

# An investigation on Not Safe For Work adult content in Reddit

(Discussion Paper)

Francesco Cauteruccio<sup>1</sup>, Enrico Corradini<sup>2</sup>, Giorgio Terracina<sup>1</sup>, Domenico Ursino<sup>2</sup> and Luca Virgili<sup>2</sup>

<sup>1</sup>DEMACS, University of Calabria, Italy

<sup>2</sup>DII, Polytechnic University of Marche, Italy

## Abstract

Reddit is one of the few social platforms that handles NSFW (Not Safe For Work) content in an explicit and well-structured way. Despite this fact, such an issue has been very neglected in the past by researchers who have studied this social network. In this paper, we aim at providing a contribution in this setting by proposing an approach to extract and analyze text patterns from NSFW content in Reddit. An important peculiarity of our approach is that patterns are extracted not only based on their frequency (as it generally happens in the past literature), but also, and especially, on one or more utility measures.

## Keywords

Reddit, NSFW posts and comments, Text patterns, Pattern utility measures, Social Network Analysis

## 1. Introduction

The term “Not Safe For Work” (hereafter, NSFW) is used by many social media to refer to content within them that cannot be viewed in public or professional settings. The study of the phenomenon of NSFW content in social media has attracted many authors (e.g., [1]). One of the social platforms that has adopted the concept of NSFW in an explicit and well-structured way is Reddit. However, despite this, few researchers have investigated the features of NSFW content in Reddit [2, 3, 4]. In particular, in [4], the authors use Social Network Analysis to investigate NSFW posts on Reddit. They focus on the structural features of the posts without analyzing their content. In this paper, we want to continue the research efforts of [4] and, once again, we use Social Network Analysis to study NSFW posts and comments on Reddit. However, here, we shift the research focus from structure to content.

More specifically, we propose an approach for extracting and analyzing text patterns present in NSFW adult content on Reddit. In our context, we use the term “pattern” to refer to a set of words in posts or comments that satisfy certain properties. Our approach consists of three main steps, namely: (i) Data Cleaning and Annotation, (ii) Pattern Extraction and Enrichment,

---


SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

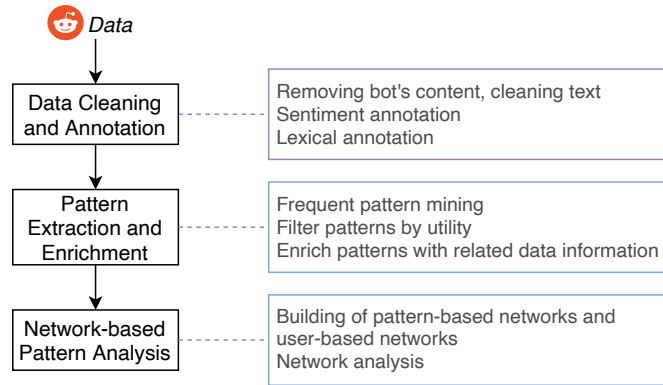
✉ cauteruccio@mat.unical.it (F. Cauteruccio); e.corradini@pm.univpm.it (E. Corradini); terracina@mat.unical.it (G. Terracina); d.ursino@univpm.it (D. Ursino); l.virgili@pm.univpm.it (L. Virgili)

🆔 0000-0001-8400-1083 (F. Cauteruccio); 0000-0002-1140-4209 (E. Corradini); 0000-0002-3090-7223 (G. Terracina); 0000-0003-1360-8499 (D. Ursino); 0000-0003-1509-783X (L. Virgili)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



**Figure 1:** The general workflow of our approach

and (iii) Network-based Pattern Analysis. Applying our approach on Reddit allowed us to make several contributions to this research scenario. They involve: (i) discovering that traditional approaches to sentiment computation are unreliable in the context of NSFW adult content; (ii) defining and finding opinion leaders in real communities sharing NSFW adult content; (iii) discovering text patterns representing the building blocks of NSFW posts and comments on Reddit; (iv) determining new virtual communities of users sharing NSFW adult content; (v) identifying opinion leaders who could influence such communities.

The rest of this paper is structured as follows: Section 2 provides a general description of our approach and the dataset used for our experiments. Section 3 illustrates the Pattern Extraction and Enrichment step. Section 4 describes the Network-based Pattern Analysis step. Finally, Section 5 presents our conclusions and possible future developments of our research efforts.

## 2. General overview of our approach

The general workflow of our approach is shown in Figure 1, which highlights the three steps composing it (i.e., Data Cleaning and Annotation, Pattern Extraction and Enrichment, and Network-based Pattern Analysis).

The *Data Cleaning and Annotation* step removes irrelevant content and standardizes text representations. It also performs lexical (e.g., part-of-speech and named entities) and sentiment annotations. These latter highlight the polarity of sentiments expressed in the texts, represented in terms of a compound score, computed by applying Vader [5]. Due to space limitations, we do not illustrate this step in detail in this paper.

The *Pattern Extraction and Enrichment* step extracts a set of text patterns from the posts and comments identified in the previous step; they are the basis for the next Network-based Pattern Analysis step. To this end, it first extracts frequent patterns. Then, it associates each pattern with a rich set of features regarding the posts and comments it derives from, as well as the users who published them. Afterwards, it defines some utility measures and associates the corresponding values with each pattern. Finally, it selects only those patterns with high frequency and high utility. Our approach allows the definition of different concepts and utility

measures and, consequently, the selection of different sets of useful patterns based on them. This allows us to analyze the available NSFW content from very different perspectives, yet adopting a uniform methodology.

The *Network-based Pattern Analysis* step applies the concepts and approaches of Social Network Analysis to the patterns obtained during the previous step with the goal of extracting information and knowledge from them. Specifically, it constructs and uses three social networks, namely: (i) *User Interaction Network*, in which a node  $n_i$  represents a user  $u_i$ , who published at least one post or comment. An arc  $(n_i, n_j, w_{ij})$  denotes that  $u_i$  commented a post of  $u_j$ ;  $w_{ij}$  indicates how many times this happened. (ii) *Pattern Network*, in which a node  $n_i$  represents a pattern  $p_i$  extracted in the previous step. An arc  $(n_i, n_j, w_{ij})$  indicates that  $p_i$  and  $p_j$  were adopted by at least one user in common;  $w_{ij}$  indicates the number of users who adopted both  $p_i$  and  $p_j$ . (iii) *User Content Network*, in which a node  $n_i$  represents a user  $u_i$ , who published at least one post or comment. An arc  $(n_i, n_j, w_{ij})$  indicates that there is at least one comment posted by  $u_i$  and at least one comment posted by  $u_j$  containing the same pattern;  $w_{ij}$  denotes the number of times this happened. Once these networks are built, this step proceeds by applying Social Network Analysis concepts and approaches to them for extracting information and knowledge on Reddit users publishing, commenting and reading NSFW adult posts and on the content these users exchange. Due to space limitations, in this paper, we focus only on the *User Interaction Network*.

To perform our experiments for evaluating our approach, we downloaded a dataset from the `pushshift.io` [6] website, which represents one of the main data repositories for Reddit. Specifically, we considered 449 NSFW adult subreddits listed at <https://www.reddit.com/r/ListOfSubreddits/wiki/nsfw> and extracted all posts and the corresponding comments published on them from January 1<sup>st</sup>, 2020 to March 31<sup>st</sup>, 2020. The number of posts on the dataset is 3,064,758, while the number of comments is 11,627,372.

### 3. Pattern Extraction and Enrichment

This step extracts text patterns from posts and comments in the dataset and, then, enriches them with additional information concerning their frequency and utility. Pattern mining plays a key role in this activity. It is a well known task in the literature, which extracts from a dataset some (hopefully interesting and/or unexpected) information that can be understood by humans.

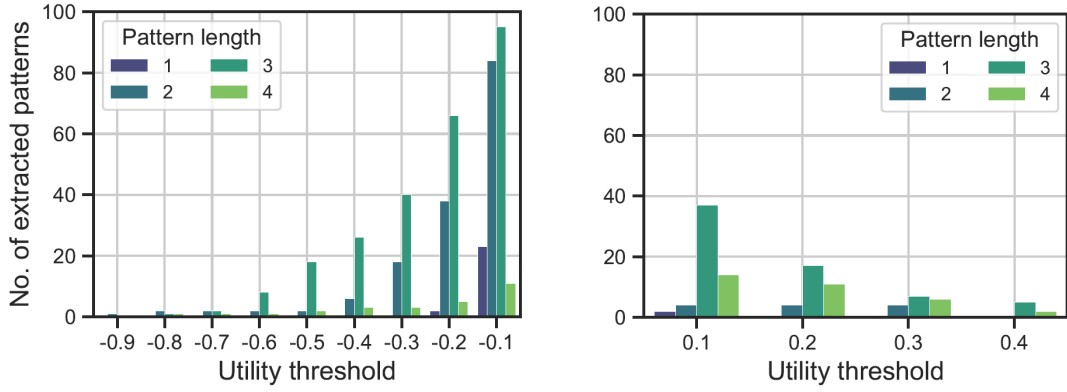
Many pattern mining approaches are based on the concept of *pattern frequency* and aim at identifying the most frequent patterns in the texts received as input. They are based on the assumption that frequent patterns are interesting [7, 8]. This is true in many application contexts. However, there are cases where it does not hold. To handle these cases, the notion of *pattern utility* has been introduced. It shifts the emphasis from frequent pattern mining to High Utility Pattern Mining (hereafter, HUPM) [9, 10]. In this case, a utility function is defined; the patterns with a high value of this function are considered interesting. Recall that a utility function denotes a user preference ordering over a set of choices [9]. It is clearly a subjective measure allowing us to state the usefulness of a text pattern from different perspectives, depending on our preferences and/or needs.

After having introduced the concepts of frequency and utility of a pattern, we can illustrate

our approach to pattern extraction and the model which it operates on. Let  $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$  be a set of lemmatized comments, obtained at the end of the Data Cleaning and Annotation step. Each comment  $c_i \in \mathcal{C}$  corresponds to a post and is written by a Reddit user. We can represent  $c_i$  as a set of lemmas  $c_i = \{l_1, l_2, \dots, l_m\}$ . Therefore, if we denote by  $\mathcal{L} = \{l_1, l_2, \dots, l_q\}$  the set of all possible lemmas, then  $c_i$  is a subset of  $\mathcal{L}$ . From the HUPM perspective, each lemma is an item. A pattern  $P_j$  is a set of items and, therefore,  $P_j \subseteq \mathcal{L}$ .  $P_j$  can occur in zero, one or more comments in our dataset. We denote by  $\mathcal{C}_j \subseteq \mathcal{C}$  the subset of the comments of  $\mathcal{C}$  in which  $P_j$  is present, and by frequency of  $P_j$  the cardinality of  $\mathcal{C}_j$ .  $P_j$  inherits the set of features characterizing the comments of  $\mathcal{C}_j$ , and the utility of  $P_j$  can be defined as an appropriate function of all or some of these features. The choice of the features and the utility function to adopt determine the perspective one wishes to consider in the analysis of patterns.

For example, consider the features `score_comm` (denoting the score of a comment) and `compound` (indicating the sentiment value extracted from the text of a comment). Suppose that the utility function is the Pearson's correlation [11] between them, which allows us to say whether there is a form of correlation between the two features, such that a high (resp., low) score of a comment arouses a positive (resp., negative) sentiment about it. This function allows us to select those patterns whose presence in comments with high (resp., low) scores is flanked by a positive (resp., negative) sentiment. We point out that this correlation between score and sentiment is not obvious for comments because there could exist comments with high (resp., low) score and null or negative (resp., positive) sentiment. In the following, we call  $f_p$  the utility function computing the Pearson correlation. It is worth investigating both patterns having a positive value of  $f_p$  and those having a negative value of that function. Indeed, a positive (resp., negative) value of  $f_p$  indicates that there is a direct (resp., inverse) correlation between the sentiment aroused by a comment and the score it obtains. Consequently, we denote by  $f_p^+$  (resp.,  $f_p^-$ ) the function that selects those patterns having a value of  $f_p$  greater (resp., lesser) than a threshold  $th_p^+$  (resp.,  $th_p^-$ ).

Figure 2 shows the trend of the number of extracted patterns as  $th_p^-$  (resp.,  $th_p^+$ ) decreases (resp., increases). Patterns are also grouped based on their length. This figure provides us with non-obvious and extremely interesting knowledge. In fact, the number of patterns extracted by  $f_p^-$  is much greater than the one extracted by  $f_p^+$ . This allows us to say that, given a pattern, a positive (resp., negative) sentiment of it is not necessarily accompanied by a high (resp., low) score of the comments where it is present. This phenomenon is very evident for moderately positive or negative values of  $f_p$ , while it reduces strongly for extreme values. It can be explained by considering that, given the nature of the reported texts, NSFW posts and comments tend to be associated with a negative sentiment by any sentiment analysis tool. This happens even when such terms are used in goliardic comments, which are actually appreciated by this type of audience. For instance, consider the text pattern  $\{hot, fuck\}$ , possibly accompanied by an emoticon with two little hearts instead of eyes. Applying the sentiment analysis tools available in the literature to our dataset [5], we obtained a sentiment value of -0.1280 for this pattern. Instead, the corresponding comments have a very high score. As a consequence, the value of  $f_p$  is negative. This allows us to obtain a first important outcome of our approach, namely that traditional sentiment computation tools do not work well in presence of NSFW posts and comments. As for the choice of the values of  $th_p^+$  and  $th_p^-$ , all the reasoning above, and the presence of a low number of extracted patterns, led us to choose low values for the two



**Figure 2:** Number of extracted patterns against negative (at left) and positive (at right)  $f_p$

thresholds. In particular, we set  $th_p^+ = 0.1$  and  $th_p^- = -0.1$ .

We end this discussion by pointing out that many other utility functions could be defined on many different features concerning posts and users. This important peculiarity of our approach allows us to analyze the phenomenon of NSFW content from many different perspectives.

## 4. Analysis of User Interaction Networks

In this section, we formally introduce the User Interaction Network and show how information can be derived from it. Let  $\mathcal{P}_f$  be the set of patterns obtained by applying the utility function  $f$ , and let  $\mathcal{U}_f$  be the set of users who have written at least one comment or post containing at least one of the patterns of  $\mathcal{P}_f$ . A User Interaction Network  $\mathcal{N}^{ui}$  can be defined as:  $\mathcal{N}^{ui} = \langle N^{ui}, A^{ui} \rangle$ .  $N^{ui}$  is the set of nodes of  $\mathcal{N}^{ui}$ . There exists a node  $n_i \in N^{ui}$  for each user  $u_i \in \mathcal{U}_f$ .  $A^{ui}$  is the set of arcs of  $\mathcal{N}^{ui}$ . An arc  $(n_i, n_j, w_{ij}) \in A^{ui}$  denotes that  $u_i$  made comments on a post published by  $u_j$ ;  $w_{ij}$  measures how many times this task happened.

$\mathcal{N}^{ui}$  allows us to characterize the behavior of users, who interact with each other by posting and commenting NSFW adult content.

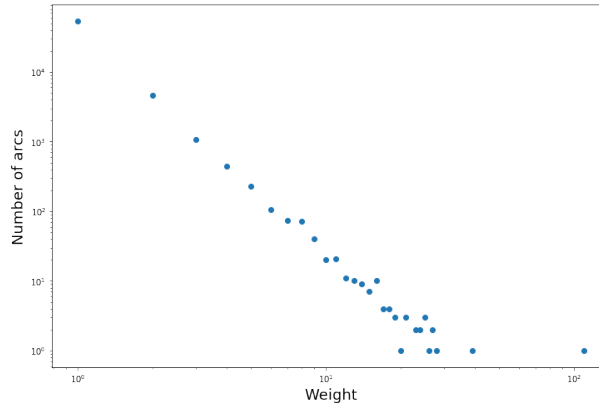
Table 1 lists some parameters of the User Interaction Networks built exploiting the dataset outlined in Section 2, the utility functions  $f_p^+$  and  $f_p^-$ , and the thresholds  $th_p^+$  and  $th_p^-$ . This table confirms the information derived in Section 3. For instance,  $\mathcal{N}_{f_p^-}^{ui}$  contains a larger number of nodes and arcs than  $\mathcal{N}_{f_p^+}^{ui}$ . Note that the cardinalities of  $\mathcal{P}_{f_p^-}$  and  $\mathcal{P}_{f_p^+}$  follow the same trend as the number of nodes (and users) in the User Interaction Networks, even though they represent patterns instead of users. It is worth observing the high density coupled with a high clustering coefficient characterizing  $\mathcal{N}_{f_p^+}^{ui}$ . This tells us that, in this network, users tend to form very tight communities, whose structure resemble that of a clique. Instead,  $\mathcal{N}_{f_p^-}^{ui}$  has a high density but a low clustering coefficient; this suggests the presence of very strong power users, i.e., users receiving comments from many other ones, who do not actually interact with each other. In both networks, the number of nodes composing the maximum connected component is very high, i.e., about 60%.

**Table 1**  
Values of some parameters for User Interaction Networks

Parameter	$\mathcal{N}_{f_p^-}^{ui}$	$\mathcal{N}_{f_p^+}^{ui}$
Nodes	27,160	1,452
Arcs	60,662	7,925
Density	8.224e-05	376.15e-05
Clustering coefficient	0.004	0.129
Number of connected components	10,939	506
Size of the maximum connected component	16,030	891
Average weight of arcs	1.205	1.935
Average Indegree (weight $\geq 2$ )	16.634	14.393
Average Outdegree (weight $\geq 2$ )	6.03	7.473
Average Clustering coefficient (weight $\geq 2$ )	0.011	0.139
Average Indegree (All)	1.921	1.973
Average Outdegree (All)	1.912	1.987
Average Clustering coefficient (All)	0.003	0.039

The average number of times a user comments the content of another one is slightly higher than 1; this is confirmed by the very low average weight of arcs. In Figure 3, we investigated this aspect and found a power law distribution of the number of arcs against weights for  $\mathcal{N}_{f_p^-}^{ui}$ . It implies that the distribution of the number of comments of a user against the contents posted by another one follows a power law. An analogous result holds for  $\mathcal{N}_{f_p^+}^{ui}$ . We determined the parameters  $\alpha$  and  $\delta$  of these power law distributions; they are:  $\alpha = 1.371, \delta = 0.062$  for  $\mathcal{N}_{f_p^-}^{ui}$  and  $\alpha = 1.507, \delta = 0.063$  for  $\mathcal{N}_{f_p^+}^{ui}$ . The results of these analyses show that there is a low number of pairs of users in which one of them comments a post of the other at least twice (we call them *interacting users* in the following). This can be considered a minimum condition to detect non-random relationships between pairs of users. We compared the values of the average indegree, outdegree and clustering coefficient of the interacting users, on one hand, and all the users, on the other hand. The bottom of Table 1 shows this comparison for the two considered networks. It is easy to observe that, in both networks, interacting users have indegrees and outdegrees much higher than the other users. Therefore, they can be considered power users. Also, their clustering coefficient is very high, indicating that they are able to promote the generation of communities. Therefore, it can be said that they are community leaders in the distribution of NSFW adult content in Reddit.

At this point, we found it interesting to investigate the possible existence of mutual relationship between interacting users. For this purpose, we determined the fraction of interacting users such that a user  $u_i$  comments the posts of a user  $u_j$ , and vice versa. We found that this fraction is low (i.e., 0.141) for  $\mathcal{N}_{f_p^-}^{ui}$ , whereas it is higher (i.e., 0.433) for  $\mathcal{N}_{f_p^+}^{ui}$ . Furthermore, although the number of nodes of  $\mathcal{N}_{f_p^+}^{ui}$  is considerably lower than that of  $\mathcal{N}_{f_p^-}^{ui}$ , the two networks have similar average indegree and outdegree for both normal and interacting users. Moreover,  $\mathcal{N}_{f_p^+}^{ui}$  has a much higher clustering coefficient and a much higher fraction of interacting users than  $\mathcal{N}_{f_p^-}^{ui}$ . Based on all these outcomes, we can state that, although the number of users of  $\mathcal{N}_{f_p^-}^{ui}$  is much greater than that of  $\mathcal{N}_{f_p^+}^{ui}$ , these last ones have a higher attitude to be opinion leaders. Taking also the utility function underlying  $\mathcal{N}_{f_p^+}^{ui}$  into account, we can deduce that the users of this network are particularly able to maintain a positive correlation between the sentiment of



**Figure 3:** Distribution of arcs against weights for  $\mathcal{N}_{f_p}^{ui}$  (log-log scale)

their comments and the associated scores. Finally, the results obtained so far allow us to state that the users of  $\mathcal{N}_{f_p}^{ui}$  are the most dynamic ones, as they publish posts attracting interest (since they are commented by other users), and comment the posts of the other users. This feature makes them particularly important, as they are not only content producers, but also dynamic participants who contribute to maintain their communities active, and act as opinion leaders for these communities. We call them *proactive* users.

## 5. Conclusion

This paper presented an approach to analyze NSFW users, comments and posts on Reddit, taking into account and exploiting the knowledge extracted by investigating text patterns. The methods and results considered in this paper can represent the basis for many new applications. In fact, posts and subreddits of other target categories (e.g., vegetarian or vegan users) could be examined by means of the same methodology. Moreover, the analysis of text patterns could represent the engine of an automatic classifier aiming at tagging posts and comments containing unsuitable content. Furthermore, our approach can be adapted to other social networks managing NSFW content less explicitly than Reddit. Finally, we plan to design an automatic tool that exploits a knowledge base built by integrating utility patterns and semantic analysis tools to automatically classify new contents, and, then, suggest the most pertinent communities which they should be directed to.

## References

- [1] K. Tiidenberg, Boundaries and conflict in a NSFW community on tumblr: The meanings and uses of selfies, *New Media & Society* 18 (2016) 1563–1578. Sage Publications.
- [2] J. Matias, Going dark: Social factors in collective action against platform operators in the Reddit blackout, in: *Proc. of the International Conference on Human Factors in Computing Systems (ACM CHI 2016)*, San Jose, CA, USA, 2016, pp. 1138–1151. ACM.



- [3] B. K. Narayanan, M. Nirmala, Adult content filtering: Restricting minor audience from accessing inappropriate Internet content, *Education and Information Technologies* 23 (2018) 2719–2735. Springer.
- [4] E. Corradini, A. Nocera, D. Ursino, L. Virgili, Investigating the phenomenon of NSFW posts in Reddit, *Information Sciences* 566 (2021) 140–164. Elsevier.
- [5] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: *Proc. of the International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, Ann Arbor, MI, USA, 2014, pp. 216–225.
- [6] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The pushshift Reddit dataset, in: *Proc. of the International AAAI Conference on Web and Social Media (ICWSM'20)*, volume 14, Atlanta, GA, USA, 2020, pp. 830–839. AAAI Press.
- [7] P. Fournier-Viger, J. C.-W. Lin, B. Vo, T. Chi, J. Zhang, H. Le, A survey of itemset mining, *WIREs Data Mining and Knowledge Discovery* 7 (2017) e1207. doi:<https://doi.org/10.1002/widm.1207>, Wiley.
- [8] C. Aggarwal, M. Bhuiyan, M. A. Hasan, Frequent pattern mining algorithms: A survey, in: J. H. C. Aggarwal (Ed.), *Frequent Pattern Mining*, 2014, pp. 19–64. doi:[10.1007/978-3-319-07821-2\\_2](https://doi.org/10.1007/978-3-319-07821-2_2), springer, Cham.
- [9] P. Fournier-Viger, J.-W. Lin, R. Nkambou, B. Vo, V. Tseng, *High-Utility Pattern Mining*, 2019. Springer.
- [10] W. Gan, C. Lin, P. Fournier-Viger, H. Chao, V. Tseng, P. Yu, A survey of utility-oriented pattern mining, *IEEE Transactions on Knowledge and Data Engineering* 33 (2021) 1306–1327. doi:<https://doi.org/10.1109/TKDE.2019.2942594>, IEEE.
- [11] K. Pearson, Note on regression and inheritance in the case of two parents, *Proceedings of the Royal Society of London* 58 (1895) 240–242. The Royal Society.