Journal of Computational Social Science
https://doi.org/10.1007/s42001-024-00269-4

RESEARCH ARTICLE

Content and interaction-based mapping of Reddit posts related to information security

Konstantinos Charmanas[1] · Nikolaos Mittas[2] · Lefteris Angelis[1]

Received: 27 October 2023 / Accepted: 9 March 2024

Abstract

Ensuring the privacy and safety of platform users has become a complex objective due to the emerging threats that surround any type of network, software, and hardware. Scams, malwares, hackers, and security vulnerabilities form the epicenter of cyber threats causing severe damage to the affected systems and sensitive data of users. Thus, users turn to online social networks to report cyber threats, discuss topics of their interest, and obtain knowledge concerning the various perspectives of information security. In this study, we aim to address the concepts of social interactions surrounding information security-related content by retrieving and analyzing Reddit posts from 45 relevant subreddits. In this regard, a word clustering approach is employed, based on the Affinity Propagation algorithm, that leads to the extraction and interpretation of 54 concepts. These concepts are relevant to information security and some more generic areas of interest including social media, software vendors, and labors. Furthermore, to provide a more comprehensive overview of users' activity in the different Reddit communities/subreddits, a knowledge map associating subreddits and concepts based on their conceptual similarities is also established. The analysis shows that the descriptions of the examined subreddits are strongly related to their underlying concepts. At the same time, the outcomes also assess the conceptual associations between the different subreddits, offering knowledge related to similar and distant communities. Ultimately, two post metrics are utilized to explore how the concepts may impact user interactions. This allows us to differentiate between concepts associated with posts typically endorsed by communities, resulting in increased information exchange (via comments), or contributing as news/announcements. Overall, the findings of this study can be used as a knowledge basis in determining user interests, opinions, perspectives, and responsiveness, when it comes to cyber threats, attacks, and malicious activities. Also, the respective outcomes can contribute as a guide for identifying similar communities/subreddits and themes. Regarding the methodological contributions of this study, the proposed framework can be adapted to similar datasets and research goals as it does not depend on the special characteristics of the imported data, offering, in turn, a practical approach for future research.

Extended author information available on the last page of the article

Published online: 13 April 2024

Keywords Information security · Social networks · Reddit · Hierarchical clustering · Word clustering · Community mapping · Behavioral analysis

Introduction

Information security has become an important component of systems and applications in the digitalized era, which is characterized by rapid development. The emerging network and application resources brought to the surface capabilities and security issues that can benefit cyber-attackers. These circumstances lead to the discovery of new gaps and demands concerning information security. In general, information security is described as the procedures that protect information in all kinds of forms and systems [1]. One specific category of information security that attracts the interest of users and industries is defined as cyber security. Cyber security contains all the practices that aim to mitigate potential threats leveraging internet resources to gain access to an attacked system [2]. Cybersecurity is a complex area of interest and contains numerous perspectives associated with the following general categories [3–7]:

- Security Vulnerabilities: Software, hardware, or network weaknesses that create opportunities for unauthorized access and malicious behavior.
- Threats: A security threat is described as any occurrence that can harm a system or cause damage to an asset. Threats are usually associated with security vulnerabilities and may benefit an attacker to infiltrate a system. Note that this category is more related to application or system aftereffects than the actions of an attacker.
- Attacks: An attack is a set of procedures that exploit specific weaknesses to harm or disrupt a system. These actions are executed by malicious third parties aiming to gain access to the sensitive areas and main workflow of a system. The ultimate objective behind these actions is to achieve personal or collective goals related to personal files, codes, and rewards.
- Mitigations and Countermeasures: A mitigation can be defined as the overall effort and relevant actions taken to restrict the consequences and damages of successful attacks. A countermeasure is a procedure, tool, or change applied to a system or software to prevent a possible exploitation or attack.
- Exposures: An exposure is defined as a system or a software flaw that allows information leaks to a potential attacker and makes the target system more exploitable to him, as he, the attacker, can gather knowledge about existing vulnerabilities.
- Attack cost: Attack cost is related to the attacker's effort and expenses which are defined by his resources, experience/knowledge, and motivation.
- Incident: An incident can be described by the outcomes and impact of a completed/successful attack i.e. bank account exposures, file retrieval, etc.

Even though weaknesses that lead to successful attacks usually originate from existing liabilities that are related to system accessibility, complexity, and capacity, many cyberattacks require human interaction [8]. As a result, user knowledge is an important factor as the security of online activities has become more complex [9]. These circumstances along with the intriguing technological capabilities and the immediate online connectivity lead individuals, both experts and amateurs, to discuss the different information security perspectives in many available forms of social networks, e.g. social media [10], Q&A sites [11], forums [12].

In recent years, social networks have grown into multifarious data sources, covering a plethora of topics and concerns that are useful for social analysis. The increasing popularity of relevant platforms like Facebook, Twitter, Stack Overflow, and Reddit affects the overall user interactions and the variety of the themes being discussed. In contrast to developer Q&A websites, Reddit contains posts with high topic diversity by allowing users to establish topic-driven communities, called subreddits, that maintain their identity and unique purpose according to appropriate rules defined by their moderators [13]. A Reddit post comprises properties that provide information regarding the author, the relevant subreddit, and the main content of the post, i.e. title, description, and attached media components. At the same time, a Reddit post is associated with meta-characteristics that are relevant to user interactions including scoring systems, i.e. upvotes and downvotes, and comment sections [14]. Until now, the existing studies that analyze information from social networks are directed to a variety of experimental goals, where many of the aforementioned cyber security perspectives are explored in the process. These goals are associated with user awareness [10], vulnerability disclosure [15–17], exploitation [18–20], threat popularity [21], discussion growth [22], post popularity [23], and topic analysis [19].

Thus, motivated by the need to investigate interesting aspects of information security and social network discussions, in this study, we focus on addressing the main concepts that characterize user discussions from multiple subreddits that are relevant to information security. The main purpose of this study is to provide insights through data and text mining techniques that could assist both users and researchers in assessing user interests, opinions, responsiveness, and experiences as well as in identifying the representative subreddits per concept and the content similarities between these communities. To meet these objectives, we (a) employ a methodology for concept extraction, (b) establish a knowledge map that determines the potential relationships between concepts and subreddits and finally, (c) implement an approach assessing the relationships between concepts and user interactions. The proposed framework aims to provide answers to three Research Questions (RQs), presented in the section "Research questions", that are strongly associated with the purpose of this study. The first step towards this framework is to employ a word clustering approach based on the Affinity Propagation [24] algorithm to unveil the underlying concepts from the titles of Reddit posts. In addition, by leveraging the outcomes of the proposed approach, we further assess word distributions over concepts and subreddits. Moreover, the Shannon–Jensen Divergence [25] between all the extracted distributions is calculated, pair wisely, in order to establish the knowledge map through hierarchical clustering [26] and offer comprehensive information concerning their content relationships. In the last phase of the framework, the potential effects of the different concepts to user interactions are evaluated, with respect to two post metrics described as number of comments and upvote ratio. The outcomes lead to concepts with relatively high public acceptance and small or big discussion threads. The findings can be used as a guide for understanding what users discuss about information security-related issues, threats, incidents, jobs, skills, and scams. Also, the outcomes of this study can be studied as an overview describing the content-wise identity of the different subreddits/communities and the interaction profiles of the different investigated concepts/themes. Until now, to the best of our knowledge, we have not encountered a study that analyzes the concepts, interactions and hierarchy of information security-related subreddits from this perspective.

Overall, study contributes in:

- Helping researchers and platform users at identifying the various major and minor concepts discussed on information security subreddits (RQ_1)
- Offering content-based guidelines through concept and subreddit similarities that can be used for filtering queries and searches (RQ_2)
- Assisting individuals at distinguishing controversial or widely approved concepts (via upvote ratio) as well as determining concepts associated with frequent responsiveness, i.e. large numbers of comments, in the Reddit communities (RQ_3)

The rest of the paper comprises the Related Work (The section "Related work") and three Research Questions of this study (The section "Research questions"), while the methodology and results of the experiments are presented in the sections "Methodology" and "Results", respectively. Furthermore, in the section "Discussion" we discuss the main findings of the study and their potential usage while in the section "Threats to validity", the main threats affecting the validity of this study are evaluated. Finally, the section "Conclusions and future work" concludes the paper and proposes future directions.

Related work

This section provides an overview of the related literature on social networks and information security, focusing especially on Reddit data. The section "Research studies related to information security and social networks" discusses some common research practices associated with the different aspects of information security and the usage of information security-related data, discussions and posts from social networks. The section "Research directions and frameworks using Reddit data" provides information on multiple use cases relevant to the Reddit platform, discussing the various areas of interest, properties and methodologies that have been previously investigated to provide valuable insights and solutions.

Research studies related to information security and social networks

In recent years, the importance of cyber security has drawn the attention of researchers to develop tools for enhancing system security as well as providing knowledge regarding specific use cases and emerging threats. Three common cyber security threats that are frequently investigated in studies of this nature are related to denial of service, malware, and phishing attacks, and the reason of this increased interest is that these threats often appear in the real world and common computers, up until recently, hence causing severe damage to both users and systems [4]. The existence of threats of this nature in most systems dictates the demand to explore approaches for threat prioritizing and mitigation through the adoption of a wide variety of data-driven approaches including graph theory, linear programming, machine learning, and statistics [27]. To succeed at threat prioritizing and mitigation, researchers develop guidelines or tools concerning the identification, adaption, and assessment of security vulnerabilities and malicious activities, thus encircling a variety of topics [28]. The aforementioned proactive practices serve as countermeasures and mitigation techniques, which are usually integrated into the general categories of application and element hardening, activity detecting, isolation of system parts, deceiving practices, and eviction decisions [29].

The different practices and aspects of information security are nowadays attracting the interest of both domain experts and amateur individuals to discuss or post relevant events, ideas, and questions on different social networks. Through the integration of data from social network platforms, researchers have previously offered important information by studying user interactions with security issues/vulnerabilities [30] and by associating conversations with specific weaknesses [31, 32]. Regarding user awareness and system protection, prior studies make use of social network data to produce user alerts [10, 16, 33], assess exploitability indicators [15, 18], and evaluate threat evidence [19] using machine learning algorithms, statistical analysis, networks, and text mining approaches. Moreover, other studies aim to protect social network users and software applications as well as prevent potential incidents by detecting malicious behavior [34–36].

Shields et al. [37] address cyber security virality factors from information security-related Reddit posts to provide insights on the most important characteristics and timestamps that elevate post popularity. They reveal that the post content and time variables affect the virality of a discussion thread. Similarly, Horawalavithana et al. [30] show that both Twitter and Reddit can offer valuable information for the early detection of security threats, while it is also mentioned that the relevant Reddit posts have higher virality than tweets, which is a characteristic that leads to faster information spread. Regarding information security-related discussions and questions, Wu et al. [23] show that the Reddit platform contains more subjective/personal rather than technical content compared to developer Q&A sites, i.e. Stack Exchange, and suggest that Reddit communities are more appropriate for recommending career paths. Finally, recent trends in analyzing cybersecurity communities are related to vector spaces that capture the semantics of large corpora i.e., words and documents, with the main goal of offering flexible data structures for topic analysis and machine learning tasks [32, 38, 39].

Research directions and frameworks using Reddit data

Even though many social network data sources offer different alternatives for research, each source has its advantages and disadvantages. By comparing Reddit and Twitter, which are close in terms of popularity and usage, someone can discover several key differences related to information spread and interaction characteristics/capabilities [40]. Up until now, Reddit constituted an appropriate source for various objectives [41] focusing on two general perceptions associated with user interactions and post analysis [42]. One efficient extraction tool for retrieving the desired data from Reddit posts and communities is called the Pushshift social media data platform and offers real-time data collection for potential research [43].

Prior studies on Reddit data collect information by defining keywords against post titles or by simply exploring appropriate communities-subreddits. The common research purposes are associated with post-topic popularity [44–46] and controversiality [47], user behavior [48, 49], user networks [50], and conversation modeling [51, 52]. The available public venues provide knowledge for all kinds of discussions and areas of interest, leading researchers to focus on various directions like immigration [53], products [54], human habits [55] and treatment-health [56–58], social behavior [59–61] along with topics closer to computer science like personal data and user privacy [12]. In summary, according to Proferes et al. [62], many relevant studies focus on politics and mental health while the commonly utilized methodologies are often relative to computational and statistical procedures.

Moreover, several studies aimed at discovering the main underlying topics of Reddit posts covering a specific domain to offer insights according to user interests. Okon et al. [63], employed a framework for topic extraction based on topic modeling methods to extract knowledge from posts that are relevant to dermatology. In the process, the topic extraction framework helped identify trends, from seven subreddits, reflecting user interests and engagement, thus offering important information to both patients and experts on therapies and issues. Similarly, Ruan and Lv [64] employ topic modeling to assess the different perceptions of electric vehicles, hence revealing trends across the years and the sentiment on Reddit across the different perceptions and different subreddits. In general, researchers achieve success in addressing trends within Reddit data related to healthcare [65], privacy [66], parenting issues [67], programming [68], diseases [69], anesthesia [70], and other subjects by focusing on the discovery of patterns and topics in the textual information of Reddit posts, through approaches similar to those in the previously mentioned studies.

Research questions

In this section, we present the three RQs reflecting the main scope and objectives of this study. The scope of this study is to provide a multi-perspective view of user interactions that are associated with information security Reddit posts by assessing the concepts that characterize the investigated communities/subreddits and their potential effects to interaction metrics. The first objective is to assist individuals and organizations in understanding the interests of Reddit platform users through a word clustering approach. The second objective is to guide future researchers and platform users, via a knowledge map, into identifying communities according to their interests, acknowledging the conceptual interconnections between subreddits, and distinguishing the representative subreddits of each concept. The final objective is to provide an overview of the concepts with respect to user interactions and approval based on the number of comments and upvote ratio of the posts. Overall, the scope and objectives of this study lead to the following three RQs:

RQ_1: How can information security Reddit posts be thematically profiled?

RQ_1 focuses on extracting knowledge from the textual content of information security-related Reddit posts. Each post has a title that reflects its content/theme and affects the overall interactions of a subreddit's users. It should be mentioned that only the post titles were analyzed as a large proportion of the retrieved posts did not include "selftext", i.e. textual information containing additional details. To provide the appropriate outcomes, text mining and *Natural Language Processing* (NLP) techniques are employed to extract the main terms/words and concepts from these posts, hence offering an outline of user interests. In fact, a word clustering approach based on the Affinity Propagation [24] algorithm is implemented to uncover patterns of frequently co-occurring words encompassing distinct concepts. Afterwards, word distributions over concepts are also assessed to characterize their structure. By studying the top terms of each cluster/concept, we provide a representative description for each concept as well. According to Westrupp et al. [67], topic extraction techniques are commonly employed to extract patterns from Reddit data and contribute to the discovery of important knowledge related to the investigated area of interest. The general motivation behind RQ_1 is to offer an outline of information security-related discussions and establish a baseline that can contribute to the better understanding of concerns, interests, perceptions, facts, and opinions, which is a task that constitutes the primary purpose of existing studies [63, 65, 66]. The practical importance of the respective outcomes is the detection of the diverse areas of interest, which are directly or indirectly associated with information security, attracting and affecting users on daily basis like common types of threats, e.g. malwares, vulnerabilities, skills, and jobs, email and application scams.

RQ_2: What are the content relationships between concepts and subreddits?

RQ_2 extends RQ_1 by first evaluating the linkage strength between each subreddit and each word (included in post titles). To achieve this, word distributions over subreddits are first assessed in a similar manner to the word distributions over concepts that were discussed previously. By leveraging these two types of distributions, the similarities between the subreddits and concepts are discovered in order to offer additional guidelines that may assist users and researchers. As mentioned in the introductory part of this study, the Shannon–Jensen Divergence [25] is employed to calculate these similarities by comparing the word distributions over concepts and subreddits pair wisely. These similarities are first used to evaluate the prevalence of each concept in the examined subreddits. Thus, the identification of the representative subreddits per concept provides insights regarding the communities that a user should follow depending on his/her interests. One of the main goals and outcomes of this study are strongly associated to RQ_2, as the respective distributions are also utilized to establish a subreddit-concept knowledge map through hierarchical clustering. The objective of this map and RQ is to help researchers and platform users identify groups of information security-related communities and concepts matching their queries and scopes.

RQ$_3$: How do the different information security concepts affect user interactions on Reddit?

RQ$_3$ is dedicated to offering insights into the sentiment and responsiveness of users across the different information security-related concepts. To evaluate these two factors, the potential effects of the concept frequencies on the retrieved posts are studied by fitting two regression models for two relevant post metrics namely *upvote ratio* (continuous variable ranging from 0 to 1) and number of comments (*no comments* – count variable). In these models, the post metrics constitute the dependent variables while the frequencies of the concepts are denoted as independent variables. The first goal is to provide insights regarding the concepts that are associated with increased responsiveness on Reddit, in terms of no comments. The outcoming results will indicate the type of discussions in which the Reddit platform can be used as a social network to address related queries/questions. Similarly, by making use of the upvote ratio of the posts, we aim to unveil concepts that are associated with significant positive or negative sentiment. Thus, the latter analysis will help at distinguishing concepts that are linked with posts that either offer trustworthy information or lead to possible misleading and controversial discussions, as evaluated by user acceptance, agreement, and sentiment. By reviewing the outcomes and findings of the respective experiments, both researchers and platform users can acquire valuable knowledge related to the overall responsiveness and sentiment of the Reddit users in the investigated information security communities and related concepts. Overall, the main purpose of RQ$_3$ is to assist these target groups in accurately filtering both their queries and potential research on the Reddit platform through a knowledge basis describing the expected misinformation and user engagement across the different information security-related concepts.

Methodology

The main framework of this study consists of distinct phases and subphases that finally lead to the establishment of concepts, the construction of the subreddit-concept knowledge map and the assessment of the potential effects of the extracted concepts to two post interaction metrics. The first phase is entitled *Post Retrieval and Preprocessing* and contains all the necessary steps that are followed to collect data from information security subreddits. First, multiple subreddits that have been previously explored as information security-related communities were selected [22, 23, 37]. In addition, we manually searched and included subreddits that are relevant to cybersecurity careers and skills as well as to red[1] and blue[2] teams. At the same time, the subreddit list was also enriched with communities that are similar to the ones that were selected previously in the existing literature. In Appendix A, the full list of the 45 subreddits that were explored to form the dataset of this study is provided, where the additional communities are marked with boldface text. Moreover, in this phase, preprocessing techniques are applied to deduplicate replicated posts, i.e. posts that have the same title and belong to multiple subreddits.

The next phase is dedicated to RQ$_1$ and contains the text mining approaches that are employed for concept extraction and interpretation (*Concept Assessment*). First, text preprocessing techniques are applied to establish a Document Term Matrix (DTM), with raw term frequency weighting, where the post titles correspond to documents and their words correspond to terms. These techniques include lowercase transformation, stopword and punctuation removal as well as token stemming. To form the final DTM, rare words that occurred in less than 0.1% of the retrieved posts and words that carried insignificant content for the analysis, e.g. *also, ever, will, start, else, anyway*, were excluded as well. Thus, the final DTM is formed by 721 different stemmed words/terms. Next, the Inclusion Index (II) [71] between the words included in the DTM are calculated to assess the pairwise word similarities

$$II_{i,j} = \frac{P(i \cap j)}{min(P(i), P(j))}, for\, i = 1, 2, \ldots ..m\, and\, j = 1, 2, \ldots \ldots m \tag{1}$$

where m is equal to the number of words included in the DTM (721 words overall), $P(i \cap j)$ is the ratio of all the posts (in the dataset) that include both i and j and $P(i)$ is the ratio of the posts that include the word i. Overall, $II_{i,j}$ ranges from 0 to 1 and measures the probability of i occurring in a document given that j appears in it and $P(j) < P(i)$. As a result, frequent words that can represent rare words will carry high values of II in the analysis. Note that, the II matrix constitutes the input to the Affinity Propagation clustering algorithm.

Affinity Propagation is a clustering method that groups a set of observations, in our case words, with the main idea of message passing between these observations [24]. The objective of this method is to identify exemplars, based on similarity measures, which quantify as appropriate at representing a set of observations based on two measures described as *Responsibility* and *Availability*.

Responsibility $r(i,k)$ measures the fitness of an observation k serving as the exemplar of observation i against other candidate exemplars while *Availability* $a(i,k)$ measures the support from the data points where observation k is expected to be an exemplar. The availabilities are initialized as 0 while $r(i,k)$ is initialized and updated on each iteration with respect to the pairwise similarity between observation i and k denoted as $s(i,k)$:

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k} \left\{ a\left(i,k'\right) + s(i,k') \right\} \tag{2}$$

[1] https://en.wikipedia.org/wiki/Red_team.

[2] https://en.wikipedia.org/wiki/Blue_team_(computer_security).

Similarly, the pairwise availability is updated iteratively and gathers evidence from other observations on whether observation k is indeed a fitting exemplar, while at the same time, the self-availability is calculated differently:

$$a(i,k) \leftarrow min\left\{ 0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} max\left\{ 0, r(i',k \right\} \right\} \tag{3}$$

$$a(k,k) \leftarrow \sum_{i' s.t. i' \notin \{i,k\}} max\left\{ 0, r(i',k \right\} \tag{4}$$

Overall, for each observation i, the observation k that maximizes the sum of responsibility and availability $(r(i,k) + a(i,k))$ is defined as the exemplar of i, while i identifies as an exemplar when $i = k$. Finally, the algorithm may converge after a predefined number of iterations, when the "*changes in the messages follow below a threshold*" or when "*local decisions stay constant for some number of iterations*" [24].

The main reason that led to the implementation of this word clustering approach is the inability of topic modeling and document clustering algorithms to fit concepts from social media posts, as these posts are usually characterized as short and noisy [52]. In the process, we fit and present topic models to provide complementary information concerning the suitability and effectiveness of the employed approach, when it comes to topic/concept extraction, by highlighting similarities, benefits and potential drawbacks. Also, the reason behind the selection of the employed similarity measures and clustering algorithm, i.e. II and Affinity Propagation, is that the values of the II matrix will drive the Affinity Propagation method to identify word exemplars with a high frequency which can represent the rest words of the same cluster conceptually in the retrieved data. At the same time, rare words included in the DTM may be brought to the surface and form a unique cluster when frequently co-occurring pairs of words are apparent, hence offering information on concepts that probably represent a minority of the documents but still exist and concern the Reddit platform users. In addition, the meanings of the exemplars along with the most frequent words of each cluster will be evaluated to provide a representative description for each concept. Finally, according to similar indicators, each concept is categorized into at least one of the following categories:

- Attacks
- Mitigations/countermeasures/security
- Incidents/exposures
- Threats/vulnerabilities
- Labor/research
- Generic

Furthermore, in our approach, the weight coefficient of each word to a concept will be measured by the II between this word and the exemplar of the concept.

Thus, by evaluating these weight coefficients, in Eq. (5) we define the word distributions over concepts.

$$C_{i,j} = \frac{II_{i,j}}{\sum_{l=1}^{m} II_{i,l}}, for\, i = 1, 2, \ldots ..k\, and\, j = 1, 2, \ldots \ldots m \tag{5}$$

Similarly, to address the RQ$_2$, a probability distribution of words for each subreddit is also evaluated, similar to concepts, based on the DTM and the posts that belong to each subreddit,

$$S_{i,j} = \frac{\sum_{p \in p_i} DTM_{p,j}}{\sum_{l=1}^{m} \sum_{p \in p_i} DTM_{p,l}}, for\, i = 1, 2, \ldots ..s\, and\, j = 1, 2, \ldots \ldots m \tag{6}$$

where s is the total number of subreddits, p_i is the subset of titles that are relevant to the ith subreddit, and $S_{i,j}$ represents the frequency ratio of the jth word in the ith subreddit.

In the following step of this phase, the extracted word distributions over concepts, i.e. matrix C, and subreddits, i.e. matrix S, are used to assess their dissimilarities. As discussed previously, these dissimilarities are calculated using the Shannon–Jensen Divergence which measures the difference/divergence between two probability distributions. Furthermore, these evaluations are utilized to establish clusters of concepts and subreddits with similar distributions via Agglomerative Hierarchical Clustering [26].

The main goal of the proposed approach is to provide a hierarchy tree, where similar objects are clustered together at its lower levels while all objects are clustered together at the highest level. In this approach, each concept and subreddit starts as a singular object at the lowest level of the tree while its most similar objects can be identified at the lower levels of the tree. The knowledge extracted from these procedures leads to establishing the knowledge map, capturing similarity details, and detecting the representative subreddits per concept (*Subreddit-Concept Knowledge Map*).

In the final phase of the methodology, we make use of two post metrics, the *no comments* and *upvote ratio*, to assess the overall sentiment and interactivity of the users across the different concepts. First, each post is linked to the concepts via the Post to Concept Matrix (PCM) where $PCM_{i,j}$ denotes the number of words in the ith post belonging to the jth concept. Next, two regression models are fitted where each column of the PCM is declared as an independent variable while the two metrics are declared as the response/dependent variables, one for each model. The outcomes of these models are used to evaluate the significance and level of the potential effect of each concept to the two metrics. The ultimate purpose of this approach is to raise awareness to platform users about the information security concepts that are usually controversial or steady (*upvote ratio*) and about the ones that are associated with increased numbers of responses *(no comments)* across the Reddit platform. This purpose will be satisfied by distinguishing the most significant to each metric concepts. The flowchart of this study is presented in Fig. 1.

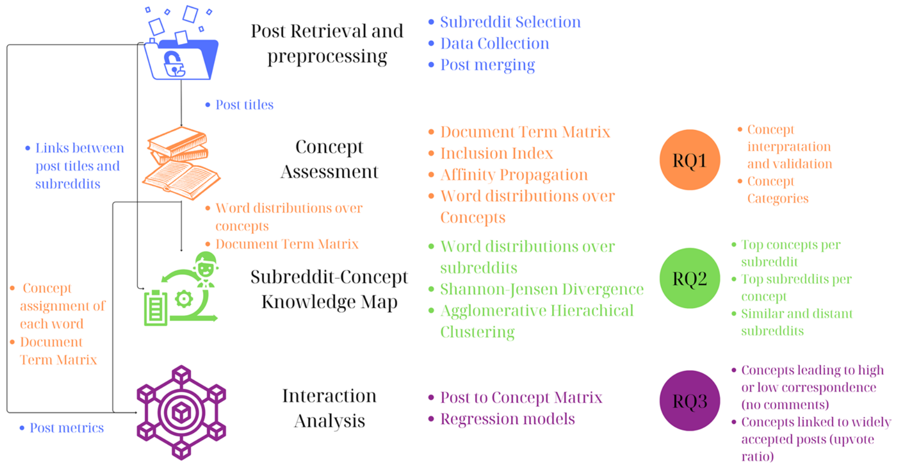
Post Retrieval and preprocessing
• Subreddit Selection
• Data Collection
• Post merging
• Post titles
• Links between post titles and subreddits
Concept Assessment
• Document Term Matrix
• Inclusion Index
• Affinity Propagation
• Word distributions over Concepts
RQ1
• Concept interpratation and validation
• Concept Categories
• Word distributions over concepts
• Document Term Matrix
Subreddit-Concept Knowledge Map
• Word distributions over subreddits
• Shannon-Jensen Divergence
• Agglomerative Hierachical Clustering
RQ2
• Top concepts per subreddit
• Top subreddits per concept
• Similar and distant subreddits
• Concept assignment of each word
• Document Term Matrix
Interaction Analysis
• Post to Concept Matrix
• Regression models
RQ3
• Concepts leading to high or low correspondence (no comments)
• Concepts linked to widely accepted posts (upvote ratio)
• Post metrics

Fig. 1 Flowchart of the study

Results

In this Section, we present the main results of this study. At this point, it should be noted that 62,091 posts from 45 subreddits were collected, covering the period from 2022-11-03 till 2023-03-07, with 53,458 unique post titles. The reason behind the exclusion of more recent posts is that the Reddit Pushshift Api has been unavailable since April 2023 for public use. The section "RQ1: How can information security Reddit posts be thematically profiled?" presents the main underlying concepts of the retrieved Reddit posts encompassing multiple themes that are directly or indirectly associated with information security. To complement the findings of the previous Section, in the section "Comparing concept extraction and topic modeling", 69 topic models are evaluated and the best model, with respect to topic coherence, is compared to the outcomes of the employed approach. Moreover, in the section "RQ2: What are the content relationships between concepts and subreddits?", we assess the main subreddits of each concept and present the subreddit-concept knowledge map which provides insights regarding the relationships between them. Finally, in the section "RQ3: How do the different information security concepts affect user interactions on Reddit?", the outputs of the regression models are interpreted while the similarities between the concepts that usually lead to relatively high/low *no comments* or *upvote ratio* are discussed as well.

Concept extraction

RQ$_1$: How can information security Reddit posts be thematically profiled?

The Affinity Propagation algorithm assessed overall 54 exemplars and clusters leading to an analogous number of distinct concepts. The respective outcomes show that Reddit posts concern a range of different perspectives related to information security

Table 1 Prevalence of categories

Category	Prevalence (%)	No concepts
Attacks	73	16
Generic	24	23
Incidents/Exposures	9	5
Labor/Research	7	10
Mitigations/Countermeasures/Security	21	11
Threats/Vulnerabilities	5	6

that were discussed in the introduction of this study while at the same time, many generic concepts related to applications, projects, careers, and skills are also apparent. In Appendix B (Table 3), the 54 concepts with their representative description and general categories/perspectives, in which each concept can be classified, are presented. In the same table, the overall occurrence of each exemplar in the documents, i.e. number of documents that the exemplar occurs at least once, is presented as well. The last column displays the number of words that declared each word as their exemplar (*No Words*). Note that the different words are stemmed, i.e. the suffixes of the exemplars are reduced to their word stem. Also, it is worth reminding that each concept may be classified into multiple categories which are separated by a semicolon in the table (";").

Both the descriptions and categories of each concept are extracted based on the core meaning of the exemplars and the most frequent words of each cluster, where the most frequently co-occurring words of each exemplar are considered as well. Moreover, Table 1 presents the frequencies of the general categories in the extracted concepts and the proportion of each category in the retrieved post titles. Note that, the aggregated values presented in the latter table sum up to more than 100 percent as a single concept may belong to multiple categories. This proportion is measured by the overall frequency of each word and the categories of their cluster. Overall, it is evident that most concepts can be classified as generic while the cyber-attacks (*Attacks*) are associated with more concepts in the investigated discussions compared to the rest information security-related categories. In addition, the table shows that cyber-attacks, is, indeed the most prevalent category in our data as 73% of the words in the posts fall into this category.

As the exemplars *scam* and *hack* represent a large proportion of the examined words, we can accept that users discuss scams and hacking in many of the retrieved posts while fewer posts are associated with the remaining concepts. Also, Table 1 shows that the generic concepts and the concepts that are related to enhancing security and research constitute the primary user interests, apart from cyber-attacks. Since cyber-attacks concern applications, accounts, and other components, the high frequency of generic concepts in the results should be highlighted as a reasonable finding as well. Also, the fact that users are frequently interested in mitigations, security, and countermeasures, means that Reddit users are interested in acquiring knowledge/skills or developing practices to enhance user and system security/

protection. Hence, the Reddit platform should be considered a valuable source for searching practical security solutions.

In summary, we observe multiple concepts that are relevant to an attacker's perception including security vulnerabilities and threats (e.g. concepts 24 and 29), cyber-attacks (e.g. concepts 1, 2, 13, 36, and 45), and incidents (e.g. concepts 32 and 46). On the other side, multiple concepts that are relevant to the enhancement of information security (e.g. concepts 9, 25, 28, 37) are also identified, meaning that users are highly interested in enhancing or learning about system and user security. At the same time, multiple concepts indicate that users are also interested in learning relevant skills, acquiring certificates, or working in related job positions and projects (e.g. concepts 8, 22, 31, 34, 35, 39). The remaining concepts concern generic themes including source codes (concept 15) and open source projects (concept 18), specific vendors and systems (concepts 10, 27, 49), applications (concept 7), web browsing (concept 33), storage systems (concept 51) and social media (concept 30) among others.

Moreover, Fig. 2 visualizes the extracted clusters in a more direct approach by showing the overall size of each cluster as indicated by the number of words represented by each exemplar. In this figure, the nodes of each cluster are presented with a unique color while the exemplars are also displayed along with their overall occurrence in the documents, i.e. number of documents that the exemplar occurs at least once. At this point it should be mentioned that the exemplar which is blurry in the figure is the word scam and occurs in 9373 documents overall.

In summary, it is evident that the retrieved Reddit posts are related to diverse concepts belonging to the general domain of information security and to other areas of interest that are indirectly associated with cyber-attacks, threats, and vulnerabilities. Overall, the retrieved post titles tend to combine two or more of the concepts and categories above. For example, multiple posts discuss threats or attacks that are relevant to websites and applications where the author either asks for advice or just reports an event. A representative post title linking websites with potential scams is the following: "Impersonator/Scammer redirecting his domain to my website". Another example linking two or more concepts is the following: "Scammed by Social Media Store Twitter account". This example is associated with two concepts and categories encompassing scams, i.e. concept 1, and social media, i.e. concept 30.

Comparing concept extraction and topic modeling

In this Section, we assess the fitness and validity of the proposed approach, by evaluating 69 topic models, which are trained using the Latent Dirichlet Allocation – LDA [72], and the employed approach under two metrics. The first metric, named Normalized Pointwise Mutual Information [73], measures the topic coherence of a model by assessing the co-occurrence strength between the top words of each topic while the second one measures the topic divergence of a model by counting the ratio of the unique words in the top words of the extracted topics [74]. Figure 3 presents the evaluations of the 69 LDA models in the range from 2 to 70 topics, where the

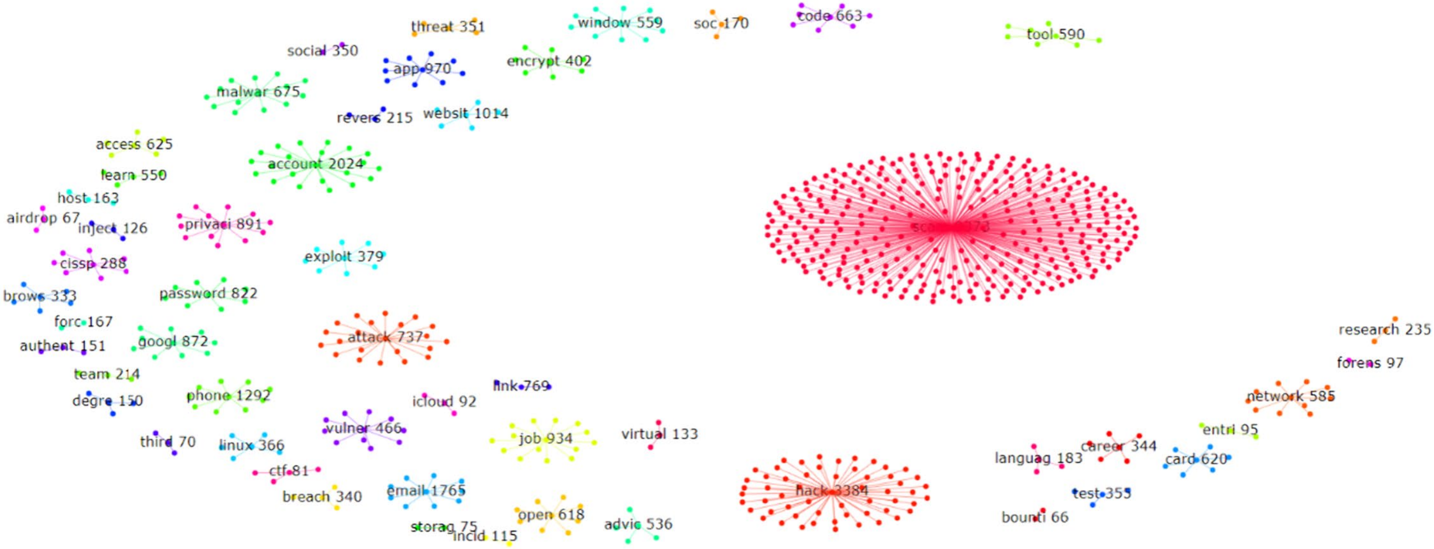

Fig. 2 Cluster visualization

number of top words was set to 3 due to the occurrence of significant reduction in topic coherence when selecting a larger number.

Overall, the optimal model, based on the topic coherence score, is the one trained for 47 topics (coherence close 0.233), while the topic divergence values show no significant improvement across the different numbers of topics. Based on the topic divergence metric, the two best models were trained for 2 and 5 topics while the rest models yielded descending values as the number of topics increased. At the same time, the employed approach was also evaluated using these metrics,, where its topic coherence evaluation was close to 0.313 while the respective topic divergence was above 0.87. Thus, when evaluating a small number of top words, the Affinity Propagation and the employed similarity metrics led to achieving the better performance for concept/topic extraction.

By thorough examining the LDA model for 47 topics, i.e. the model that maximized topic coherence, we believe it is worth summarizing the key points concerning the main differences and similarities between this model's outcomes and the outcomes of the proposed approach. Regarding the similarities, both algorithms identified the major topics/concepts of the posts including scams, hacking, vulnerabilities and exploits, jobs and careers, privacy, accounts, apps, encryption, passwords, phones, emails, attacks as well as links. However, several of the most frequent words of the dataset, e.g. *scam, hack, privaci*, were evaluated as the top words of multiple topics, having a relatively high proportion within a topic (usually more than 10%). In particular, the word *scam* was apparent in the top-3 words of 15 topics, hence creating overlaps. As discussed in the section "RQ1: How can information security Reddit posts be thematically profiled?", the word scam constitutes the exemplar for a significant number of words (322 overall), a number justifying the occurrence of overlaps in the LDA model. Due to this outcome, some minor concepts that were extracted using the Affinity Propagation algorithm are not apparent in the LDA models, including brute force, bounty hunting, ctf, forensics, red and blue teams etc.

These findings led to the conclusion that when it comes to uniqueness and divergence, the employed approach provided more unrelated and hidden concepts compared to the LDA model. In contrast, the LDA model captured topics that can be categorized to the same general theme, which is an issue concerning the occurrence of significant overlaps between the topics. The overlapping topics refer to types of scams (e.g. on Facebook, via text, money scams), or perspectives of hacking (e.g. learning to hack, hacking/hacked accounts). By summarizing, since LDA is considered a baseline method in extracting topics from textual information and several topics match the concepts extracted by the proposed approach, it is worth highlighting that the employed approach indeed captured text patterns characterizing a significant fraction of the analyzed posts.

RQ$_2$: What are the content relationships between concepts and subreddits?

In this section, we extend the findings of the section "Concept extraction" by exploiting the representative subreddits of each concept and further establishing the

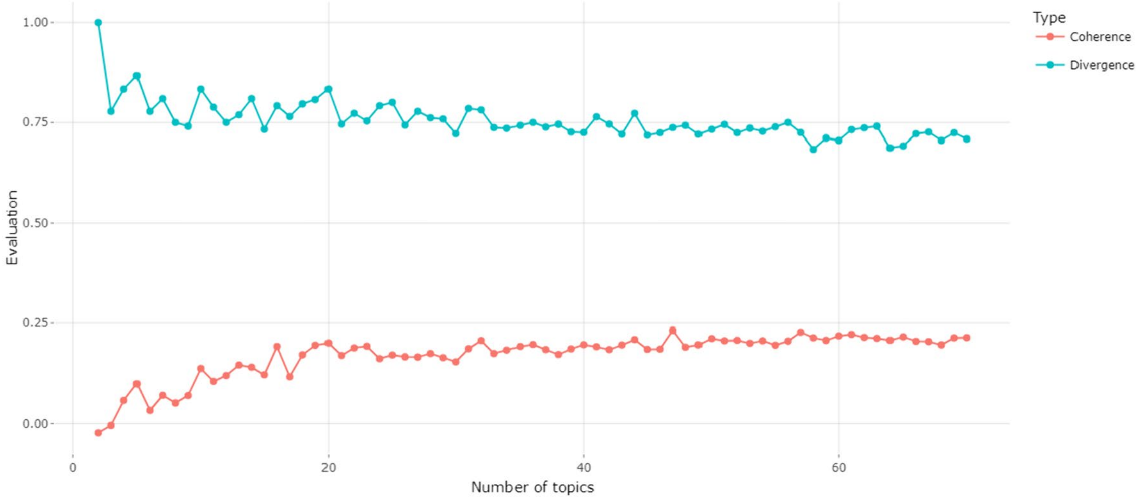

Fig. 3 Evaluation of topic models

subreddit-concept knowledge map. Initially, the pairwise divergence of the probability distributions is calculated, comparing 99 distributions for 54 concepts (matrix C) and 45 subreddits (matrix S). Then, the representative subreddits of each concept are defined as the subreddits with the lowest divergence, where values equal to 0 indicate identical distributions and values closer to 1 indicate a high degree of dissimilarity. Appendix B (Table 4) presents the top 3 subreddits of each concept along with their relative divergence, as calculated by the Shannon–Jensen Divergence.

By examining the descriptions of the different concepts and their top subreddits, it is observed that there is a conceptual association between them in multiple cases. These cases include concept 1 (i.e. Scams), concept 2 (i.e. Hacking), concept 8 (i.e. Jobs), concept 9 (i.e. Privacy and policies), concept 14 (i.e. Malwares), and concept 34 (i.e. Cissp), meaning that the employed approach captured several strongly associated pairs of concepts and subreddits. Furthermore, the subreddit-concept knowledge map is created through hierarchical clustering and based on the pairwise dissimilarities as calculated by the Shannon–Jensen divergence. In this approach, the *second Ward's method* proposed by Murtagh and Legendre [75] is employed to form the hierarchy tree. Figure 4 presents the complete hierarchy between the examined subreddits and concepts.

The subreddit-concept knowledge map extends the findings of the previous section by providing additional information regarding the hierarchy of these objects and by addressing their interconnections in the process. In more detail, we detect that concepts and subreddits that refer to similar content are clustered together in the lower levels of the hierarchy tree. A representative example associated with this finding is the relationship among Concept 14 (Malwares), r/Malwarebytes, r/computerviruses, and r/Malware which are clustered early at the lower layers of the tree, and it is obvious that they represent similar content, i.e. malwares and viruses. Similar cases include the early clustering of Concepts 8, 31, 34, 39, 43, and 48 with subreddits that are relevant to career advice, as well as the early clustering of r/Scams and Concept 1 (Scams) with relevant concepts or objects/components that are indirectly associated with scams, i.e. accounts, emails, websites, cards and links. In summary, the proposed approach captures several of the expected similarities between concepts and subreddits in multiple cases, offering, in turn, valuable information for future research and queries as well as for user interactions and engagement.

RQ$_3$: How do the different information security concepts affect user interactions on Reddit?

In this Section, we provide the outcomes of the trained regression models and discuss the main findings connecting each concept to the two investigated interaction metrics (*no comments*, *upvote ratio*). At this point, it is worth mentioning that different regression models were selected according to the specificities of these metrics which constitute the response variables in the experiments.

By investigating the *no comments* of each post, which constitute count values, a significant overdispersion was detected, since a large proportion (close to 27%) of the posts did not have a comment. As a result, the zero-inflated Poisson regression

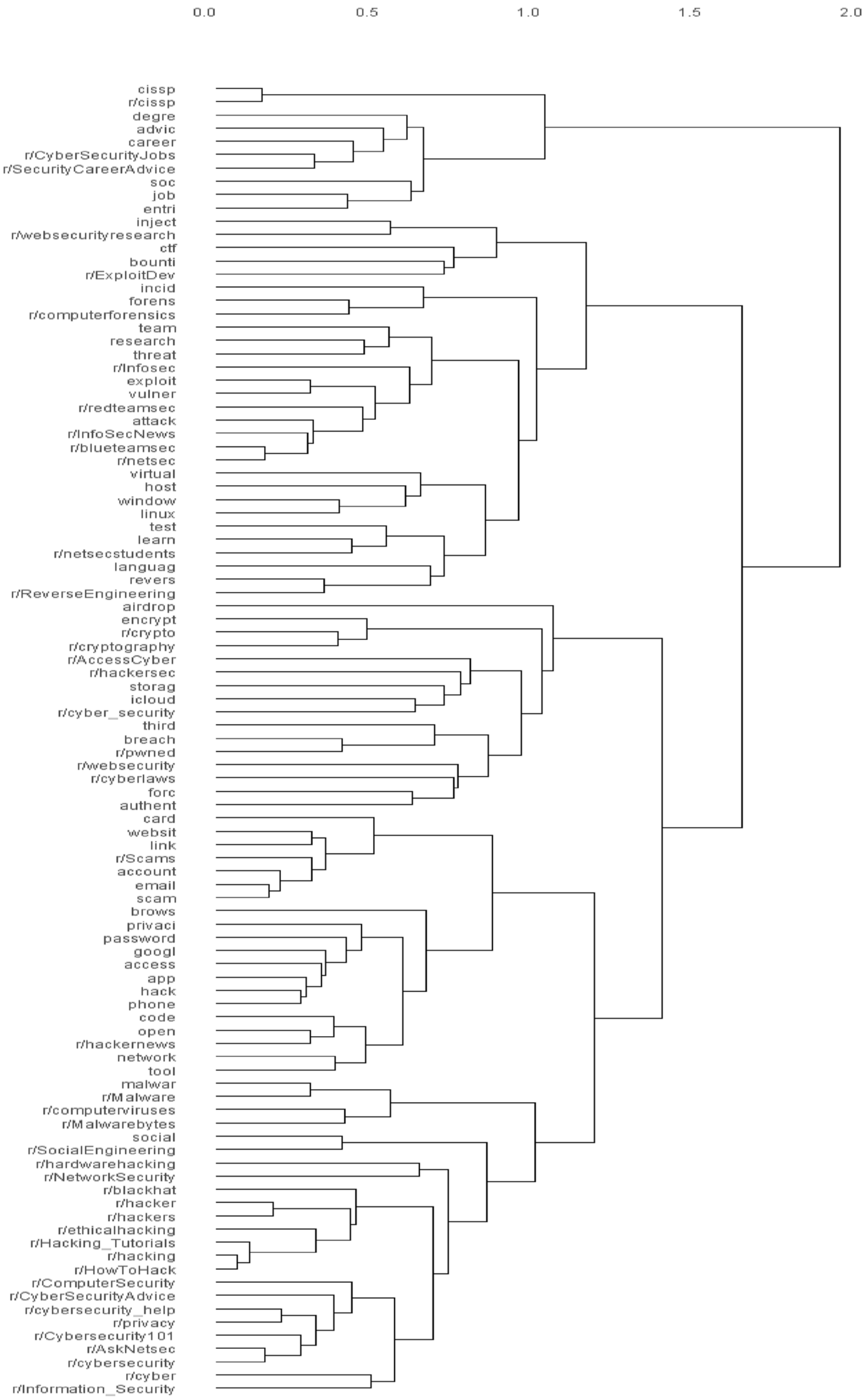

Fig. 4 Subreddit-Concept Knowledge Map

[76] is employed which overcomes this issue by fitting two models. The first one models the probability of observing a specific non-zero count (first inflation model), for every possible non-zero count, and the second one models the probability of observing a zero count (second inflation model). Regarding the *upvote ratio* of the posts, which constitute proportional values, a binomial regression model [77] is fitted and in this case, no significant overdispersion was detected. Thus, by extracting the results of the two models, a 2d map presenting the coefficients and significance of each concept is established (Fig. 5). Note that, this map presents the coefficients of the first inflation model for the *no comments* of the posts, referring to non-zero counts (y axis).

Based on the above figure, it is detected that most concepts affect the *no comments* of the posts, and that concepts *airdrop* and *language* should be considered as the concepts that lead to relatively few comments. Apart from this observation, there is not a concrete subset of concepts with high coefficients as a large proportion of the concepts were declared significant (50/54).

For this reason, the model that was fitted for the zero counts of *no comments* (second inflation model) was also inspected and led to determining that apart from *airdrop*, there are other concepts that are usually associated to posts with zero comments as well. These concepts are directly or indirectly linked to cyber threats, vulnerabilities and exploits (i.e. concepts *threat, vulner, exploit*) or to practices aiming at identifying these threats in a system (i.e. concepts *bounti* and *ctf*). In addition, the concepts *attack, malwar* and *inject* were also evaluated with similar attributes compared to the rest concepts.

On the other side, the model indicated that the posts belonging to the general category of jobs/careers, including the concept *advice*, find increased responsiveness by the information security communities in terms of comments. Similarly, concepts that are directly or indirectly related to privacy (*privaci*, *googl, password*), where the indirect relations are identified through Fig. 4, are linked with increased *no comments* or zero comments. Also, some major concepts including *scam, phone, social, encrypt* and *authent* usually receive at least one comment or lead to a relatively high *no comments*.

Regarding the second metric, it is observed that multiple concepts, i.e. *threat, vulner, exploit, malwar, attack,* that were associated with relatively few *no comments* are also linked to high values of *upvote ratio*. Overall, the binomial model indicated that only a few concepts (14 overall), compared to the previous regression models, should be considered as significant while at the same time, not a single concept was evaluated with a significant negative sentiment.

The practical meaning of the outcomes of this Section is that there are differences between the investigated concepts concerning their approval and controversy, expressed by voting/scoring systems (*upvote ratio*). This finding leads us to determine concepts finding agreement by the Reddit users. Similarly, the potential interactivity (*no comments*) of the concepts varied among the different categories, meaning that some concepts refer and contribute as news/announcements (lower values of *no comments*) while others potentially constitute discussion threads, queries and questions/answers (higher values of *no comments*). By inspecting the fitted regression models, we can conclude that the posts referring to Certified Information

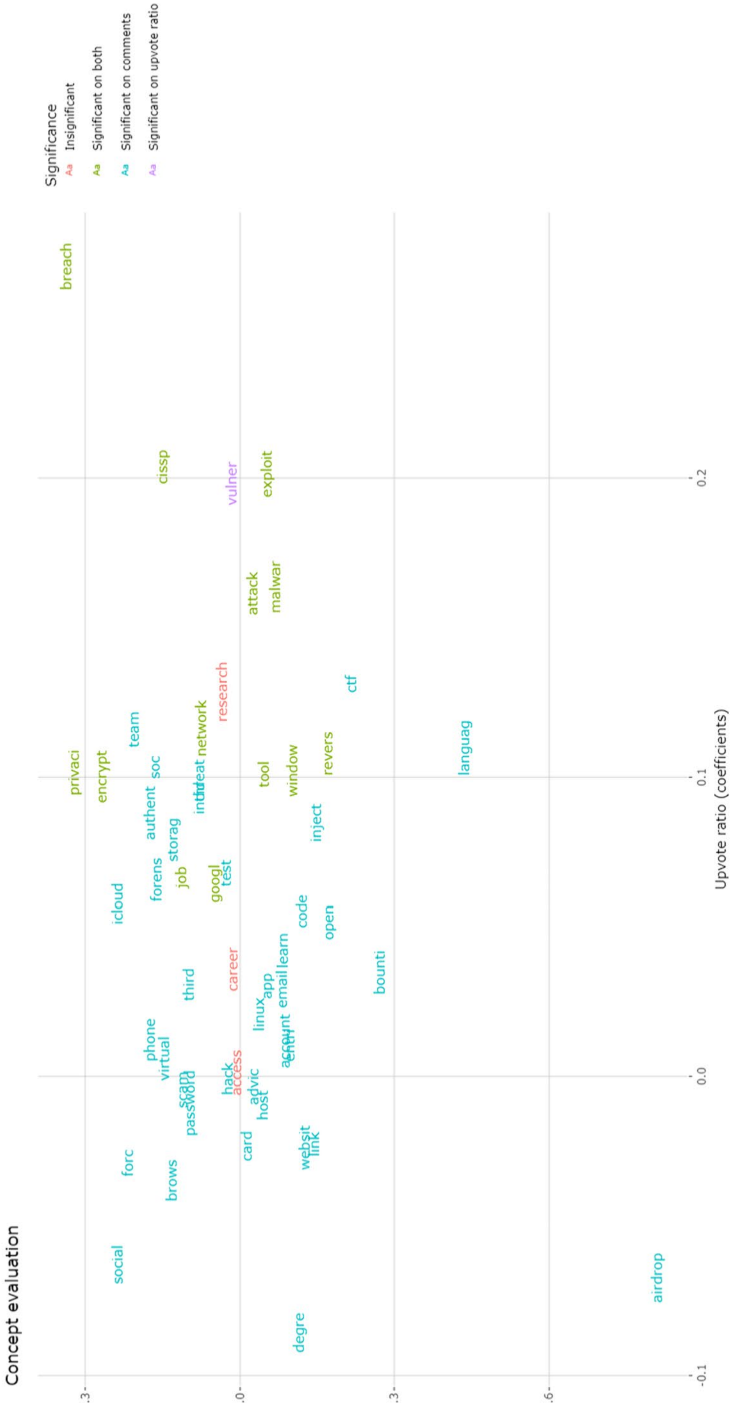

Fig. 5 Regression coefficients of concepts

Systems Security Professional (*cissp*) and security breaches (*breach*) should be considered as the most informative/interactive (*no comments*) and widely approved/liked by the Reddit users (*upvote ratio*). It should be also mentioned that *breach* may also lead to zero *no comments* according to the zero values of the *no comments* (second inflation model). To sum up, the key contributions presented in this Section are the following:

- Through the presented map, researchers and users can identify concepts driving trustworthy or misleading posts/discussions based on user opinions/sentiment

  The investigated subreddits can be used as a trustworthy news stream for vulnerabilities and exploits as well as for cyber-attack and malware related announcements

- Users seeking advice or requiring answers can review the above figure to determine whether the investigated communities would correspond to their queries

  Queries related to Jobs/careers obtain significant responsiveness

Discussion

Our analysis showed that the Reddit platform can be studied as a valuable source for exploring numerous and diverse areas of interest that are associated with user interests, concerns, opinions, and experiences. In summary, we believe that both platform users and researchers may review the outcomes of this study to understand the prevalence of each of the different extracted themes and categories as well as acknowledge specific concepts that were previously unknown to them. Also, through the presented findings, individuals who intend to discover new areas of interest or study unique concepts can find associated discussions and content in specific communities to cover their queries and needs. As shown previously, objects with similar content are clustered early in the established hierarchy tree, indicating that groups of concepts and subreddits can be studied together instead of exploring one concept or community at a time. As a large proportion of the posts were related to cyber-attacks, the Reddit platform can be seen as a source to enhance user awareness by identifying and learning about malicious activities, e.g. scams, and viruses. In addition, security practitioners and analysts can study the outcomes of this study to distinguish categories of threats, vulnerabilities, and attacks, which can be viewed as complementary material for improving system and application security. Also, individuals exploring alternatives for developing appropriate frameworks and techniques associated with countermeasures and mitigations can benefit from the outcomes of this study in a similar manner (Category Mitigations/Countermeasures/Security from the section "Concept extraction"). In addition, since several concepts that are relevant to security careers and tools (Labor/Research) were uncovered, the same groups of individuals may use the extracted knowledge map, or the results presented in the section "Concept extraction", to search the Reddit platform for new security tools, ideas and techniques as well as identify related courses, jobs and skills. When it comes to user interactions and sentiment across the different concepts, researchers and platform users can review the outcomes of this study to identify concepts that are usually associated with news/announcements or discussion threads and queries as well as determine concepts offering non-controversial information which usually contributes as factual knowledge.

Overall, the employed approach provided details rather than a generic outline of user interests by identifying both frequent and infrequent concepts, which were also cross validated using topic models. Notably, multiple rare words were declared as exemplars and helped form unique concepts that are relevant to information security, including *revers* (Reverse engineering), *breach* (Security breaches), *threat* (Cyber threats), *forens* (Digital Forensics), and more. At the same time, frequent words like *scam*, *hack,* and *privaci* were also identified as cluster exemplars through the proposed approach. In summary, an important outcome of this study is that the extracted concepts match multiple information security-related themes and categories, hence offering meaningful insights regarding user interactions (RQ_1).

Further, the subreddit-concept knowledge map offered a structure assessing the similarities between the different subreddits and concepts (RQ_2). The analysis pointed out multiple communities that are more directly to the category of cyber-attacks but at the same time, subreddits that are related to the rest categories were distinguished as well. With more detail, it was observed that r/Malware, r/computerviruses, and r/Malwarebytes are the subreddits in which users discuss about malwares and viruses (concept 14). At the same time, it was detected that security careers, jobs, and certifications are mostly discussed in r/CyberSecurityJobs and r/SecurityCareerAdvice. Similar clusters include security vulnerabilities, exploits, and threats, which are mostly associated with r/Infosec, r/netsec, r/readteamsec, and r/blueteamsec, as well as communities that mostly contain posts relevant to hacking. Furthermore, based on the main title of the subreddits, multiple pairs that were expected to be clustered together early in the hierarchy tree were also identified, e.g. r/cissp and *cissp*, r/Social Engineering and *social*, r/Scams and *scam*. In conclusion, the different hierarchy levels of the subreddit-concept knowledge map can be investigated to distinguish clusters of subreddits and concepts that are associated with specific or generic themes, depending on the selected level/layer.

Regarding the employed regression models, the outcomes indicated that posts related to cyber-attacks and security threats mostly contribute to raising awareness as news feed, having a wide acceptance by the users of the investigated communities. Also, according to the *no comments* of the posts, Reddit should be considered a social network where users can seek responsiveness or review questions/answers for security jobs, skills, and careers (especially for CISSP). Overall, groups of concepts referring to the same category are characterized with similar interaction profiles, meaning that users engage related concepts with similar sentiment and responsiveness.

As for the practical implications of the study, the findings of our approach can be studied as a basis for different purposes. Through the respective analysis, it was detected that the extracted concepts (RQ_1), knowledge map (RQ_2) and interaction profiles (RQ_3) address the interconnections and interaction effects of subreddits and concepts which can be studied for eight main purposes that are associated with the main objectives and RQs of this study:

- Recognizing user interests, concerns, and experiences;
- Identify concepts characterizing a subreddit;
- Assess representative subreddits/communities for each concept;
- Find similar subreddits to a target subreddit encompassing likewise content;
- Distinguish highly associated or distant concepts;
- Determine concepts that are usually related to news/announcements
- Classify controversial and publicly approved post types
- Detect question types that acquire responsiveness on information security-related subreddits

Finally, we believe that the proposed approach can be adapted to other similar tasks involving textual information and unique entities. The flexibility of this approach is indicated by the absence of dependencies or restrictions on the investigated domain and characteristics of the explored data source. The only requirement concerning the properties of the imported dataset is to contain two variables, one denoting the textual information of an observation and one denoting its related entities. By keeping this structure in mind, future researchers can investigate the underlying concepts and similarities of a set of books, Twitter accounts, Q&A communities, authors, etc.

Threats to validity

Although the employed approach and experiments are based on existing methodologies and a data source that is investigated for similar purposes, it is important to address potential threats to the validity of this study. Thus, this Section discusses internal and external threats, where these threats are related to the vital stages of our approach including data collection, data preprocessing, word clustering, algorithm selection, assessment methods, and outcome interpretation.

Regarding the internal validity of the approach, the are some steps that were manually handled, hence raising bias towards the proposed approach. First, the selected/applied text preprocessing techniques affect the whole of the methodology and findings accordingly. To address this concern, a set of text preprocessing techniques which are commonly utilized in research were selected. In addition, the implemented methodology for concept extraction is based on a clustering algorithm while other alternatives are considered in similar experiments as well, i.e. topic modeling, and document clustering. However, the analysis showed that the employed approach extracted multiple concepts belonging to the general domain of information security while at the same time, prior research supports the fact that the aforementioned standard methods are not always appropriate for these types of data [52]. Further, the employed performance metrics were not evaluated based on a ground truth indicator. To mitigate this threat, the choice concerning these metrics was based on existing and frequently applied methods for measuring the semantic similarity between words and the divergence between probability distributions. Finally, the extracted concepts were manually interpreted and described, thus raising bias again. To minimize the impact of this concern, the majority of the concepts were examined and matched against existing themes relevant to both information security, e.g. threats, attacks, vulnerabilities, and to more generic contexts, e.g. applications, skills, jobs, social media, etc.

The threats to external validity involve all potential issues affecting the generalization of the findings and the proposed method. Apart from Reddit, there are several sources where both expert and amateur individuals pose their questions, concerns, and queries, i.e. Stackoverflow, and Twitter. To enhance the validity of the outcomes and decrease the significance of this risk, we decided to base the analysis on multiple subreddits which are explored in similar studies as well [22, 23, 37]. Moreover, since the implemented framework contains newly proposed steps, without the direct validation of the respective outcomes with ground truth indicators, the generalization of the findings and the effectiveness of the proposed framework in future applications/research is currently not investigated. Knowing the significance of likewise issues, we employed solid algorithms and metrics that have been extensively used on generic data mining tasks. In addition, the outcomes of the study were matched against multiple information security terms while the extracted knowledge map contained reasonable information, e.g. the subreddit r/scams was strongly associated with the concept that was relevant to scams.

Conclusions and future work

In conclusion, this study presented a novel framework aiming at extracting the main underlying concepts of information security Reddit posts, characterizing their interaction profiles, and mapping these concepts and different communities, i.e. subreddits, encompassing posts of this nature. This framework utilizes a word clustering approach, an approach based on the Affinity Propagation algorithm to discover these concepts, assesses the interconnections between the aforementioned objects, i.e. concepts and subreddits, via hierarchical clustering, and finally addresses the potential effects of the concepts to the interaction metrics of the posts through regression models. The respective findings indicate that the users of the investigated subreddits are interested in many diverging themes which are categorized into multiple information security concepts and categories, e.g. cyber-attacks, scams, threats, vulnerabilities, incidents, exposures countermeasures, skills, and jobs. Moreover, the extracted subreddit-concept knowledge map helped provide details concerning the conceptual similarities/dissimilarities between subreddits and concepts, thus offering insights about the representative subreddits of the different concepts and the subreddits that share similar content. Finally, the regression models unveiled types of concepts that usually provide non-controversial information which is approved by the platform users while at the same time, pointed out types of discussions that work as either news/announcements or discussion threads.

Overall, we believe that the proposed framework can be applied to different content, data sources, and tasks as the employed algorithms and data inputs can be easily replaced or adapted depending on the goals of potential future research. For example, this framework can be used to discover and assess the relationships between researchers and themes in a specific domain. Regarding the practical usage of the main outcomes, the findings of this study can be used as a comprehensive guide in identifying user interests, assessing the main content of a subreddit, discovering similar or distant information security communities/subreddits as well as detecting concepts that lead to high responsiveness and trustworthy information.

In future research, the similarities between subreddits and concepts as well as the identification of concepts can be investigated using different algorithms and metrics. Topic modeling, document clustering, and co-word analysis are some of the approaches which constitute the more standard approaches for this task, as they provide sufficient results on many occasions. Also, the identification of user interests and themes can be investigated in multiple platforms where both expert and amateur individuals can interact with each other through information security-related discussions, queries, and events. Thus, the outcomes of this study can be cross-validated with the posted events, discussions, and questions from other platforms or contribute to assessing the diversity between two or more communities/platforms, including the Reddit platform. Regarding the interaction metrics of the Reddit posts, it is worth noting that some alternative approaches measuring the effects could also provide sufficient knowledge, in a similar manner to the employed framework, including machine learning models, e.g. Random Forest, correlation analysis and hypothesis testing.

When it comes to the communities of the Reddit platform, the different subreddits also contain information about their users. These properties can provide valuable information concerning post popularity and engagement while the subreddit properties may be evaluated to assess the similarities/dissimilarities between two subreddits, hence offering an alternative approach in clustering communities. Undoubtedly, studying user interests is a multi-perspective task as they can interact in diverse communities. Thus, by investigating the co-occurring interests of users interacting in information security communities, future research can unveil other domains that are associated with information security, e.g. computer science, social media, and applications.

Appendix A: Selected subreddits

See Table 2

Appendix B: Content based details and relevant communities of concepts

See Table 3

See Table 4

Table 2 List of the selected information security subreddits

r/AccessCyber	r/AskNetsec	r/blackhat	**r/blueteamsec**	**r/cissp**
r/computerforensics	r/ComputerSecurity	r/computerviruses	r/crypto	**r/cryptography**
r/cyber	r/cyber_security	r/cyberlaws	r/cybersecurity	**r/cybersecurity_help**
r/Cybersecurity101	**r/CyberSecurityAdvice**	**r/CyberSecurityJobs**	r/ethicalhacking	r/ExploitDev
r/hacker	**r/hackernews**	**r/hackers**	**r/hackersec**	r/hacking
r/Hacking_Tutorials	r/hardwarehacking	r/HowToHack	**r/Information_Security**	r/Infosec
r/InfoSecNews	r/Malware	r/Malwarebytes	r/netsec	r/netsecstudents
r/NetworkSecurity	r/privacy	r/pwned	**r/redteamsec**	r/ReverseEngineering
r/Scams	**r/SecurityCareerAdvice**	r/SocialEngineering	**r/websecurity**	r/websecurityresearch

*Boldface text indicates a community that was selected by the authors of this study with no external suggestions

Table 3 Main information of each concept

No.	Exemplar	Description	Category	Exemplar documents	No Words
1	scam	Scams in general	Attacks	9373	322
2	hack	Hacking (both malicious and ethical)	Attacks;Mitigations/Countermeasures/Security	3384	57
3	account	Account compromises and protection	Attacks;Mitigations/Countermeasures/Security; Incidents/Exposures; Generic	2024	22
4	email	Email scams, spams, spoofing	Generic;Attacks;Incidents/Exposures	1765	10
5	phone	Phone related discussions (security, compromises etc.)	Generic	1292	10
6	websit	Security and cyber attacks related to websites	Generic; Attacks	1014	6
7	app	Issues, scams and discussions related to web and mobile applications	Attacks;Threats/Vulnerabilities	970	10
8	job	Job related questions (requirements, skills, interviews, resumes etc.)	Labor/Research	934	15
9	privaci	Privacy, protection and policies	Mitigations/Countermeasures/Security	891	13
10	googl	Google related security discussions	Generic	872	10
11	password	Password generation, management and crack	Mitigations/Countermeasures/Security; Attacks;Generic	822	9
12	link	Questions and discussions related to malicious links and sharing links	Threats/Vulnerabilities	769	3
13	attack	Cyber attacks and attack prevention	Attacks;Mitigations/Countermeasures/Security	737	22
14	malwar	Malwares	Attacks	675	14
15	code	Source code related discussions (Malicious codes/objects, remote code execution, github source codes, code verification)	Generic	663	8
16	access	Access control and gain	Generic	625	6
17	card	Credit, gift and debit card scams and security	Generic;Attacks;Incidents/Exposures	620	7
18	open	Open source objects (projects, code etc.)	Labor/Research	618	9
19	tool	Developing and using tools (malicious, security, educational tools etc.)	Generic	590	8

Table 3 (continued)

No.	Exemplar	Description	Category	Exemplar documents	No Words
20	network	Networks in general	Generic	585	11
21	window	Posts related to Windows security	Generic	559	12
22	learn	Learning and acquiring skills	Labor/Research	550	4
23	advic	Seeking advices in general (protection/preventing advices, career advices, legal advices)	Generic	536	5
24	vulner	Security vulnerabilities	Threats/Vulnerabilities	466	11
25	encrypt	Encryption	Mitigations/Countermeasures/Security	402	8
26	exploit	Security exploits	Attacks	379	9
27	linux	Linux related discussions	Generic	366	6
28	test	Testing operations and penetration testing	Mitigations/Countermeasures/Security	355	4
29	threat	Cyber threats and threat intelligence	Threats/Vulnerabilities;Mitigations/Countermeasures/Security	351	5
30	social	Social media related discussions, scams, engineering etc	Generic;Attacks	350	2
31	career	Security careers	Labor/Research	344	6
32	breach	Security breaches	Incidents/Exposures	340	4
33	brows	Web browsers and online browsing	Generic	333	6
34	cissp	Security-related exams, studies and classes, especially related to Certified Information Systems Security Professional	Labor/Research	288	8
35	research	Cyber security research (papers, projects, labs etc.)	Labor/Research	235	3
36	revers	Reverse engineering, action reversal	Attacks	215	3
37	team	Cyber security and research teams (Red and Blue teams)	Mitigations/Countermeasures/Security	214	3
38	languag	Programming languages, Language models, Modeling and programming in general	Generic	183	4
39	soc	Security analysts and Security Operations Centers	Labor/Research	170	4

Table 3 (continued)

No.	Exemplar	Description	Category	Exemplar documents	No Words
40	forc	Brute force attacks and forced actions	Attacks	167	2
41	host	Hosting and servers	Generic	163	3
42	authent	Authentication and security bypass	Mitigations/Countermeasures/Security;Threats/Vulnerabilities	151	3
43	degree	Educational degrees	Labor/Research	150	4
44	virtual	Virtual machines, machines in general (computers, industrial machines etc.) and private assets (networks, accounts, keys, messages, browsers etc.)	Generic	133	3
45	inject	Injection attacks (especially sql injection)	Attacks	126	3
46	incid	Security incidents and incident response	Incidents/Exposures	115	2
47	forens	Digital assets (identities, twins etc.) and digital forensics	Generic	97	2
48	entri	Certifications and entry level jobs	Labor/Research	95	3
49	icloud	Apple iphone and icloud	Generic	92	3
50	ctf	Capture the flag challenges and events	Labor/Research	81	4
51	storag	Storage and cloud systems	Generic	75	2
52	third	Third party interactions and security risks	Attacks;Threats/Vulnerabilities	70	3
53	airdrop	Token airdrops (usually from cryptocurrency) and application tokens	Generic	67	3
54	bounti	Bug bounties and bug/bounty hunters	Mitigations/Countermeasures/Security	66	2

Table 4 Top subreddits of each concept

No.	Exemplar	Subreddit 1	Subreddit 2	Subreddit 3
1	scam	r/Scams 0.22	r/hackernews 0.38	r/hacking 0.39
2	hack	r/hacking 0.19	r/cybersecurity 0.24	r/HowToHack 0.25
3	account	r/hacking 0.27	r/Scams 0.3	r/hackers 0.32
4	email	r/Scams 0.27	r/hacking 0.33	r/cybersecurity_help 0.34
5	phone	r/privacy 0.28	r/hacking 0.3	r/cybersecurity_help 0.32
6	websit	r/hacking 0.33	r/cybersecurity 0.35	r/Scams 0.35
7	app	r/privacy 0.26	r/hacking 0.31	r/hackernews 0.32
8	job	r/CyberSecurityJobs 0.34	r/cybersecurity 0.36	r/SecurityCareerAdvice 0.45
9	privaci	r/privacy 0.24	r/hackernews 0.33	r/cybersecurity 0.39
10	googl	r/privacy 0.29	r/hackernews 0.32	r/hacking 0.35
11	password	r/hacking 0.32	r/HowToHack 0.35	r/cybersecurity_help 0.36
12	link	r/Scams 0.33	r/cybersecurity_help 0.35	r/hacking 0.36
13	attack	r/blueteamsec 0.22	r/InfoSecNews 0.29	r/cybersecurity 0.31
14	malwar	r/Malware 0.29	r/blueteamsec 0.3	r/netsec 0.35
15	code	r/netsec 0.32	r/hackernews 0.32	r/cybersecurity 0.35
16	access	r/hacking 0.31	r/privacy 0.32	r/cybersecurity 0.32
17	card	r/Scams 0.41	r/privacy 0.5	r/hacking 0.5
18	open	r/hackernews 0.29	r/cybersecurity 0.34	r/netsec 0.35
19	tool	r/netsec 0.29	r/blueteamsec 0.31	r/cybersecurity 0.32
20	network	r/cybersecurity 0.32	r/blueteamsec 0.36	r/AskNetsec 0.37
21	window	r/blueteamsec 0.35	r/netsec 0.37	r/cybersecurity 0.41
22	learn	r/cybersecurity 0.37	r/Hacking_Tutorials 0.37	r/HowToHack 0.39
23	advic	r/cybersecurity 0.39	r/CyberSecurityAdvice 0.46	r/CyberSecurityJobs 0.48
24	vulner	r/netsec 0.28	r/blueteamsec 0.34	r/InfoSecNews 0.4
25	encrypt	r/cryptography 0.41	r/privacy 0.42	r/hackernews 0.44
26	exploit	r/netsec 0.32	r/blueteamsec 0.35	r/InfoSecNews 0.37
27	linux	r/netsec 0.45	r/Hacking_Tutorials 0.48	r/blueteamsec 0.5
28	test	r/cybersecurity 0.42	r/AskNetsec 0.43	r/netsec 0.46
29	threat	r/blueteamsec 0.31	r/cybersecurity 0.4	r/netsec 0.43
30	social	r/SocialEngineering 0.39	r/hacking 0.41	r/privacy 0.44
31	career	r/SecurityCareerAdvice 0.39	r/CyberSecurityJobs 0.41	r/cybersecurity 0.5
32	breach	r/InfoSecNews 0.38	r/pwned 0.39	r/cybersecurity 0.43
33	brows	r/privacy 0.42	r/cybersecurity_help 0.49	r/hackernews 0.5
34	cissp	r/cissp 0.14	r/SecurityCareerAdvice 0.62	r/cybersecurity 0.63
35	research	r/cybersecurity 0.36	r/netsec 0.37	r/blueteamsec 0.37
36	revers	r/ReverseEngineering 0.34	r/hacking 0.52	r/netsec 0.52
37	team	r/cybersecurity 0.42	r/blueteamsec 0.44	r/netsec 0.46
38	languag	r/hackernews 0.47	r/netsec 0.52	r/ReverseEngineering 0.55
39	soc	r/CyberSecurityJobs 0.46	r/SecurityCareerAdvice 0.46	r/cybersecurity 0.53

Table 4 (continued)

No.	Exemplar	Subreddit 1	Subreddit 2	Subreddit 3
40	forc	r/cybersecurity 0.51	r/hacking 0.51	r/HowToHack 0.52
41	host	r/AskNetsec 0.44	r/cybersecurity 0.46	r/blueteamsec 0.47
42	authent	r/hacking 0.48	r/cybersecurity 0.48	r/cybersecurity_help 0.48
43	degre	r/CyberSecurityJobs 0.47	r/SecurityCareerAdvice 0.52	r/cybersecurity 0.57
44	virtual	r/cybersecurity_help 0.55	r/hacking 0.56	r/Hacking_Tutorials 0.56
45	inject	r/netsec 0.51	r/websecurityresearch 0.54	r/blueteamsec 0.56
46	incid	r/cybersecurity 0.51	r/blueteamsec 0.52	r/netsec 0.57
47	forens	r/computerforensics 0.41	r/cybersecurity 0.56	r/AskNetsec 0.58
48	entri	r/CyberSecurityJobs 0.42	r/SecurityCareerAdvice 0.5	r/cybersecurity 0.56
49	icloud	r/privacy 0.5	r/cybersecurity_help 0.5	r/hacking 0.51
50	ctf	r/netsecstudents 0.64	r/Hacking_Tutorials 0.66	r/HowToHack 0.67
51	storag	r/privacy 0.55	r/cybersecurity_help 0.59	r/AskNetsec 0.6
52	third	r/cybersecurity 0.52	r/privacy 0.53	r/Scams 0.53
53	airdrop	r/cryptography 0.74	r/ethicalhacking 0.82	r/Scams 0.85
54	bounti	r/Hacking_Tutorials 0.6	r/ethicalhacking 0.62	r/netsec 0.65

Acknowledgements The research in this paper is part of the PhD dissertation of the first author.

Author contributions Conceptualization: Konstantinos Charmanas, Lefteris Angelis; Methodology: Konstantinos Charmanas; Formal Analysis and investigation: Konstantinos Charmanas, Nikolaos Mittas; Writing—original draft preparation: Konstantinos Charmanas; Writing—review and editing: Nikolaos Mittas, Lefteris Angelis; Resources: Konstantinos Charmanas, Nikolaos Mittas; Supervision: Nikolaos Mittas, Lefteris Angelis; Data curation: Konstantinos Charmanas; Software: Konstantinos Charmanas, Lefteris Angelis; Visualization: Konstantinos Charmanas.

Funding Open access funding provided by HEAL-Link Greece. This research received no external funding.

Availability of data and materials The dataset, code and models of this research are not available due to policy/legal restrictions.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *Computers & Security, 38*, 97–102.
2. Goutam, R. K. (2015). Importance of cyber security. *International Journal of Computer Applications, 111*(7), 0975–8887.
3. Abomhara, M., & Køien, G. M. (2015). Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks. *Journal of Cyber Security and Mobility, 4*(1), 65–88.
4. Humayun, M., Niazi, M., Jhanjhi, N. Z., Alshayeb, M., & Mahmood, S. (2020). Cyber security threats and vulnerabilities: A systematic mapping study. *Arabian Journal for Science and Engineering, 45*, 3171–3189.
5. Shirey, R. (2000). Internet security glossary (No. rfc2828).
6. Kissel, R. (Ed.). (2011). *Glossary of key information security terms*. Diane Publishing.
7. Norman, T. L. (2014). *Integrated security systems design: A complete reference for building enterprise-wide digital security systems*. Butterworth-Heinemann.
8. Razzaq, A., Hur, A., Ahmad, H. F., & Masood, M. (2013). Cyber security: Threats, reasons, challenges, methodologies and state of the art solutions for industrial applications. In: 2013 IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS). IEEE. pp. 1–6.
9. Ben-Asher, N., & Gonzalez, C. (2015). Effects of cyber security knowledge on attack detection. *Computers in Human Behavior, 48*, 51–61.
10. Alves, F., Andongabo, A., Gashi, I., Ferreira, P. M., & Bessani, A. (2020). Follow the blue bird: A study on threat data published on Twitter. In Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25 (pp. 217–236). Springer International Publishing.
11. Yang, X. L., Lo, D., Xia, X., Wan, Z. Y., & Sun, J. L. (2016). What security questions do developers ask? A large-scale study of stack overflow posts. *Journal of Computer Science and Technology, 31*, 910–924.
12. Li, T., Louie, E., Dabbish, L., & Hong, J. I. (2021). How developers talk about personal data and what it means for user privacy: A case study of a developer forum on Reddit. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), 1–28.
13. Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New data sources in social science research: Things to know before working with Reddit data. *Social science computer review, 39*(5), 943–960.
14. Steinbaur, T. (2012). Information and social analysis of Reddit. In Proc. TROYSTEINBAUER CS. UCSB. EDU (pp. 1–12).
15. Sabottke, C., Suciu, O., & Dumitraș, T. (2015). Vulnerability disclosure in the age of social media: Exploiting twitter for predicting {Real-World} exploits. In 24th USENIX Security Symposium (USENIX Security 15) (pp. 1041–1056).
16. Bahl, A., Sharma, A., & Asghar, M. R. (2021). Vulnerability disclosure and cybersecurity awareness campaigns on twitter during COVID-19. *Security and Privacy, 4*(6), e180.
17. Schiappa, M., Chantry, G., & Garibay, I. (2019). Cyber security in a complex community: A social media analysis on common vulnerabilities and exposures. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 13–20). IEEE.
18. Chen, H., Liu, R., Park, N., & Subrahmanian, V. S. (2019). Using twitter to predict when vulnerabilities will be exploited. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 3143–3152).
19. Huang, S. Y., & Ban, T. (2020, December). Monitoring social media for vulnerability-threat prediction and topic analysis. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (pp. 1771–1776). IEEE.
20. de Sousa, D. A., de Faria, E. R., & Miani, R. S. (2020). Evaluating the performance of Twitter-based exploit detectors. arXiv preprint arXiv:2011.03113.
21. Syed, R., Rahafrooz, M., & Keisler, J. M. (2018). What it takes to get retweeted: An analysis of software vulnerability messages. *Computers in Human Behavior, 80*, 207–215.

22 Horawalavithana, S., Choudhury, N., Skvoretz, J., & Iamnitchi, A. (2022). Online discussion threads as conversation pools: predicting the growth of discussion threads on reddit. *Computational and Mathematical Organization Theory, 28*(2), 112–140.

23. Wu, M., Aranovich, R., & Filkov, V. (2021). Evolution and differentiation of the cybersecurity communities in three social question and answer sites: A mixed-methods analysis. *PLoS ONE, 16*(12), e0261954.
24. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972–976.
25. Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory, 37*(1), 145–151.
26. Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software, 53*, 1–18.
27. Lun, Y. Z., D'Innocenzo, A., Malavolta, I., & Di Benedetto, M. D. (2016). Cyber-physical systems security: a systematic mapping study. arXiv preprint arXiv:1605.09641.
28. Mohammed, N. M., Niazi, M., Alshayeb, M., & Mahmood, S. (2017). Exploring software security approaches in software development lifecycle: A systematic mapping study. *Computer Standards & Interfaces, 50*, 107–115.
29. Kaloroumakis, P. E., & Smith, M. J. (2021). Toward a knowledge graph of cybersecurity countermeasures. The MITRE Corporation, 11.
30. Horawalavithana, S., Bhattacharjee, A., Liu, R., Choudhury, N., O. Hall, L., & Iamnitchi, A. (2019). Mentions of security vulnerabilities on reddit, twitter and github. In IEEE/WIC/ACM International Conference on Web Intelligence (pp. 200–207).
31. Alperin, K., Joback, E., Shing, L., & Elkin, G. (2021). A framework for unsupervised classificiation and data mining of tweets about cyber vulnerabilities. arXiv preprint arXiv:2104.11695.
32. Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., & Shimada, H. (2019). Identification of cybersecurity specific content using the Doc2Vec language model. In 2019 IEEE 43rd annual computer software and applications conference (COMPSAC) (Vol. 1, pp. 396–401). IEEE.
33. Shrestha, P., Sathanur, A., Maharjan, S., Saldanha, E., Arendt, D., & Volkova, S. (2020). Multiple social platforms reveal actionable signals for software vulnerability awareness: A study of GitHub. *Twitter and Reddit. Plos one, 15*(3), e0230250.
34. Campbell Jr, J., Mensch, A. C., Zeno, G., Campbell, W. M., Lippmann, R. P., Weller-Fahy, D. J., & MIT Lincoln Laboratory Lexington United States. (2015). Finding malicious cyber discussions in social media (p. 0019). Technical report, MIT Lincoln Laboratory Lexington United States.
35. Azeez, N. A., Lawal, A. O., Misra, S., & Oluranti, J. (2022). Machine learning approach for identifying suspicious uniform resource locators (URLs) on Reddit social network. *African Journal of Science, Technology, Innovation and Development, 14*(6), 1618–1626.
36. Khurana, N., Mittal, S., Piplai, A., & Joshi, A. (2019). Preventing poisoning attacks on AI based threat intelligence systems. In 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1–6). IEEE.
37. Shields, T., Li, H., Lebedev, P., & Dykstra, J. (2020). Cyber Buzz: Examining Virality Characteristics of Cybersecurity Content In Social Networks. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 64, No. 1, pp. 441–445). Sage CA: Los Angeles, CA: SAGE Publications.
38. Wang, G., & Kwok, S. W. H. (2021). Using k-means clustering method with Doc2Vec to understand the twitter users' opinions on COVID-19 vaccination. In 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1–4). IEEE.
39. Mendsaikhan, O., Hasegawa, H., Yamaguchi, Y., Shimada, H., & Bataa, E. (2020). Identification of cybersecurity specific content using different language models. *Journal of Information Processing, 28*, 623–632.
40. Priya, S., Sequeira, R., Chandra, J., & Dandapat, S. K. (2019). Where should one get news updates: Twitter or Reddit. *Online Social Networks and Media, 9*, 17–29.
41. Jamnik, M. R., & Lane, D. J. (2017). The use of Reddit as an inexpensive source for high-quality data. *Practical Assessment, Research, and Evaluation, 22*(1), 5.
42. Medvedev, A. N., Lambiotte, R., & Delvenne, J. C. (2019). The anatomy of Reddit: An overview of academic research. *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches, 10*, 183–204.
43. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. In Proceedings of the international AAAI conference on web and social media (Vol. 14, pp. 830–839).
44. Lakkaraju, K. H. (2012). Demystifying content popularity on Reddit.

45. Deaton, S., Hutchison, S., & Matthews, S. J. (2017). Using Machine Learning to Predict the Popularity of Reddit Comments. seandeaton.com.
46. Stoddard, G. (2015). Popularity dynamics and intrinsic quality in reddit and hacker news. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 9, No. 1, pp. 416–425).
47. Jasser, J., Garibay, I., Scheinert, S., & Mantzaris, A. V. (2022). Controversial information spreads faster and further than non-controversial information in Reddit. *Journal of Computational Social Science, 5*(1), 111–122.
48. Thukral, S., Meisheri, H., Kataria, T., Agarwal, A., Verma, I., Chatterjee, A., & Dey, L. (2018). Analyzing behavioral trends in community driven discussion platforms like reddit. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 662–669). IEEE.
49. Weninger, T., Zhu, X. A., & Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining (pp. 579–583).
50. Buntain, C., & Golbeck, J. (2014). Identifying social roles in reddit using network structure. In Proceedings of the 23rd international conference on world wide web (pp. 615–620).
51. Zayats, V., & Ostendorf, M. (2018). Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics, 6*, 121–132.
52. Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management, 57*(2), 102034.
53. Duguay, P. A. (2022). Read it on Reddit: Homogeneity and ideological segregation in the age of social news. *Social Science Computer Review, 40*(5), 1186–1202.
54. Brett, E. I., Stevens, E. M., Wagener, T. L., Leavens, E. L., Morgan, T. L., Cotton, W. D., & Hébert, E. T. (2019). A content analysis of JUUL discussions on social media: Using Reddit to understand patterns and perceptions of JUUL use. *Drug and alcohol dependence, 194*, 358–362.
55. Sowles, S. J., McLeary, M., Optican, A., Cahn, E., Krauss, M. J., Fitzsimmons-Craft, E. E., & Cavazos-Rehg, P. A. (2018). A content analysis of an online pro-eating disorder community on Reddit. *Body Image, 24*, 137–144.
56. Lossio-Ventura, J. A., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2019). Clustering and topic modeling over tweets: A comparison over a health dataset. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1544–1547). IEEE.
57. Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80–88).
58 Al-khateeb, S., & Agarwal, N. (2019). *Deviance in social media and social cyber forensics: Uncovering hidden relations using open source information (OSINF)*. Springer International Publishing.
59. Babb, R. E. (2021). The Community Industry: An Analysis of Reddit and/r/socialism (Doctoral dissertation, Bowling Green State University).
60. Nasim, Z., & Haider, S. (2022). Cluster analysis of Urdu tweets. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 2170–2179.
61. Bin Abdur Rakib, T., & Soon, L. K. (2018). Using the Reddit corpus for cyberbully detection. In Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19–21, 2018, Proceedings, Part I 10 (pp. 180–189). Springer International Publishing.
62 Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society, 7*(2), 20563051211019004.
63. Okon, E., Rachakonda, V., Hong, H. J., Callison-Burch, C., & Lipoff, J. B. (2020). Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. *Journal of the American Academy of Dermatology, 83*(3), 803–808.
64. Ruan, T., & Lv, Q. (2022). Public perception of electric vehicles on reddit over the past decade. *Communications in Transportation Research, 2*, 100070.
65. Park, A., & Conway, M. (2017). Tracking health related discussions on Reddit for public health applications. In AMIA annual symposium proceedings (Vol. 2017, p. 1362). American Medical Informatics Association.
66. Parsons, J., Schrider, M., Ogunlela, O., & Ghanavati, S. (2023). Understanding Developers Privacy Concerns Through Reddit Thread Analysis. arXiv preprint arXiv:2304.07650.

67. Westrupp, E. M., Greenwood, C. J., Fuller-Tyszkiewicz, M., Berkowitz, T. S., Hagg, L., & Youssef, G. (2022). Text mining of Reddit posts: Using latent Dirichlet allocation to identify common parenting issues. *PLoS ONE, 17*(2), e0262529.
68. Liu, Y., & Anwar, M. (2022). Learning Programming in Social Media: An NLP-powered Reddit Study. In 2022 Fourth International Conference on Transdisciplinary AI (TransAI) (pp. 55–58). IEEE.
69. Johnson, A. K., Bhaumik, R., Nandi, D., Roy, A., & Mehta, S. D. (2022). Sexually transmitted disease-related Reddit posts during the COVID-19 pandemic: latent Dirichlet allocation analysis. *Journal of Medical Internet Research, 24*(10), e37258.
70. El-Jack, K., Henderson, K., Andy, A. U., & Southwick, L. (2022). Reddit users' questions and concerns about Anesthesia. *International Journal of Medical Students, 10*(4), 370–374.
71. He, Q. (1999). Knowledge discovery through co-word analysis.
72. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.
73. Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL, 30*, 31–40.
74. Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics, 8*, 439–453.
75. Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of classification, 31*, 274–295.
76. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1–14.
77. Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Konstantinos Charmanas[1] · **Nikolaos Mittas**[2] · **Lefteris Angelis**[1]

✉ Konstantinos Charmanas
kcharman@csd.auth.gr

[1] School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

[2] Hephaestus Laboratory, Department of Chemistry, School of Science, Democritus University of Thrace, Kavala, Greece