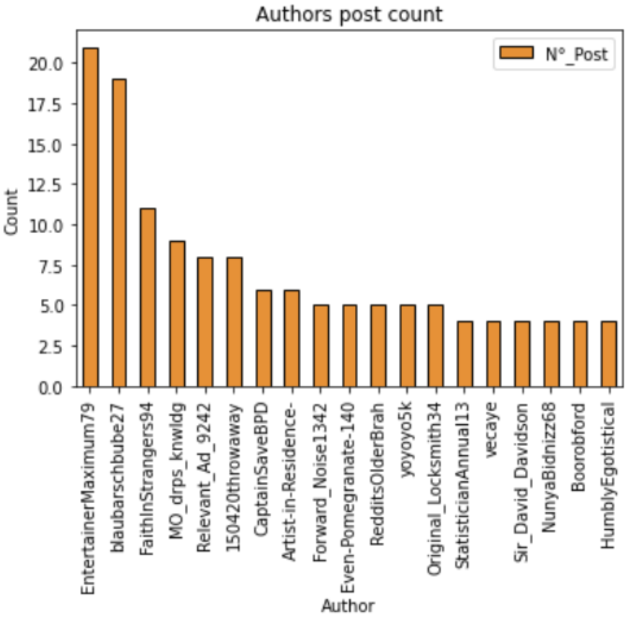

Figure 2. Author of the post.

Taking into consideration the comments dataset, the authors with the most successful and the most criticised comments are, respectively, ”thatstotallyracist” and ”Environmental_Cress2”. The first one collected 552 votes with the comment ”Are you telling me that there are hot, single women waiting to meet me in my area?” and the other ”Environmental_Cress2” who collected -63 votes with the comment ”prolly. I prefer nerdy girls. But I’m mostly around party tiktok type of girls. So that affects my perception”.

Furthermore, for the total number of comments a mean of 6 was computed, with a standard deviation of 22.

The second analysis regards the distribution of the comments published during the period of time starting from October 1st to October 31st, as shown in the graph below.

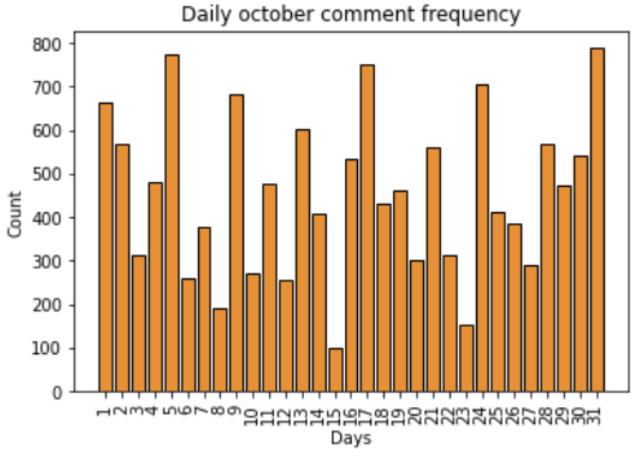

Figure 3. Comments of october.

Throughout the month of October, the number of comments ranged from a minimum of 99 on October 15th to a maximum of 789 on October 31st. On average, there were 450 comments published per day, with a standard deviation of 186.

5. Graph

In order to gain insights about the r/seduction community, the interactions between the authors in the subreddit were modelled into a graph. An edge from the commenter to the author of the commented content was created every time a comment was made. The weight of this link was determined by the number of upvotes the comment received. However, not all weights turned out being greater than or equal to zero; in those cases, the weight was set by default to be 0,1. The result was a directed, weighted, multigraph, composed by 4145 different nodes, or authors, and 13690 edges, whose disconnected components were removed.

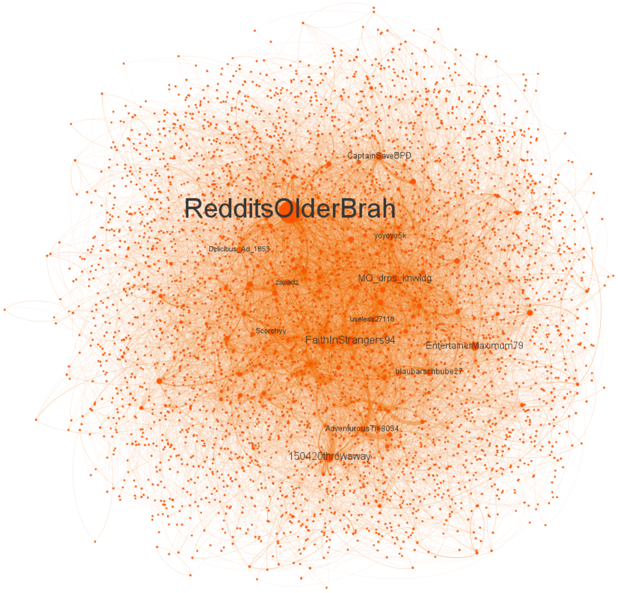

Figure 4. Community’s graph

5.1 Metrics

With an assortativity coefficient of -0.07, the graph is neither assortative nor disassortative, but it can be categorized as neutral, meaning that nodes are connected to each other with probabilities consistent with randomness.

The network density is 0.0008, which indicates that there are very few connections in the network compared to the total number of possible connections; the network can be categorized as sparse.

Other metrics such as diameter and radius of the graph could not be calculated, since the graph was not strongly connected. Hence, there is at least one subgraph in which a specific node can be reached by another one, but not the opposite.

In order to understand who the main influencers of the network are, four normalized centrality metrics were computed: in-degree and out-degree centrality, betweenness centrality and closeness centrality. The in-degree of a node is the number of edges that point towards it, while the out-degree of a node is the number of edges that point away from it. These measures allow us to understand the connections and influence of a node within the graph.

User	In-degree centrality
RedditsOlderBrah	0.084
FaithInStrangers94	0.053
MO_drps_knwldg	0.048
150420throwaway	0.044
EntertainerMaximum79	0.039

Table 1. Top 5 highest in-degree centralities

As it emerges from the tables 1-2, it’s possible to notice that a user, namely “RedditsOlderBrah”, has the most connections in the whole network, both incoming and outgoing; this means that he’s very active in the community and also receives many replies, making him a very popular member.

User	Out-degree centrality
RedditsOlderBrah	0.098
zapadz	0.027
g3wb3r	0.024
incognito3107	0.024
Canadian-Seductioner	0.023

Table 2. Top 5 highest out-degree centralities

The betweenness centrality of a node is a measure of how often that node lies along the shortest path between other nodes. A node with high betweenness centrality has a high potential to control the flow of information or resources through the network, because it lies on many shortest paths.
On the other hand, the closeness centrality of a node is a measure of how close the node is to all other nodes in the network. A node with high closeness centrality has a short average distance to all other nodes, so it is able to reach other nodes quickly and efficiently; moreover, a node with this characteristic is often considered to be well-connected and influential within the network. These metrics are shown in the below tables.

User	Betweenness centrality
RedditsOlderBrah	0.106
FaithInStrangers94	0.034
MO_drps_knwldg	0.026
150420throwaway	0.020
EntertainerMaximum79	0.020

Table 3. Top 5 highest betweenness centralities

User	Closeness centrality
RedditsOlderBrah	0.303
FaithInStrangers94	0.287
MO_drps_knwldg	0.278
EntertainerMaximum79	0.273
blaubarschbube27	0.272

Table 4. Top 5 highest closeness centralities

It is reasonable to confirm the importance of the user ”RedditsOlderBrah”; he is also well-connected to other users and lies on many shortest paths between other nodes.
However, “FaithInStrangers94” can also be considered as a central node, who is close to other users, since he has a closeness centrality only 0.15 lower than the main hub.

6. Text analysis

In order to get more information about the subreddit's community and analyze it in more detail, sentiment analysis and topic modeling were carried out.

6.1 Sentiment Analysis

Sentiment Analysis was performed to detect the polarity (positive, negative, neutral) of posts and comments' texts. The model used is VADER (Valence Aware Dictionary for Sentiment Reasoning) [3], that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. More precisely, it relies on a dictionary that maps lexical features to emotion intensities, known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text, which ranges between -1 and 1. A score close to -1 indicates a very negative sentiment, a score close to 1 indicates a very positive sentiment, and a score close to 0 indicates a neutral sentiment.
For both datasets the preprocessing steps consisted in removing numbers, URLs, emails, newlines, extra whitespaces, hashtags, other comments quotations, Reddit's text formatting characters and texts whose author is 'seduction-ModTeam', the subreddit's moderators account. Texts from this account are formal and pre-written messages that indicate which rule a user has violated and that the post has been removed, and were not useful for the analysis.
Stopwords removal, instead, was not applied in order to avoid loosing meaningful information for the sentiment analysis. Considering that the model performs very well with emojis, slangs and emoticons, they were not removed from the text and some punctuation that can have a great influence on the sentiment analysis was kept too, i.e. those necessary for emoticons as well as question and exclamation marks, that can increase the magnitude of the intensity without modifying the semantic orientation. Furthermore, lower case folding was not applied either, because the use of upper case letters to emphasise a sentiment-relevant word in the presence of other non-capitalized words.
Finally, the lemmatization technique was applied, using the SpaCy library [4]. This library includes a built-in lemmatizer, which is a function that reduces a word to its base or root form. The lemmatizer has the ability to take into account the context of the word, which helps to improve the accuracy of the lemmatization. Therefore, making it easier to identify words that express the same sentiment, even if they appear in different forms in the text. This ultimately improves the accuracy of the sentiment analysis.
The results are represented in figures 5 and 6. Considering the posts dataset, 72.6% of them are positive, 4.2% are neutral and 20.2% are negative. Considering the comments dataset, 59.4% of them are positive, 16.6% are neutral and 23.9% are negative.
The sentiment analysis of both the posts and comments datasets reveals a similar trend, with a higher prevalence of positive sentiment compared to negative sentiment. Neutral sentiment is the least prevalent among the data set.

6.2 Topic modeling

To understand the community and identify the interests and main topics of discussion among its users, a topic modeling

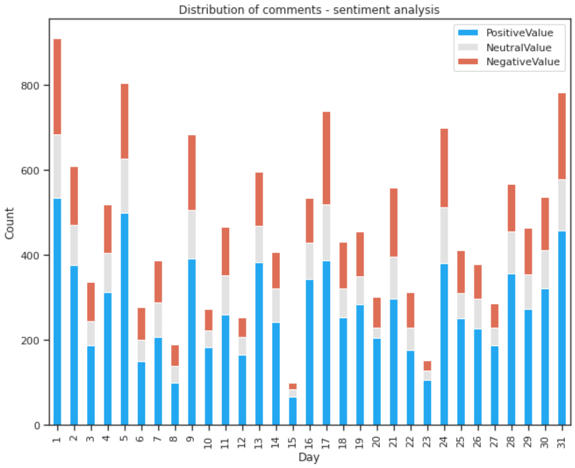

Figure 5. Sentiment Analysis performed on the distribution of the comments. Month October 2022.

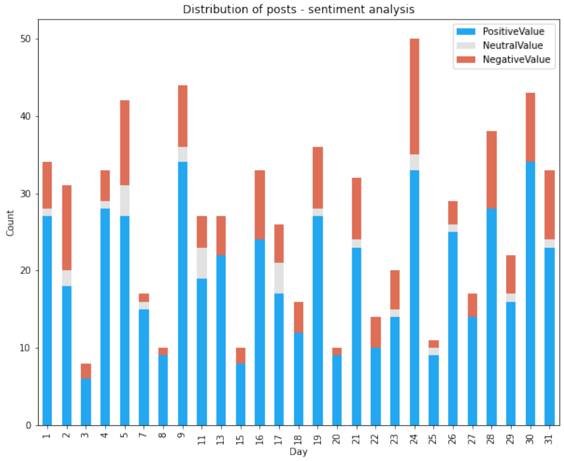

Figure 6. Sentiment Analysis performed on the distribution of the posts. Month October 2022.

task was performed. This was done using the Latent Dirichlet Allocation (LDA) on text data that had undergone these preprocessing steps:

- Lowercasing
- Removing special characters, numbers, and punctuation marks
- Removing stopwords and URLs
- Removing characters repetition
- Tokenization and lemmatization
- Removing words that are either too rare or too common, which happear in less than 4 documents and more than the 35% of them.

The LDA run on text data represented in a Bag-of-Words (BOW) format allowed to recognize topics such as:

- Approaching girls
- Ways to improve appearance
- Approaches and rejections
- First dates
- Relationships
- Sexual relations
- Body language

In the below figure, the most frequent words and topics are visualized in a wordcloud.

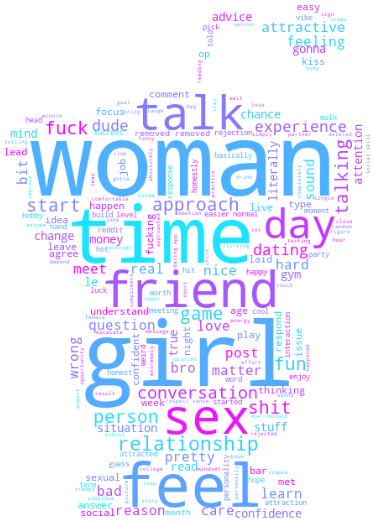

Figure 7. Comments wordcloud

7. Community detection and analysis

An overall graph view and analysis provides a broad understanding of the structure and organization of a network as a whole. Nevertheless, discovering hidden patterns and structures that are not immediately apparent is not straightforward; for this reason, the community detection task was performed.

To discover communities, the greedy_modularity_communities algorithm from networkx library [5] was implemented. A greedy community detection algorithm is a method that iteratively improves the modularity of a network division by displacing nodes between communities. The algorithm terminates when no more improvement in modularity can be achieved. At the end of this process, the number of communities found was 26 and the modularity coefficient was equal to 0.43. Among these communities, only the biggest 4 were analyzed, which were respectively composed by 390, 381, 343 and 329 users.

Community	Positive	Neutral	Negative
1	58.1%	14.4%	27.4%
2	58.9%	16.7%	25.4%
3	58.7%	17.8%	23.5%
4	54.7%	20.9%	24.5%

Table 5. Four biggest communities’ sentiment

As it's possible to note, the proportions of positive, negative, and neutral comments are relatively similar across all clusters. However, to deepen into this aspect, the most negative and most positive communities were examined and the results are presented in Table 6. The initially identified most positive community, with a 76% ratio of positive comments, was not considered due to its small sample size of only 17 individuals. Instead, the second-largest one, with 365 members, was taken into account. On the other hand, the most negative cluster was composed of 75 individuals.

	Communities	
	Most Positive	Most Negative
Positive	69.3%	52.0%
Neutral	14.5%	13.3%
Negative	16.2%	34.7 %

Table 6. Most positive and most negative communities' sentiment

The results show that the proportion of sentiments remains uniform across all clusters, meaning that the ratio of positive, negative, and neutral comments is mostly steady among all communities.

8. Conclusions

The objective of this project consisted in analyzing the subreddit r/seduction, its contents, discussed topics and their relative sentiment, interaction within the network, identify the most influential users, the communities, as well as their characteristics. Data of posts and comments were collected for the month of October 2022. In order to gain insights about the r/seduction community, the interactions between the authors were modelled into a graph. Sentiment Analysis was performed to detect the polarity (positive, negative, neutral) of posts and comments' texts. Finally, to better understand the community and to identify the users interests and main discussed topics, topic modeling was performed. The analyses provide interesting results; among all the users, "RedditsOlderBrah" can be described as both the most influential and central one, having the highest score in all analyzed metrics. The most prevalent sentiment emerged from the analysis is the positive one; this result goes against the negative perceptions towards the community, that guided the development of this project. Additionally, the most discussed topics deals with ways to improve appearance, approaches to girls, rejections, first dates, relationships, sexual relations and body language. Finally, 26 communities were found; among the four biggest communities analyzed, the proportions of positive, negative, and neutral comments are relatively similar. The most negative community reached 34.7% of negative comments, instead the most positive one reached 69.3% of positive comments; it was proven again that even the most negative cluster is not as extremely polarized as initally imagined. The findings of this study serve as a preliminary examination of the content and discourse within the r/seduction subreddit. Further exploration and validation of these results could be conducted through additional data collection and analysis over an extended period of time. This would provide a more robust and comprehensive understanding of the community and its interactions. These insights could serve as a foundation for future research and analysis.

References

[1] Reddit. r/seduction. https://www.reddit.com/r/seduction/.

[2] David Marx. PSAW: Python Pushshift.io API Wrapper. https://psaw.readthedocs.io/en/latest/.

[3] E.E. Hutto, C.J. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 2014.

[4] spaCy API Documentation. Lemmatizer. https://spacy.io/api/lemmatizer.

[5] networkx. greedy_modularity_communities. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.modularity_max.greedy_modularity_communities.html.