Expert Systems With Applications 176 (2021) 114808

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Large-scale analysis of grooming in modern social networks

Nikolaos Lykousas [a], Constantinos Patsakis [a,b,*]

[a] Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., 18534 Piraeus, Greece
[b] Information Management Systems Institute, Athena Research Center, Artemidos 6, Marousi 15125, Greece

ARTICLE INFO

Keywords:
Online grooming
Social networks
LDA
Text analysis
Emoji

ABSTRACT

Social networks are evolving to engage their users more by providing them with more functionalities. One of the most attracting ones is streaming. Users may broadcast part of their daily lives to thousands of others world-wide and interact with them in real-time. Unfortunately, this feature is reportedly exploited for grooming. In this work, we provide the first in-depth analysis of this problem for social live streaming services. More precisely, using a dataset that we collected, we identify predatory behaviours and grooming on chats that bypassed the moderation mechanisms of the LiveMe, the service under investigation. Beyond the traditional text approaches, we also investigate the relevance of emojis in this context, as well as the user interactions through the gift mechanisms of LiveMe. Finally, our analysis indicates the possibility of grooming towards minors, showing the extent of the problem in such platforms.

1. Introduction

The recent advances in telecommunications have unleashed the potentials of sharing and exchanging content, changing radically the way we interact with others online. By lifting many bandwidth barriers, users may generate and share arbitrary content and disseminate it instantly to millions of users. As a result, we see Social Networks and Media's dominance in various aspects of our daily lives.

This radical shift and penetration of mobile devices have led millions of people and youngsters to use them on a daily basis. While most social networks have specific policies about use from minors, in practice, this policy is bypassed. Minors declare fake ages to register to service providers and end up using the services as normal users. While this might not be noticed or overseen by service providers, this is not the users' case. Unfortunately, thousands of users maliciously target minors. Of specific interest is the case of *grooming*. Grooming refers to the process by which an offender prepares a victim for sexually abusive behaviour. More precisely, according to Craven et al. (2006):

> *[Grooming is]…a process by which a person prepares a child, significant others, and the environment for the abuse of this child. Specific goals include gaining access to the child, gaining the child's compliance, and maintaining the child's secrecy to avoid disclosure. This process serves to strengthen the offender's abusive pattern, as it may be used as a means of justifying or denying their actions…*

Apparently, child grooming is of extreme importance due to the impact that it can have in the children's lives. In fact, despite the measures that social networks might have already taken, they do not seem to be successful at all.[1] To this end, it is necessary to investigate how grooming in social networks works and how groomers manage to bypass the policies and filters set by social networks. In terms of verbal content, currently, there is only one available dataset from the Perverted Justice website.[2] The organisation behind this website, Perverted Justice Foundation, Inc., has recruited volunteers to carry out sting operations. They appear as minors to several online services and record the interactions with them. Their operations have made a tremendous positive impact as they have led to the conviction of more than 620 offenders. While undoubtedly, this is a huge contribution, the problem persists, and the provided dataset is rather old to be used for modern filters.

1.1. Motivation

The past few years, there is a steady increase of reports in mainstream media and officials[3] regarding the exploitation of social networks for grooming. The problem regardless of the age factor is rather big and

* Corresponding author.
E-mail addresses: nlykousas@unipi.gr (N. Lykousas), kpatsak@unipi.gr (C. Patsakis).
[1] https://www.bbc.com/news/uk-47410520.
[2] http://perverted-justice.com/.
[3] https://www.nspcc.org.uk/what-we-do/news-opinion/3000-new-grooming-offences/.

https://doi.org/10.1016/j.eswa.2021.114808
Received 1 April 2020; Received in revised form 15 December 2020; Accepted 28 February 2021
Available online 6 March 2021

stigmatises the life of thousands of people. The emergence of new social networks, allowing live streaming to potentially thousands of users along with traditional chatting and appraisal methods of traditional social networks can be further exploited for grooming.

The findings discussed in Lykousas et al. (2018) demonstrated that the moderation systems used by the LiveMe platform at that time were highly ineffective in suspending the accounts of deviant users producing adult content. Notably, in the same year, FOX 11; a major mainstream media outlet, reported that (Melugin, 2018):

> *A FOX 11 investigation has found that pedophiles are using the popular live streaming app LiveMe to manipulate underage girls into performing sexual acts, reward them with virtual currency, and then post screen captures or recordings of the girls online to be sold and distributed as child porn.*

As such, it is reasonable to assume that the adult content problem and the sexual grooming behaviours identified by FOX 11 are related to some extent. In this work, we aim to unveil communication patterns of sexual groomers in the context of social live streaming services.

1.2. Main contributions

Our work primarily aims to identify and disentangle the mechanics of grooming and predatory behaviours in the context of Social Live Streaming Services, by analysing the behavioural and communication patterns of viewers, at a broadcast-level. Therefore, user-level detection of groomers falls beyond the scope of our work. It has to be noted though that a distinctive difference of grooming in Social Live Streaming Services is that is not performed through one-to-one interaction with the victim, but many-to-one.

Based on the above, the contributions of this work are multifold. First, we facilitate research in this field and the generation of new filters and algorithms to detect such predatory behaviour through the release of a large-scale dataset of both verbal and non-verbal interactions (e.g. likes and rewards) in a Social Live Streaming Service. Due to its nature, the dataset is available only to researchers and law enforcement agencies upon request via Zenodo.[4] Second, we analyse the basic characteristics of the verbal content. Our analysis illustrates how such predatory behaviour bypasses the filters of service providers by, e.g. altering some “bad words”, or by using emojis. Notably, to the best of our knowledge, this is the first work highlighting the role of emojis in grooming. Then, based on our analysis, we manage to identify chats where grooming is performed. Moreover, we analyse non-verbal interactions between users that differentiate chats where grooming is performed from the others. Finally, our analysis shows that it is possible to identify illegal actions, such as the grooming of minors.

1.3. Organisation of the article

The rest of this work is organised as follows. In the next section, we provide an overview of the related work on the detection of deviant behaviour and grooming. Then, we provide the legal and ethical justification of collecting such data and GDPR compliance assessment. Section 4 provides an overview of our dataset. In Section 5, we analyse our dataset and provide some insight into it. Afterwards, in Section 6, we investigate possible modelling of grooming behaviours using both verbal and non-verbal features. Finally, the article concludes summarising our contributions and discussing ideas for future work.

2. Related work

Coletto et al. (2017) aimed at going beyond previous studies that considered deviant groups in isolation by observing them in context. In particular, they attempted to answer questions relevant to the deviant behaviours related with pornographic material in the social media context, such as i) how much deviant groups are structurally secluded from the rest of the social network, and what are the characteristics of their subgroups who build ties with the external world; ii) how the content produced by a deviant community spreads and what is the entity of the diffusion which reaches users outside the boundaries of the deviant community who voluntarily or inadvertently access the adult content, and iii) what is the demographic composition of producers and consumers of deviant content and what is the potential risk that young boys and girls are exposed to it. Very interestingly, they find that while deviant communities may have limited size, they are tightly connected and structured in subgroups. Moreover, the content which is first shared in these groups soon reaches a broad audience of not previously considered deviant users.

The proliferation of the Internet has transformed child sexual abuse into a crime without geographical boundaries. Child sex offenders turning to the Internet as a means of creating and distributing child pornography has allowed the creation of a network of support groups for child sex offenders, when historically, this was an offence that occurred in isolation (Westlake et al., 2016). This concern was echoed by Mitchell et al. (2010), who recognised that a small percentage of offenders used social networking sites (SNS) to distribute child pornography. While there is scientific debate on whether the online predator is a new type of child sex offender (Quayle et al., 2000) or if those with a predisposition to offend are responding to the opportunities afforded by the new forms of social media (Cooper, 1998), empirical evidence points to the problem of Internet-based paedophilia as endemic. Recent work, such as Winters and Jeglic (2017), Zambrano et al. (2019), shows that nearly half of the offenders who had committed one or more contact offences, i.e., they had directly and physically abused children, had displayed so-called “grooming behaviour”.

However, when investigating the possibility of developing automated methods to detect grooming online, researchers are confronted with many issues. First, only one benchmark dataset contains (English) chat conversations written by child sex offenders, the PAN 2012 Sexual Predator Identification dataset, which leverages data from PJ. Concretely, PJ data comprises a single class of chats in the context of PAN 2012 data. Yet, because the victims were actually adult volunteers posing as children, it is likely that these conversations are not entirely representative of online predator-victim communications (Pendar, 2007). Moreover, since the seduction stage often shows similar characteristics with adults’ or teenagers’ flirting, initial studies trying to detect predatory behaviour directly on the user level typically resulted in numerous false positives when they were applied to non-predatory sexually-oriented chat conversations in the PAN 2012 dataset (Inches and Crestani, 2012).

For machine learning algorithms to identify online sexual predators effectively, they need to be trained with both illegal conversations between offenders and their victims and sexually-oriented conversations between consenting adults (Pendar, 2007). Since such data are rarely made public, initial studies (Pendar, 2007; McGhee et al., 2011) only experimented with the PJ data. The k-NN classification experiments based on word token n-grams performed in Pendar (2007) achieved up to 93.4% F-score (trigrams with $k = 30$) when identifying the predators from the pseudo-victims. Miah et al. were the first to include additional corpora in the non-predatory class (Miah et al., 2011). They included 85 conversations containing adult descriptions of sexual fantasies and 107 general non-offensive chat logs from websites likehttp://www.fugly.com andhttp://chatdump.com. When distinguishing between 200 PJ conversations and these additional chat logs, the Naïve Bayes classifier outperformed the Decision Tree and the Regression classifier, which resulted in an F-score of 91.7% for the PJ class. In Bogdanova et al. (2014), Peersman et al. (2012), Morris and Hirst (2012), Hidalgo and Díaz (2012), the researchers used a corpus of cybersex chat logs and the

[4] https://zenodo.org/record/3560365.

Naval Postgraduate School (NPS) chat corpus and experimented with new feature types such as emotional markers, emoticons and imperative sentences and computed sex-related lexical chains to detect offenders directly in the PJ dataset automatically. Their Naïve Bayes classifier yielded an accuracy of 92% for PJ predators *vs* NPS and 94% for PJ predators *vs* cybersex based on their high-level features. However, both Miah et al. (2011) and Bogdanova et al. (2014) did not filter out any cues that were typical of the social media platforms from which the additional corpora were extracted, which could entail that their models were (to some degree) trained on detecting these cues rather than the grooming content. Moreover, because the high-level features described by Bogdanova et al. (2014) were (partially) derived from the PJ dataset itself, these experiments may have resulted in overestimated accuracy when detecting predators from the same dataset.

Recently, the detection of Internet child sex offenders has been extensively investigated in the framework of the PAN 2012 competition, during which efforts have been made to pair the PJ data with a whole range of non-predatory data, including cybersex conversations between adults (Inches and Crestani, 2012). Because the PAN 2012 benchmark dataset was heavily skewed towards the non-predatory class, most participants applied a two-stage classification framework in which they combined information on the conversation level to the user level (Villatoro-Tello et al., 2012). Moreover, apart from one submission that used character-gram features, all other studies used (combinations of) lexical (e.g., token unigrams) and "behavioural" features (e.g., the frequency of turn-taking or the number of questions asked). Morris and Hirst (2012) achieved the best results using a Neural Network classifier combined with a binary weighting scheme in a two-stage approach to first identify the suspicious conversations and, secondly, distinguish between the predator and the victim. Their system achieved an F-score of 87.3%. However, during their study, they assumed that "*predators usually apply the same course of conduct pattern when they are approaching a child*" (Morris and Hirst, 2012), which is in contrast with research by Gottschalk (2011), which resulted in three different types of predators and, hence, of grooming approaches. Moreover, the PJ dataset was also not cleansed of platform-specific cues, which could again have led to overestimated F-scores during the competition. A more detailed overview of the PAN 2012 International Sexual Predator Identification Competition results can be found in Inches and Crestani (2012).

Concerning the content of predatory chat conversations, McGhee et al. were the first to investigate the possibility to detect different stages in the grooming process automatically (McGhee et al., 2011). Based on an expanded dictionary of terms they applied a rule-based approach, which categorised a post as belonging to the stage of gaining personal information, grooming (which included lowering inhibitions or reframing and sexual references), or none. Their rule-based approach outperformed the machine learning algorithms they tested and reached up to 75.1% accuracy when categorising posts from the PJ dataset into one of these stages. A similar approach was used by Michalopoulos and Mavridis whose Naïve Bayes classifier achieved a 96% accuracy when categorising predatory PJ posts as belonging to either the gaining access, the deceptive relationship or the sexual affair grooming stage (Michalopoulos and Mavridis, 2011). The second task of the PAN 2012 competition consisted of detecting the specific posts that were most typical of predatory behaviour from the users that were labelled suspicious during the first task. To this end, most participants either created a dictionary-based filter containing suspicious terms (Morris and Hirst, 2012; Parapar et al., 2012) or used their post-level predictions from the predator identification task (Kontostathis et al., 2012; Hidalgo and Díaz, 2012). The best F-score was achieved by Peersman et al. (2012), who used a dictionary-based filter highlighting the utterances that referred to one of the following grooming stages: sexual stage, re-framing, approach, requests for data, isolation from adult supervision and age- and child-related references. Their approach resulted in a 35.8% precision, a 26.1% recall and a 30.2% F-score. Finally, Elzinga et al. (2012) proposed a method based on Temporal Concept Analysis using Temporal Relational Semantic Systems, conceptual scaling and nested line diagrams to analyse PJ chat conversations. Their transition diagrams of predatory chat conversations seemed to be useful for measuring the level of threat each offender poses to his victim based on the presence of the different grooming stages.

Although these studies showed promising results, the issue remains that these methods are applied to a corpus that contains conversations between offenders and pseudo-victims. Hence, the adult volunteers that were posing as children could not accede to requests for "cammin", sending pictures, etc. As a result, the PJ dataset contains hardly any conversations by groomers, because this type of offender typically does not invest much time in the seduction process and switches to a different victim when his needs are not fulfilled quickly. Moreover, it is highly likely that children would have responded differently to the grooming utterances than the adult volunteers did, which could have influenced the offenders' language use.

2.1. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) is a type of generative probabilistic model proposed by Blei et al. (2003). It comprises an endogenous NLP technique, which as highlighted in Cambria and White (2014) "*involves the use of machine-learning techniques to perform semantic analysis of a corpus by building structures that approximate concepts from a large set of documents*" without relying on any external knowledge base. As the name implies, LDA is a latent variable model in which each item in a collection (e.g., each text document in a corpus) is modelled as a finite mixture over an underlying set of topics. Each of these topics is characterised by a distribution over item properties (e.g., words). LDA assumes that these properties are exchangeable (i.e., ordering of words is ignored, as in many other "bag of words" approaches in text modelling), and that the properties of each document are observable (e.g., the words in each document are known). The word distribution for each topic and the topic distribution for each document are unobserved; they are learned from the data.

Since LDA is an unsupervised topic modelling method, there is no direct measure to identify the optimal number of topics to include in a model. What LDA does is to assign to documents probabilities to belong to different topics (an integer number k provided by the user), where these probabilities depend on the occurrence of words which are assumed to co-occur in documents belonging to the same topic (Dirichlet prior assumption). This exemplifies the main idea behind all unsupervised topic models, that language is organised by latent dimensions that actors may not even be aware of McFarland et al. (2013). Thus, LDA exploits that even if a word belongs to many topics, occurring in them with different probabilities, they co-occur with neighbouring words in each topic with other probabilities that help define the topics better. The best number of topics is the number of topics that helps the most human interpretability of the topics. This means that if the topics given by LDA can be well-distinguished by humans, then the corresponding number of topics is acceptable. Researchers have recommended various approaches to establish the optimal k (e.g. Cao et al., 2009; Arun et al., 2010; Deveaud et al., 2014; Röder et al., 2015; Zhao et al., 2015). These approaches provide a good range of possible k values that are mathematically plausible. However, according to DiMaggio et al. (2013), when topic modelling is used to identify themes and assist in interpretation (like in the present study), rather than to predict a knowable state or quantity, there is no statistical test for the optimal number of topics or the quality of a solution. A simple way to evaluate topic models is to look at the qualities of each topic and discern whether they are reasonable (McFarland et al., 2013). In addition, the topic number selection was guided by the model's ability to identify a number of substantively meaningful and analytically useful topics. In fact, the increase in fit is sometimes at the expense of interpretability due to overfitting (Dyer et al., 2017). Increasing the number of topics, producing ever-finer partitions can result in a less useful model because it becomes almost

impossible for humans to differentiate between many of the topics (Chang et al., 2009). Ultimately, the choice of models must be driven by the questions being analysed. DiMaggio et al. (2013) suggest that the process is empirically disciplined, in that, if the data are inappropriate for answering the analysts' questions, no topic model will produce a useful reduction of the data. To the best of our knowledge, the topic coherence measure with the most considerable correlation to human interpretability is the C_v score defined in Röder et al. (2015), which we also adopt in this study to establish the optimal number of topics, see Section 6.

3. Ethical and legal compliance

Data scraping from the web is extensively used by academic researchers to track the web, and companies to gain information about their customers. The philosophy of crawling is to index the web and the Internet as a whole, to make information available to the public, and to extract information for different business and research purposes. Yet, due to the invasive practices used for extracting large amounts of information, there is an ongoing debate on the ethical and legal aspects of web data crawling.

According to Internet advocates, if web crawling were to be unethical, then the whole web would not have been discoverable since the entire expansion of the Internet is based on web crawling. As a matter of fact, web scraping has benefited the web so much that virtually everyone on the net is directly or indirectly involved in web scraping. Even big service providers like Google scrap the Internet to be able to provide qualified and verified data in the search results. However, for web data crawling to be ethical, there must be some rules to be followed (like those imposed in the `robots.txt` file of every web site) to not infringe on the security and the rights of the users. In fact, there are already several professional web scraping service providers who abide by the general rules and regulations to get adequate and appropriate authorisation from the concerned web resource.

As a matter of fact, many scholars advocate that it is the application of the data that have been scrapped and not the web scrapping *per se, that may be unethical or illegal. For instance, there might be issues when data that are not meant to be made public are scraped and reused for commercial or other purposes. The legal issues of web scraping are widely discussed in the context of the copyrighted and data protection law. The latter is expressed in the EU by the GDPR, which defines the privacy and data protection rights and the rules to be respected when the processing of personal data takes place. While the GDPR is applicable even for research purposes, it states that for meeting "the specificities of processing personal data for scientific research purposes, specific conditions should apply in particular as regards the publication or otherwise disclosure of personal data in the context of scientific research purposes" (recital 159). Inevitably, when web crawling collects the personal data of web users to facilitate specific research purposes, this processing needs to be aligned with the data protection principles enshrined in the GDPR.*

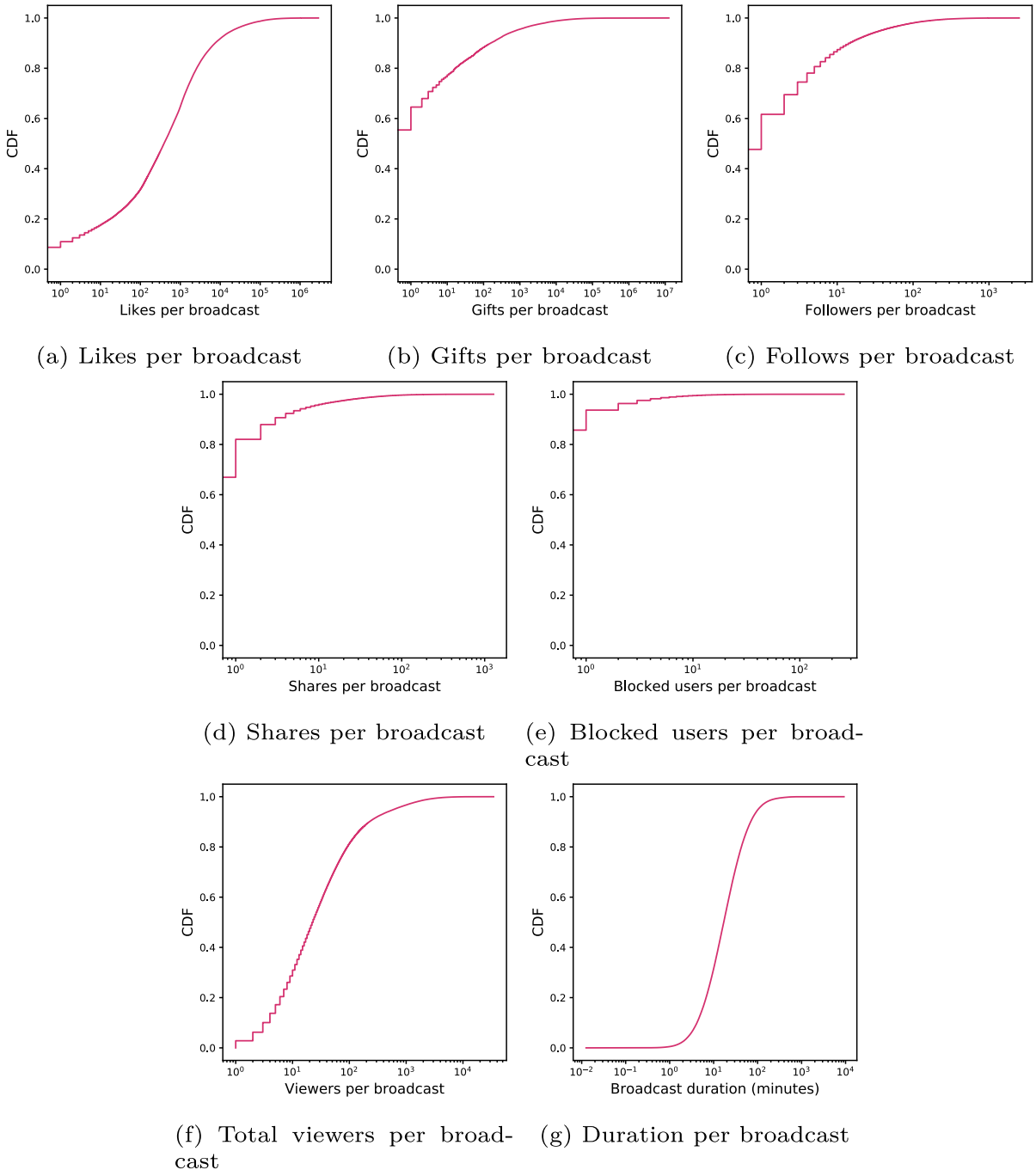
The GDPR requires a specific lawful basis for the processing of the personal data of individuals, with the consent to be the most commonly advertised among them. Beyond consent, however, the GDPR defines some other bases so as the processing of the personal data to be lawful: when the processing is necessary to protect the vital interests of the data subject or of another natural person (Article 6(1)(d)); when the processing is necessary for the performance of a task carried out in the public interest (Article 6(1)(e)); or when the processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party (Article 6(1)(f)). Therefore, while research is not explicitly designated as its own lawful basis for processing, in some cases it may qualify under Articles 6(1)(d)(e)(f) compatible with some of the already foreseen lawful bases. When a controller collects personal data under a lawful basis, Article 6(4) allows it to process the data for a secondary research purpose. Thus, while the GDPR explicitly permits re-purposing collected data for research, it also may permit a controller to collect personal data initially for research purposes, without requiring the data subject's consent.

Furthermore, although research is not mentioned explicitly as a lawful basis for personal data processing, Recital 157 identifies the benefits associated with personal data research, subject to appropriate conditions and safeguards. These benefits include the potential for new knowledge when researchers "*obtain essential knowledge about the long-term correlation of a number of social conditions*". The results of the research "*obtained through registries provide solid, high-quality knowledge which can provide the basis for the formulation and implementation of knowledge-based policy, improve the quality of life for a number of people, and improve the efficiency of social services.*".

Moreover, the GDPR foresees derogations for the secondary processing of personal data for research purposes as long as there is a lawful basis for such processing (Article 5, Recital 50). Article 89 sets out the "*appropriate safeguards*" that controllers must implement to further process personal data for research. It mandates controllers explicitly to put in place "*technical and organisational measures*" to ensure that they process only the personal data necessary for the research purposes, in accordance with the principle of data minimisation outlined in Article 5 (c). Article 89(1) provides that one way for a controller to comply with the mandate for technical and organisational measures is through the deployment of "*pseudonymisation.*" Pseudonymisation is "*the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution to an identified or identifiable individual*" (Article 4(3b)).

Taken the above into consideration, one may consider that the case of personal data scraped from social media sites without the consent of the user that added them, may raise serious concerns regarding its ethical and legal consequences. Yet, these concerns can be easily removed, as we demonstrate below, when data scrapping is performed by researchers to facilitate the mitigation of malevolent uses such as those of pedophile and sex exploitations. More specifically, a team of researchers scraped a well-known social media site that attracts millions of teenagers (even if the web site's terms of service forbid its use by people under the age of 18). They found that the exchanged text chats among its participants include numerous instances of discussions involving sexual harassment and pedophile actions, all covered up under seemingly innocent words and terminologies that are impossible to be tracked by conventional software tailored to identify specific words for sex abuse. To facilitate research on advanced and innovative ways of tracking down suspicious cases of child abuse and harassment, the researchers, after scrapping the chats on the site referring to the coded malevolent conversations, published a dedicated corpus including these suspicious words, strings and emoticons. All user data, namely the user's nickname, have been anonymised with masking techniques whereas every single user was always masked with the same string. Taking into account that the identification of the users could be potentially possible when additional information (held by the researchers) is used, this masking technique is, in fact, a pseudonymisation in GDPR terms. Since pseudonymised data are still personal, they still fall under the scope of the GDPR. Therefore, researchers had to ensure that the processing of the personal data contained in the scrapped chats is compatible with the data protection provisions of the GDPR, and in particular with at least one of the six lawful purposes of processing enshrined in GDPR Article 6. Taking into account that the undertaken data crawling of the personal data can protect the vital interests of the children participating in the social media site so an not to be fooled by pedophile users, as well as that this processing is beyond any doubt carried out in the public interest, the data scrapping and subsequent analysis of the concerned data by the researcher are in accordance with the GDPR.

Particular attention should be paid for the processing of users data, given that the processed information most likely refers to the sexual preferences of the data subjects, a piece of information considered to be



(a) Likes per broadcast (b) Gifts per broadcast (c) Follows per broadcast

(d) Shares per broadcast (e) Blocked users per broadcast

(f) Total viewers per broadcast (g) Duration per broadcast

Fig. 1. Cumulative distribution functions (CDFs) of broadcast metadata features and interactions.

among the special categories of personal data referred to as "*sensitive*" for which stricter provisions apply (Article 9). Yet, derogating from the prohibition on processing special categories of personal data "*should also be allowed when provided for in Union or Member State law and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where it is in the public interest to do so*" (recital 52, Article 9(2)(j)). And beyond any doubt, protecting children from pedophile actions and sexual harassment is, above all, of substantial public interest and has been foreseen to all domestic legislations. Therefore, provided that the processing of the personal data of the data subjects is proportionate to the aim pursued, respects the essence of the right to data protection and provides for suitable and specific measures, i.e. pseudonymisation, to safeguard the fundamental rights and the interests of the data subject, the derogations for processing sensitive information under the Article 9 (2) are fulfilled.

Finally, the GDPR Article 12(1) requires controllers to "*take appropriate measures*" to inform data subjects of the nature of the processing activities and the rights available to them. Controllers are required to provide this information in all circumstances, regardless of whether consent is the basis for processing, "*in a concise, transparent, intelligible and easily accessible form, using clear and plain language*" (Article 12(1)). Nevertheless, a researcher may be exempted from the notice requirement if she received the personal data from someone other than the data subject, such as where the data came from a publicly available source. Article 14 exempts controllers in these circumstances, if "*the provision of such information proves impossible or would involve a disproportionate effort,*" which "*could in particular be the case*" in the research context (Recital 62). A researcher also may claim an exemption if providing

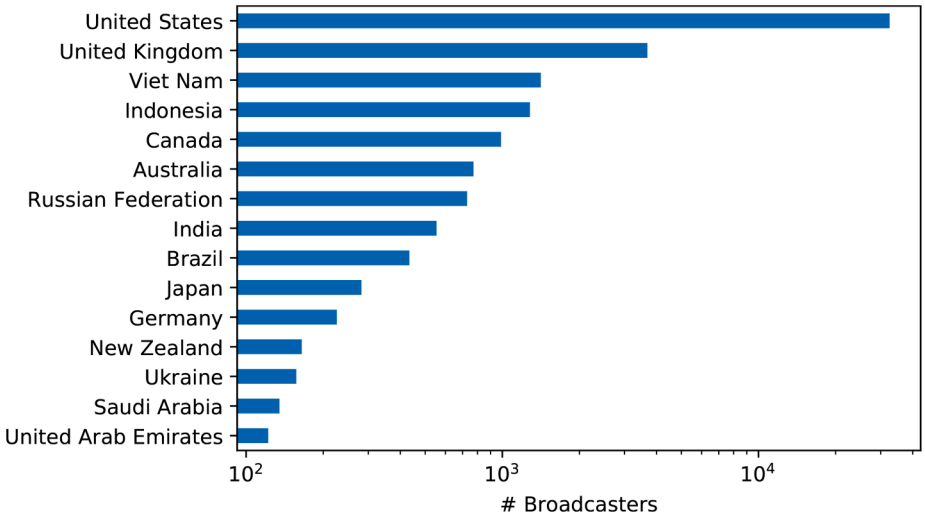
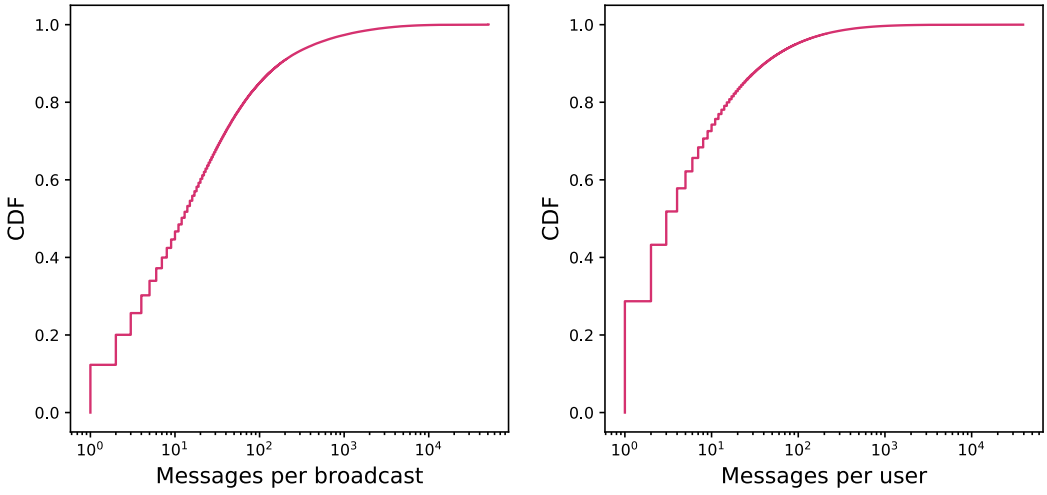


Fig. 2. Broadcasters count per country.



(a) chat messages per broadcast (b) chat messages per user

Fig. 3. Cumulative distribution functions (CDFs) of the chat messages per broadcast and per user.

notice would be "*likely to render impossible or seriously impair the achievement of the [research] objectives," provided there are appropriate safeguards in place, "including making the information publicly available*" (Article 14(5)(b)).

In summary, scrapping personal data from social media sites and publishing them in pseudonymised form for research purposes is legal and ethical as long as it is performed to protect the vital interests of the data subjects or others and it is in the public interest to do so.

4. The dataset

In what follows, we analyse a large-scale dataset that we created based on the public interactions between streamers and viewers during the live broadcasts of users identified as adult content producers in Lykousas et al. (2018), from the LiveMe[5] platform, a major Social Live Streaming Service (SLSS). The dataset comprises 39,382,838 chat messages exchanged by 1,428,284 users, in the context of 291,487 live broadcasts during a period of approximately two years, from July 2016 to June 2018. Each broadcast effectively functions as a temporary chatroom. The audience can interact with the streamers via text messages and reward them with virtual rewards, e.g. points, gifts, badges (some of which are purchasable) even virtual money. Apart from the chat messages, the dataset contains a wide range of user interactions along with metadata. We describe the features below:

- **Metadata (broadcast)**

  - Total Viewers: total number of viewers who joined the livestream as viewers.
  - Duration: duration of stream in seconds

- **Metatadata (broadcaster)**

  - Country Code

- **Interactions**

  - Likes: Viewers who liked the broadcast & the number of likes given.

[5] https://www.liveme.com/.

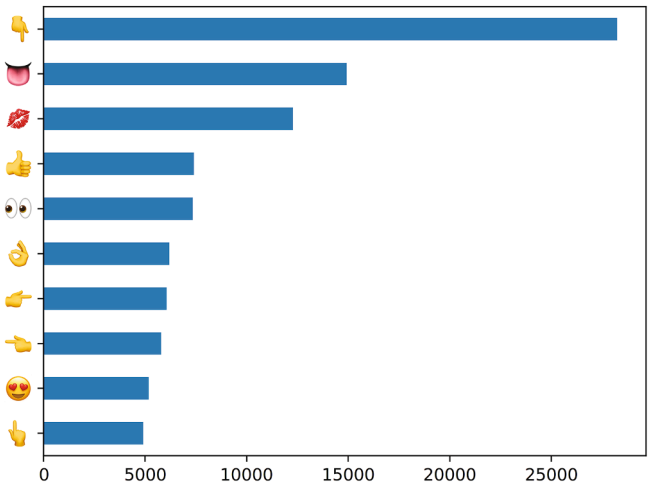


Fig. 4. Top emoji collocations for clothing related emojis.

- Follows: Viewers who followed the broadcaster during the livestream.
- Gifts: Viewers who sent virtual gifts to the broadcaster, along with value (in virtual currency) for each gift.
- Shares: Viewers who shared the broadcast (via a link so others can join).
- Blocks: Viewers who have been blocked by the broadcaster (i.e. banned from a stream).

To better understand how the features mentioned above are distributed, we plot the cumulative distribution functions (CDFs) in Fig. 1. We observe that every broadcast in our dataset had viewers (143.5 on average) and sizeable duration (31.4 min on average). While most of the broadcasts received likes (92%), the 55% did not receive any gifts (since they cost money, contrary to likes). Furthermore, 47% for the broadcasts did not generate any new followers for the broadcasters. At the same time, the interactions of sharing and blocking are relatively rare in our dataset (i.e. they are zero for 0.67% and 0.86% of the broadcasts, respectively). Next, to understand the geographical distribution of adult content producers, we plot the distribution of the broadcasters per country of the whole dataset, focusing on the 15 countries with most broadcasters, in Fig. 2.

5. Large-scale grooming analysis

By plotting the CDFs of the chat messages per broadcast and per user in Fig. 3,4, we notice that around 82% of the broadcasts of adult content producers receive less than 100 chat messages. Moreover, out of the unique users chatting during these broadcasts, only 30% send more than ten messages in total. Both distributions are particularly heavy-tailed, meaning that the majority of chat messages in our dataset are exchanged during a few highly popular broadcasts.

To identify sexual grooming behaviour in the chat messages, we adopted the approach followed by several authors in the most recent relevant works (Drouin et al., 2017; Lorenzo-Dus and Kinzel, 2019; Lorenzo-Dus et al., 2020) analysing the Perverted-Justice Dataset (PJ), which although dated and relatively small-scale, was the only publicly available dataset of chats produced by online groomers to date. To this end, we search the chat messages comprising our dataset for sexual content keywords defined in Linguistic Inquiry and Word Count (LIWC) corpus (Pennebaker et al., 2015). More precisely, the 2015 version of the LIWC dictionary for the sexual content variable comprises a total of 131 words. These include a wide range of terms about sexual matters, including sexual orientation (e.g. bi-sexual, heterosexual), sexual organs (e.g. penis*, vagin*, womb), slang terms, sexually transmitted diseases and infections, sexual violence and assault terms and sex enhancements. The most frequently occurring sexual terms in the PJ dataset, had a very low number of occurrences in the LiveMe chats (less than five exact matches in most occasions). The very low occurrence of such words implies the existence of an automated filtering mechanism in place. Nonetheless, relevant literature about online chat has demonstrated that users with previous exposure to text-based automatic moderation

Table 1
Top 15 verbs (simple and phrasal) associated with clothing items.

Verb	Count
Wear	6553
Show	5817
Remove	3947
See	3765
Get	3157
Open	3105
Like	2914
Love	2913
Dare	2844
Lift	2159
Change	2154
Want	1811
Take	1412
Go	1267
Say	1237
Put_on	6083
Take_off	4229
Pull_down	1930
Pull_up	1625
Have_on	1214
Take_of	731
Get_on	728
Lift_up	690
Put_in	507
Dress_up	433
See_without	371
Look_in	336
Put_down	312
Look_like	295
Change_into	274

Table 2
Top 10 nearest neighbors (cosine distance) of the word “pussy”.

Term	Distance	Count	#Broadcasts	#Users
Pusy	0.828122	956	513	432
Pus	0.768473	416	305	259
Pushy	0.741119	267	185	158
Bussy	0.799563	209	128	100
püssy	0.810713	198	133	101
Puzzy	0.753680	195	122	113
pûssy	0.781377	184	110	90
Pussycat	0.818996	169	138	141
Pissy	0.702024	160	141	142
Pssy	0.812888	135	103	79

Table 3
Top 10 nearest neighbors (cosine distance) of the word “boobs”.

Term	Distance	Count	#Chatrooms	#Users
bobs	0.752709	14728	5720	5754
boos	0.756812	670	490	444
booms	0.759904	638	305	189
boobes	0.868892	578	315	182
bobbs	0.794095	494	341	292
boops	0.803665	452	276	177
boody	0.784702	400	285	190
boobz	0.858590	389	256	161
bobss	0.787802	267	175	113
boobd	0.896997	159	146	150

Fig. 5. Most frequent semantically-similar words to LIWC sexual terms, as learned by FastText.

Fig. 6. Top collocates of sexual words in LiveMe dataset.

techniques can easily circumvent them by introducing noise such as typos, grammatical errors, uncommon abbreviations and out-of-vocabulary words (Papegnies et al., 2019; Hosseini et al., 2017). To determine whether this is relevant in our dataset, we use Facebook's FastText library (Bojanowski et al., 2017) to train subword-informed word representations on the LiveMe chats, which we then leverage to identify the semantically-similar adversarial misspellings of filtered terms (such as pussy, boobs, dick, etc.), by querying their nearest neighbours. Our results indicate that indeed this is the case in LiveMe chats, as illustrated in Tables 1–3. To illustrate the sexual word misspellings better, we plot the word cloud of the closest neighbours for the relevant LIWC terms in Fig. 5.

Next, to understand the contexts where the aforementioned terms are used, we plot the word cloud of their top collocates in Fig. 6. We notice that the 3 most frequently collocated words are the verbs *show* (13, 329 collocation occurrences), *open* (3032 collocation occurrences), and *see* (3028 collocation occurrences). To further investigate the imperative meaning of such words in the context of the grooming problem, in Fig. 7, we plot the top collocates in the whole dataset of chat messages for the most frequent one: `show` (203, 230 total occurrences), clearly indicating the existence of sexually predatory behaviours. Similarly, the word `open` is most frequently collocated with words denoting positive politeness (such as please, plz) and endearment (e.g. baby, dear), as well as sexually connoted words, mostly related to clothing (e.g. underwear, clothes, top, shirt, pants, dress), and emojis representing clothing items (e.g., 👙, 👕, 👖, 👗).

While emojis are present in many published datasets, to the best of our knowledge, this is the first study to highlight their relevance in the context of grooming, especially the ones referring to clothing. To this end, we use the same embeddings-based approach as previously described to capture similar clothing terms, see Table 4. Using this

Fig. 7. Collocates of the word "show".

Table 4
Top 10 most frequent clothing terms.

Term	Count	#Chatrooms	#Users
Shirt	36306	19070	16969
Shorts	17449	7635	7635
Dress	12319	7267	6154
Pants	11693	6597	6479
Short	10682	6379	6416
Clothes	10504	5940	5905
Underwear	6055	2490	3022
Bottoms	4768	2997	3272
Bikini	4621	1928	1993
Socks	4563	1855	2581

method, we assemble a list of 300 unique terms, appearing in the chat messages of 45,086 live streams. Next, to examine the intentions underlying these messages, we performed dependency parsing on every chat message the clothing terms appear in, using the spaCy parser (Honnibal and Johnson, 2015). From the extracted parse trees, we collected the simple and phrasal verbs. Table 1 contains the 15 most frequently occurring simple and phrasal verbs in their base forms obtained using spaCy lemmatizer,[6] after removing the verbs contained in the NLTK (Loper and Bird, 2002) stopwords list (e.g. be, can, do, have) to reduce noise in the results. We plot a word cloud of the extracted verbs and verb phrases in Fig. 8. The latter is a clear indication that predators are requesting streamers to perform inappropriate acts involving the removal of their clothes. These findings highlight the imperative nature of the predators' communications related to clothing items.

Additionally, we explore the use of clothing-related emojis[7] in our dataset, occurring in 153,797 chats. We note that 83.6% (128,604) of these messages contain only emojis, without any text. We extract the singular emojis co-occurring with clothing-related emojis since it has been shown that in text messages, emoji sequences tend to have a high level of repetition (McCulloch and Gawne, 2018). Plotting the 10 most frequent emojis, see Fig. 4, we may observe that first came the "backhand index pointing down" (👇) emoji with 28,248 occurrences, followed by the "tongue" emoji (👅) appearing 14,919 times. Considering the high co-occurrence of emojis depicting hand gestures, we speculate that the use of such emoji combinations comprises a novel nonverbal communication pattern adopted by predators to convey to potential victims their requests for sexually inappropriate and suggestive acts, involving the removal of clothes.

6. Topic modelling

In this section, we investigate the extent to which grooming behaviours can be modelled mainly using the textual content of chat messages in broadcasts. To this end, we consider a class of probabilistic techniques called "topic models", comprising a method well suited to studying high-level relationships between text documents.

In this study, all the chat messages sent by users during a broadcast are considered to represent a *document*, similar to the notion of chat log documents; described in Basher and Fung (2014). The topics learned from LDA trained on the chat log documents from our dataset could highlight specific terms associated with latent communication patterns emerging within the broadcasts, that will help us to understand and identify the modus operandi of sexual groomers in the context of SLSS better, by providing meaningful interpretations of different aspects of user behaviour within the chats. Additionally, we investigate the connection of user interactions beyond chatting to grooming, to shed light on the mechanics of sexual predatory behaviours in SLSS.

6.1. Preprocessing

To reduce noise and variation in the text data, we focus only on the chat messages produced by streamers in English-speaking countries appearing in our dataset, i.e. United States, Great Britain, Australia, Canada and New Zealand. This resulted in 209,624 broadcasts, produced by 38,099 unique users. Next, for each of the selected broadcasts, we preprocess the content of each chat message individually, according to the following procedure: First, we apply standard text-normalisation techniques, including tokenization, whitespace trimming, capital-letter reduction, and discarding tokens of lengths > 15 and < 2. Next, provided the prevalence of misspellings related to sexual or clothing terms, as well the abundant use of emojis, we collect the 100 most semantically similar neighbours of each LIWC sexual term, by querying the learned FastText model (417 terms in total), with a single token *SEX_TERM*. We

[6] https://spacy.io/api/lemmatizer.
[7] https://unicode.org/emoji/charts-12.0/emoji-ordering.html#clothing.

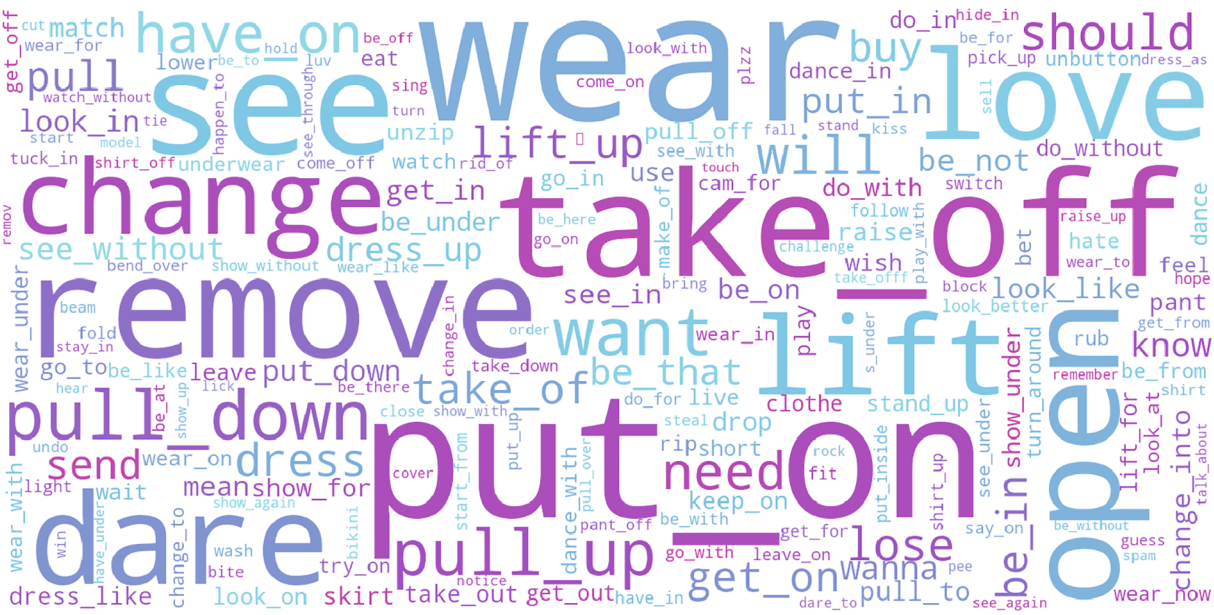


Fig. 8. Verbs extracted from chats containing clothing terms.

repeat the same procedure for the clothing-related terms (see Table 4), collecting 334 terms in total, to which we add the clothing-related emojis, as previously described since they are relevant for our analysis. Any term occurring in the set of clothing terms and emojis is similarly replaced by a single token, namely *CLOTHING_TERM*. To further reduce the noise of the chat data, we repeat the same by substituting the 100 most semantically similar neighbours of the words `show` and `open`, which as previously discussed comprise the top collocates for sexual terms. Furthermore, we remove English stopwords defined in the NLTK (Loper and Bird, 2002) stopwords list, and we additionally detect and remove *gibberish* text, i.e. character sequences that do not reflect a real word, but they are like a random compilation of characters instead. This is a typical spamming behaviour, e.g. misbehaving users clogging online communication channels with gibberish (Yin et al., 2009). More precisely, the detection of gibberish strings is handled by a software library by Rob Neuhaus,[8] implementing a two-state Markov chain which learns how likely two characters of the English alphabet are to appear next to each other. The training of the model is done on a large-scale corpus consisting of English texts available on the Project Gutenberg.[9] A previous study (Doll et al., 2019) assessed the gibberish detection performance of the library and reported an F1-Score of 0.90, which we consider sufficient for the scope of this work. For the remaining words found in each chat message, we obtain their base forms using the spaCy lemmatizer. Finally, we discard broadcasts with less than ten messages, since short texts usually contain few meaningful terms. Thus, the word co-occurrence information is difficult to be captured by conventional topic models like LDA (Hong and Davison, 2010; Zhao et al., 2011). After following these steps, our dataset was reduced to a total of 64,104 broadcasts (30% of all streams produced by English-speaking broadcasters).

6.2. LDA models

For training LDA models, we employed the implementation provided by Machine Learning for Language Toolkit (MALLET).[10] To obtain the most coherent topic model for our data, we vary the number of topics k from 5 to 50 with a step of 5, and train LDA models with 1,000 Gibbs sampling iterations and priors $\alpha = 5/k$ and $\beta = 0.01$. For each trained model, we compute the $C_v(k)$ metric using the implementation provided by Gensim library (spsacctoremoveAPAŘehuřek and Sojka, 2010). We find that $k = 20$ is the optimal topic number according to the C_v metric ($C_v(20) = 0.52$), see Fig. 9.

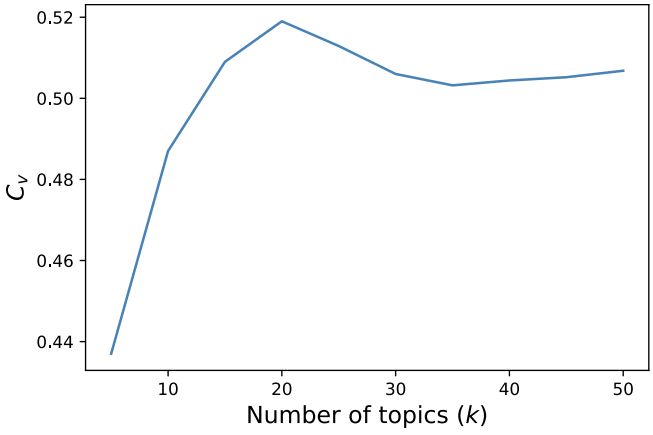


Fig. 9. C_v metric according to no of topics.

In Table 6, we present the topics learned by our best LDA model, including the most relevant terms describing each topic and the number of chats where each topic is dominant. To obtain the most descriptive terms for topic interpretation, we adopted the approach of ranking individual terms within topics presented in Sievert and Shirley (2014) and set $\lambda = 1$.

6.3. Topic interpretation and analysis

From Table 6, it is evident that the most prevalent topic across all broadcasts in our LDA experiment is topic #18, dominating the topic mixture proportions in 12,209 chat log documents (19% of the modelled documents). This topic is clearly related to sexual grooming, with key terms including *CLOTH_TERM, show, open, SEX_TERM*, and various other relevant terms previously identified in our grooming behaviour analysis (e.g. remove, wear, top - which in this case refers to a clothing item,

[8] https://github.com/rrenaud/Gibberish-Detector.
[9] https://www.gutenberg.org/.
[10] http://mallet.cs.umass.edu/.

Table 5
Top 5 interaction features relevant for characterising grooming broadcasts.

Rank	Feature	MDI
1	New followers to viewers	0.36
2	Likers to viewers	0.16
3	Total new followers	0.10
4	Chat messages per user	0.10
5	Total chat messages	0.05

etc.).

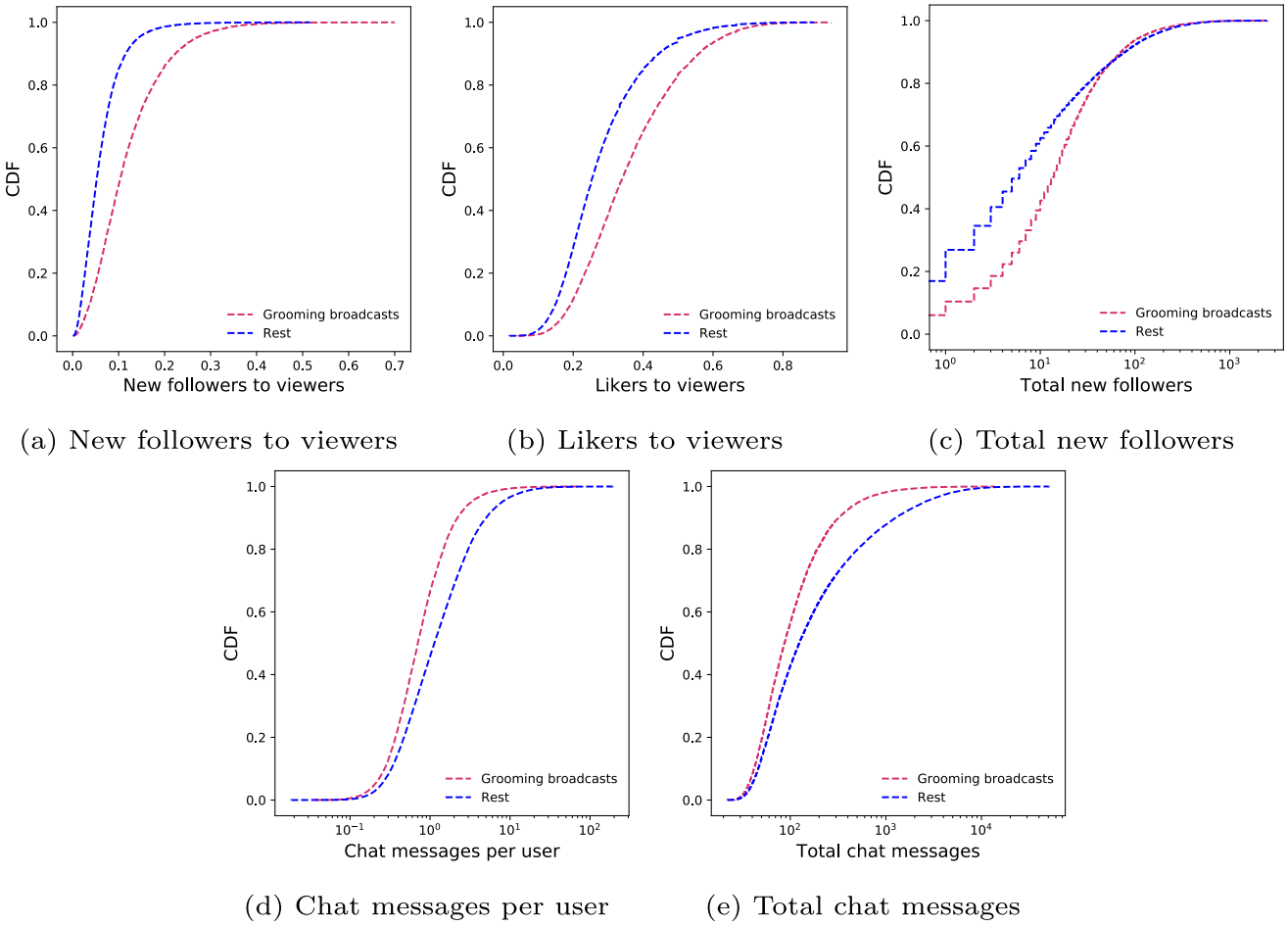
The second most dominant topic (#11) reflects flirtatious behaviours, including many endearment terms (e.g. love, nice, pretty, kiss, cute, gorgeous, hot), and words associated with appearance features (e.g. eye, lip, hair, smile, tattoo). Topic #11 is the most representative of 8,477 chat log documents (13% of modelled broadcasts). The rest of the topics describe a wide range of behaviours occurring in the context of live streams, including virtual currency and gifts of LiveMe (i.e. coin, coindrop, castle, diamond, wand), dancing, singing, eating, social media, etc. An interesting observation is the emergence of a topic containing mostly Spanish words (topic #2). We speculate that a proportion of the US viewers are using Spanish to communicate within the broadcasts, something we did not consider in the preprocessing stage. It should be noted that Spanish are the second most spoken language in the US and widely used in some states. Nonetheless, provided that it dominates only 2142 chats (3% of modelled broadcasts), we expect that its impact will be negligible for the rest of our analysis.

Next, we assess the degree to which user interactions other than chatting can be characteristic of grooming. For this, we leverage the interaction and metadata features of our dataset. Additionally, we normalise the interaction features by the total number of viewers of each stream, considering them as additional features. Next, we employ the Mean Decrease Impurity (MDI) (Breiman, 2001; Breiman, 2002) measure obtained in the process of random forest growing to assess the importance of the described features for discriminating between the broadcasts where topic #18 is dominant in the topic mixture and the rest.

Table 5,6 reports the ranking of the top five most important features according to the normalised MDI metric. To understand how these features are distributed across the two latent classes of broadcasts, we plot their cumulative distribution functions (CDFs) in Fig. 10. We note that the most characterising feature is the fraction of viewers who started following the broadcaster during the stream (Fig. 10a), which in the case of the broadcasts where the grooming topic dominates is much higher than the ones where it does not. Moreover, in Fig. 10c we observe that only around 6% of the grooming broadcasts have not generated any followers for the broadcaster, while the same is true for 17% of the rest of broadcasts. This behaviour is in line with the findings of Lykousas et al. (2018), where the adult content producers of LiveMe were found to have an exceptionally high number of followers which are characterised by their tendency to follow users who have broadcasted adult content systematically, labelled as *adult content consumers*. A possible explanation could be that in broadcasts where the grooming behaviour is prevalent, broadcasters are coerced into performing sexual acts requested by the viewers, as previously outlined. This could justify why the number of new followers they gain in such broadcasts is significantly higher since the viewers might expect that the broadcasters will stream more nude/adult content in the future, and following them is the only way to be notified when they start a new broadcast. Similarly, the fraction of viewers who have liked a broadcast is higher when the grooming behaviour is dominant (Fig. 10b, which is consistent with the findings of Lykousas et al. (2018) where adult content producers are observed to have received higher amounts of praise than the users found in their ego-networks (i.e. followers and followees). This further exemplifies the predatory behaviour of viewers who use likes/praise to coerce broadcasters into inappropriate acts or reward them when they have achieved their objective. Interestingly, for the *Chat messages per user* and *Total chat messages* features which were also found to be important (albeit considerably less impactful in a classification setting), we observe the opposite behaviour: In grooming broadcasts users exchange fewer chat messages, both per-user and at the broadcast level.

Table 6
Topics.

Topic	Keywords	#Docs
18	CLOTH_TERM, show, open, SEX_TERM, nice, dare, dance, hot, stand, leg, put, kiss, turn, pull, wear, cam, camera, remove, foot, top, snapchat, gift, girl, rub, low, hand, lift, finger, message, tease	12,209
11	Love, nice, pretty, kiss, cute, eye, girl, gorgeous, hot, SEX_TERM, sweet, lip, hair, dear, tattoo, smile, single, dance, friend, number, stand, beauty, lovely, cutie, face, boyfriend	8477
1	Sleep, phone, bed, tired, cool, car, wake, cold, drive, smoke, hour, fall, hear, talk, asleep, stay, high, long, house, game, guess, chill, goodnight, sound, money, iphone, fun, pay	5731
19	Talk, hear, happen, friend, leave, wrong, true, cool, mad, sad, care, sound, hurt, smile, fight, stay, fine, dude, person, funny, break, hard, nice, head, long, boy, army, problem, lose, girl, yep	4432
7	Block, admin, girl, message, leave, report, show, talk, account, kid, creep, young, shut, ban, rude, fake, nasty, send, perv, truth, boy, lie, hater, wrong, police, child, unblock	3328
16	Drink, cat, food, pizza, laugh, eat, funny, face, water, put, chicken, dead, hair, head, challenge, roast, cream, leave, SEX_TERM, apple, taco, chocolate, pet, bob, hand, candy, mouth, cheese, nose	3180
12	Cute, snapchat, send, instagram, rate, clown, dab, hot, number, insta, hair, play, text, friend, pretty, love, put, single, phone, kik, profile, eye, cutie, chat, ghost, boy, girl, fake, girlfriend	3080
3	Send, gift, spam, castle, diamond, share, top, level, broadcast, win, giveaway, broadcaster, wand, number, stream, boat, enter, entry, star, love, feature, join, porsche, coin, awesome, comment, fan	2953
5	Coin, drop, coindrop, follower, send, win, feature, shout, fan, castle, love, wand, dab, thot, number, gift, shoutout, giveaway, diamond, stream, iphone, lag, goal, dude, pumpkin, andy, light, level	2798
4	Song, play, sing, love, voice, rap, singing, amazing, nice, dance, awesome, beat, put, hear, panda, listen, cool, singer, sound, closer, juju, black, job, girl, talent, guitar, boy, heart, hit, drake	2687
8	Love, stream, friend, accent, talk, remember, guess, cool, speak, leave, sleep, skype, long, cute, funny, nice, number, meet, hair, mate, lot, person, dad, class, cat, joke, jenni, kat, join, change	2682
20	Light, turn, gang, love, stay, queen, squad, hit, chill, number, king, slay, fact, thot, level, savage, rock, party, dead, boy, mad, play, homie, ight, lot, black, nun, show, petty, dope, top, sum	2366
2	Hola, mami, como, cute, hermosa, eres, show, spanish, amor, SEX_TERM, pretty, bella, donde, hot, bonita, bien, lip, kiss, speak, tienes, espanol, gorgeous, stand, rico, jada	2142
13	Girl, love, cute, play, blue, twin, pretty, red, hot, black, dance, snapchat, green, pink, makeup, hair, lady, white, friend, cool, color, game, face, team, texas, nice, CLOTH_TERM, batman, favorite	1954
14	Kate, love, kid, nice, awesome, cool, tree, country, santa, dad, boy, level, show, broadcast, send, amazing, hear, wolf, talk, lot, son, king, falcon, grim, happen, stream, matt, house, long, rock	1831
9	Beam, love, lag, send, king, cris, stream, castle, fletch, broadcast, show, level, dude, awesome, nick, game, amazing, feature, remember, joey, gift, beem, roll, diamond, join, happen, rip, rackbar	1663
15	Ready, love, spam, feature, stay, game, number, win, boy, tre, read, letter, chat, duck, turtle, greg, cat, spamme, fun, red, ugh, play, controller, send, coin, hehe, cool, high, comment, gift, party	1027
17	Love, fan, favorite, youtube, shout, meet, dab, channel, pickle, song, shoutout, canada, movie, awesome, fav, twerk, vote, magic, food, subscribe, notice, tattoo, cool, texas, win, vid, hair, ily	908
6	Race, love, family, human, unity, amen, put, country, draw, earth, whiskey, broadcast, peace, block, thre, lucky, princess, spam, britt, join, general, respect, coin, barbie, send, level, lag, brit	642
10	President, kira, criticize, article, essay, literary, loco, fard, natur, fward, lag, foard, riot, ward, folard, kilo, follrd	14



(a) New followers to viewers (b) Likers to viewers (c) Total new followers

(d) Chat messages per user (e) Total chat messages

Fig. 10. Cumulative distribution functions (CDFs) of the features of Table 5.

Table 7
Illustrative chat messages from broadcasts where Topic #7 is dominant. Key terms of Topic #7 are in bold.

Chat message
A predator is a person who asks **kids** to undress in front of the camara
And his bio said he likes meeting **young girls**
Block foot fetish **creep**
Don't show the **creeps** anything
Everyone **report** chat **police**
Leave her alone **creep**
Pervs. This **kid** is like 12
Report that **creep** too the **police**
Report the users asking **kids** to undress; to authorities not LiveMe
Show your **kids**
So if they didn't **ban** people for nuditys you would show?
This needs to be **reported** what sort of sick people are ye. She is only 11
YOUR MOM WILL NOW GET A CALL TO KNOW YOU TALK TO 40 years old **creeps**
You **pervs** are **nasty** as f***
block & **report nasty** stuff
creeps make **kids** do **nasty** stuff
he's following lots of **young girls**
pervs stop asking her to undress
report these pedos to **police** mate
she is a **child** stop asking that
she not leting you **creeps** or sick **perv** seeing her dress or undress ok
she's a **kid**. …**perv**
they can't **ban** you if you delete your video after you show
too **young** this is illegal nd worng lol
try not to undress on stream, it will draw in a lot of **creeps**
you have **creeps** who made you do **nasty** stuff
you look very **young**.there are lots of pedos on here. be careful

6.4. Topic relatedness

In this section, we aim to explore the relatedness of the dominant grooming topic and other topics learned by LDA, which could unveil different aspects of this deviant behaviour, beyond our initial analysis. To this end, we use a frequent itemset mining approach to examine the co-occurrence of prevalent topics within the chat log documents. More precisely, we first selected the three topics with the highest probability in the mixture assigned to each broadcast. Then, we applied the FP-growth algorithm (Han et al., 2004) to discover frequent patterns of size two. In Table 8, we show the 10 most frequent patterns extracted following the described approach. As expected, the top result includes the two most prevalent topics in the mixture. Notably, the second most frequent pattern includes the grooming topic, and topic #7, which contains terms related to the (self) moderation of broadcasts (i.e. block,

Table 8
10 most frequent prevalent-topic patterns.

Topic pattern	Occurrences
(11, 18)	11,150
(7, 18)	5536
(1, 19)	4892
(13, 18)	3843
(1, 16)	3416
(1, 11)	3357
(7, 11)	3338
(3, 5)	3331
(2, 11)	3269
(12, 18)	2987

report, ban, shut), terms indicating young age (i.e. kid, young, child, girl, boy), terms of hostility (i.e. creep, perv, hater), words bearing negative sentiment according to LIWC (nasty, wrong, fake, lie). Moreover, the key term that possibly contributes the most towards the interpretation of this topic is **police**. Thus, we expect this topic to be indicative of the criminal dimension of sexual grooming of minors in LiveMe, a large-scale deviant behaviour, also attested by popular media (Melugin, 2018).

To test this speculation, we manually examined a portion of chat messages from broadcasts where Topic #7 is dominant where the aforementioned key terms appear, and we present some illustrative examples in Table 7,8. What we observe is that a part of the users expresses their discontent and anger towards the predators/groomers and their harassment targeting minors. We argue that the above illustrates the extent of deviant behaviour in SLSS, something that beyond the media is also reported by users in, e.g. their feedback for the app. Moreover, the high ranking of this pair indicates that such phenomena, despite the app's moderation mechanisms, are often and known to many users. Finally, the fourth pair (13,18), beyond the common keywords of both topics, shows that some users request further engagement through other platforms, and a primary phase of praise of clothing and body parts, possibly preceding the grooming phase.

7. Conclusions

Social live streaming services due to the continuous use of live streams and immediate user interaction are continuously expanding their user base. As expected, these platforms have attracted the interest of deviant users which try to exploit the new features on these platforms. Obviously, grooming is not only performed in SLSS, nor it is the only thing done on these platforms. Nonetheless, the different possible user interactions coupled with the live streaming nature, create a novel and less explored field.

This work performs an in-depth analysis of the chats of thousands of users and identifies characteristics of the grooming behaviour in the verbal and non-verbal context. To facilitate further research in the field, we responsibly share a massive dataset and provide ethical and legal justification for the collection and processing of such a dataset. Moreover, we illustrate in an automated way how users bypass the word filters that service providers use in their platforms. We also highlight the importance of emojis for the first time in the context of grooming. Finally, our work illustrates that more deviant behaviours may be performed on these platforms.

We believe that this scientific work constitutes a significant contribution towards understanding the deviant behaviours on social networks. The latter should be considered in the light of the role that social networks have in our daily lives and the potentials that the emergence of SLSS have. Our work implies that additional risks exist due to the inefficiency of current moderation mechanisms. Therefore, further measures must be taken to secure the content of what is broadcast, from whom, and to whom. Undoubtedly, due to the size and rate of exchanged information, moderation mechanisms may be difficult to be performed in real-time. However, our work illustrates how deviant behaviours can be detected effectively without resolving to the use of multimedia which require heavy processing. Therefore, we believe that the grooming and other predatory actions will be soon identified better and addressed more effectively by service providers.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Nikolaos Lykousas: Conceptualization, Methodology, Investigation, Software, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Constantinos Patsakis:** Methodology, Investigation, Data curation, Validation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the European Commission under the Horizon 2020 Programme (H2020), as part of the projects CyberSection 4Europe (https://www.cybersec4europe.eu) (Grant Agreement No. 830929) and *LOCARD* (https://locard.eu) (Grant Agreement No. 832735). The authors would also like to thank NVIDIA Corporation for their GPU donation supporting their research.

The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

References

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 391–402). Springer.

Basher, A. R. M., & Fung, B. C. (2014). Analyzing topics and authors in chat logs for crime investigation. *Knowledge and Information Systems, 39*(2), 351–381.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research 3 (Jan), 993–1022.

Bogdanova, D., Rosso, P., & Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. *Computer Speech & Language, 28*(1), 108–120.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L. (2002). *Manual on setting up, using, and understanding random forests, v3* p. 1). CA, USA: Statistics Department University of California Berkeley, 58.

Cambria, E., & White, B. (2014). Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine, 9*(2), 48–57.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing, 72*(7–9), 1775–1781.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.

Coletto, M., Aiello, L. M., Lucchese, C., & Silvestri, F. (2017). Adult content consumption in online social networks. *Social Network Analysis and Mining, 7*(1), 28.

Cooper, A. (1998). Sexuality and the internet: Surfing into the new millennium. *CyberPsychology & Behavior, 1*(2), 187–193.

Craven, S., Brown, S., & Gilchrist, E. (2006). Sexual grooming of children: Review of literature and theoretical considerations. *Journal of Sexual Aggression, 12*(3), 287–299.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61–84.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics, 41*(6), 570–606.

Doll, C., Sykosch, A., Ohm, M., & Meier, M. (2019). Automated pattern inference based on repeatedly observed malware artifacts. In *Proceedings of the 14th international conference on availability, reliability and security* (pp. 1–10).

Drouin, M., Boyd, R. L., Hancock, J. T., & James, A. (2017). Linguistic analysis of chat transcripts from child predator undercover sex stings. *The Journal of Forensic Psychiatry & Psychology, 28*(4), 437–457.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics, 64* (2–3), 221–245.

Elzinga, P., Wolff, K. E., & Poelmans, J. (2012). Analyzing chat conversations of pedophiles with temporal relational semantic systems. In *2012 European intelligence and security informatics conference* (pp. 242–249). IEEE.

Westlake, G. B. & Bouchard, M. (2016). Criminal careers in cyberspace: Examining website failure within child exploitation networks. Justice Quarterly, 33 (7), 1154–1181.
Gottschalk, P. (2011). A dark side of computing and information sciences: Characteristics of online groomers. *Journal of Emerging Trends in Computing and Information Sciences, 2*(9).
Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery, 8* (1), 53–87.
Hidalgo, J. M. G. & Díaz, A. A. C. (2012). Combining predation heuristics and chat-like features in sexual predator identification. In CLEF (Online Working Notes/Labs/ Workshop). Citeseer.
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).
Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1373–1378).
Hosseini, H., Kannan, S., Zhang, B. & Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138.
Inches, G., & Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. *CLEF (Online working notes/labs/workshop)* (Vol. 30).
Kontostathis, A., Garron, A., Reynolds, K., West, W. & Edwards, L. (2012). Identifying predators using chatcoder 2.0. In CLEF (Online Working Notes/Labs/Workshop).
Loper, E. & Bird, S. (2002). Nltk: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pp. 63–70.
Lorenzo-Dus, N., & Kinzel, A. (2019). 'So is your mom as cute as you?': Examining patterns of language use by online sexual groomers. *Journal of Corpora and Discourse Studies, 2*(1), 1–30.
Lorenzo-Dus, N., Kinzel, A., & Di Cristofaro, M. (2020). The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. *Journal of Pragmatics, 155*, 15–27.
Lykousas, N., Gómez, V., & Patsakis, C. (2018). Adult content in social live streaming services: Characterizing deviant users and relationships. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 375–382). IEEE.
McCulloch, G., & Gawne, L. (2018). Emoji grammar as beat gestures. In *Proceedings of the 1st international workshop on Emoji understanding and applications in social media*.
McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., & Jurafsky, D. (2013). Differentiating language usage through topic models. *Poetics, 41*(6), 607–625.
McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., & Jakubowski, E. (2011). Learning to identify internet sexual predation. *International Journal of Electronic Commerce, 15*(3), 103–122.
Melugin, B. (2018). Pedophiles using app to manipulate underage girls into sexual acts, sell recordings as child porn. URL:https://www.foxla.com/news/pedophiles-using-app-to-manipulate-underage-girls-into-sexual-acts-sell-recordings-as-child-porn. [Online; last accessed 12-March-2020].
Miah, M. W. R., Yearwood, J., & Kulkarni, S. (2011). Detection of child exploiting chats from a mixed chat dataset as a text classification task. *Proceedings of the Australasian Language Technology Association Workshop, 2011*, 157–165.
Michalopoulos, D., & Mavridis, I. (2011). Utilizing document classification for grooming attack recognition. In *2011 IEEE symposium on computers and communications (ISCC)* (pp. 864–869). IEEE.
Mitchell, K. J., Finkelhor, D., Jones, L. M., & Wolak, J. (2010). Use of social networking sites in online sex crimes against minors: An examination of national incidence and means of utilization. *Journal of Adolescent Health, 47*(2), 183–190.
Morris, C., & Hirst, G. (2012). Identifying sexual predators by svm classification with lexical and behavioral features. *CLEF (Online Working Notes/Labs/Workshop), 12*, page 29.
Papegnies, E., Labatut, V., Dufour, R., & Linarès, G. (2019). Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems, 6* (1), 38–55.
Parapar, J., Losada, D. E. & Barreiro, A. (2012). A learning-based approach for the identification of sexual predators in chat logs. In P. Forner, J. Karlgren & C. Womser-Hacker (Eds.), CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012, volume 1178 of CEUR Workshop Proceedings. CEUR-WS.org.
Peersman, C., Vaassen, F., Van Asch, V., & Daelemans, W. (2012). Conversation level constraints on pedophile detection in chat rooms. *CLEF (Online Working Notes/Labs/ Workshop)*, 1–13.
Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In International Conference on Semantic Computing (ICSC 2007) (pp. 235–241). IEEE.
Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin.
Quayle, E., Holland, G., Linehan, C., & Taylor, M. (2000). The internet and offending behaviour: A case study. *Journal of Sexual Aggression, 6*(1–2), 78–96.
Řehuřek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). Valletta, Malta. ELRA.
Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408).
Sievert, C. & Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces (pp. 63–70).
Villatoro-Tello, E., Juarez-Gonzalez, A., Escalante, H. J., y Gomez, M. M. & Villasenor, L. (2012). A two-step approach for effective detection of misbehaving users in chats. In P. Forner, J. Karlgren & C. Womser-Hacker (Eds.), CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, Rome, Italy.
Winters, G. M., & Jeglic, E. L. (2017). Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters. *Deviant Behavior, 38*(6), 724–733.
Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. In *Proceedings of the content analysis in the WEB* (pp. 1–7).
Zambrano, P., Torres, J., & Flores, P. (2019). How does grooming fit into social engineering?. In *Advances in computer communication and computational sciences* (pp. 629–639). Springer.
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y. & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. In BMC bioinformatics (Vol. 16, p. S8). Springer.
Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338–349). Springer.