Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.sciencedirect.com/journal/methods-in-psychology

Comparing manual and computational approaches to theme identification in online forums: A case study of a sex work special interest community

Pelham Carter, Matt Gee, Hollie McIlhone, Harkeeret Lally, Robert Lawson *

ARTICLE INFO

Keywords:
Sex work
Online forums
Language, gender and sexuality
Mixed methods
Corpus linguistics

ABSTRACT

Online forums afford individuals opportunities to take part in a community with shared interests and goals. This involves the sharing of experiences and advice (Attard and Coulson, 2012) and can lead to positive effects (Pendry and Salvatore, 2015). Online forums also afford access to rich sources of detailed data, personal experiences, and hard-to-reach or taboo communities. Such online research, though well-suited to qualitative analysis, leads to a number of practical problems in terms of range, depth, and ease of access to data. Even extensive data collection and manual analysis often only engage with a small percentage of the data available in online communities.

In this article, we present a traditional manual collection and thematic analysis of data (2631 posts across 60 different threads, approximately 300,000 words) from forums where sex workers and men who pay for sex discuss matters relating to prostitution. This analysis revealed five themes of forum use: preference sharing, personal narrative sharing, practical advice, philosophical issues, and community maintenance. Further automated data collection and corpus analysis, such as keyness and topic modelling, are presented as a potential innovation within online qualitative research. This approach allowed for the analysis of a larger dataset of 255,891 posts, across 14,232 threads (16,472,006 words), revealing additional themes such as sexual hygiene, desire, legality, and ethnicity, as well as differences in the use of terms of address and slang by punters and sex workers. The automated methods presented allow for more comprehensive investigations of online communities than traditional approaches, but we also note that manual interpretation should still be incorporated into the analysis.

1. Introduction

Online forums are spaces where a vast range of individuals separated by geography and circumstance can socialise, share information, and support one another. Issues with geographical proximity, mobility or the scarcity of other like-minded individuals in the physical locality can be overcome through use of online forums (Rodriquez, 2013). A common reason for accessing a special interest forum is to gain advice and knowledge. This can be, for example, accessing a forum to get specific health information about medical treatments. Sinha et al. (2018) discuss how individuals accessing a forum for chronic coughs would do so for the purposes of gaining additional information about their condition and sharing treatment advice. In a similar fashion, individuals might access non-health related forums to gain information from specialists with experience within that community (Pendry and Salvatore, 2015).

Like offline research, there are online communities that are under-researched. There are also a number of online communities that cater to socially taboo topics and interests, including those wanting to discuss sex work, from both the side of the service user and service provider.[1] Online communities are potentially very attractive for interests that are non-mainstream, due to the relative degree of protection from stigma and social judgement afforded by anonymity. While there has been some research regarding online communities for sex work, this has not always considered the use or role of the forum separate from the broad context of the topic of sex work. For example, Pettinger (2011, 2015) examines the commercialisation of sex work online, as well as the morality of sex work and the use of review forums (Pettinger, 2013). In other work, Holt and Blevins (2007) focus on forums comprising almost entirely of punters,

* Corresponding author. School of English, 3rd Floor Millennium Point, City Centre Campus, Birmingham City University, Birmingham, B4 7BD, UK.
E-mail address: robert.lawson@bcu.ac.uk (R. Lawson).

[1] In discussing the sex work community, there are, as might be expected, questions of terminology (see Stella Montreal, 2013 for further discussion). Those who pay for sex colloquially refer to themselves as *punters* (although *client* and *patron* are also sometimes used), while there are a range of terms used for sex workers, including *working girl*, *escort*, and *prostitute*. In this article, we use the terms *punter(s)* and *sex worker(s)* throughout.

https://doi.org/10.1016/j.metip.2021.100065
Received 16 December 2020; Received in revised form 2 July 2021; Accepted 10 July 2021
Available online 14 July 2021

while Castle and Lee (2008) concentrate on a number of online sites used by sex workers to advertise their services. Qualitative research in the area of online sex work forums could provide more detail about the experiences of both punters and sex workers, groups that are traditionally hard to access due to the stigmatised nature of paying for sex.

While qualitative research into online communities has great potential, there are a number of practical considerations that limit access and analysis (Carter and Kondor, 2020). For example, most online forums comprise of several million words worth of user-generated data, but while this far outstrips the volume of data provided in traditional face-to-face interviews or focus group approaches, the intensive nature of collecting, cleaning, and analysing such data typically means that forums are not often considered in their entirety. The criteria around what is selected to form part of the sample also has the potential to be contentious and open to selection bias. When the data already exists, clear justification for some data being selected over the rest is needed. Finally, timeframes or collection windows are frequently used to create the sample, but these are often arbitrary in nature and will exclude the majority of shared experiences and data that falls outside of this window.

However, with the application of methods from other fields not traditionally used in online qualitative research, greater coverage can be achieved. We can also mitigate issues surrounding the selection of the data. For example, corpus linguistics is an approach that analyses large bodies of text (a ‘corpus’, plural ‘corpora’) using computerised methods to reveal information about the frequency and use of terms (McEnery and Hardie, 2011). The datasets can be extremely large and their analysis can demonstrate differences in style, word usage, phrasing, and topic. This approach has been used to investigate a range of social issues such as political discourse (Orpin, 2005), sexuality (Baker, 2018), and the use and representation of the word *Muslim* in newspapers (Baker et al., 2012).

Methods developed in computer science can be used to scrape large quantities of data from online forums in a fraction of the time taken by manual methods, while the application of topic modelling can provide a starting point for further qualitative analysis that is derived from web-scale datasets. Additional coding of this mass-collected data can also allow for corpus-assisted discourse or thematic analysis. The analysis still retains the core qualitative strengths and applications and keeps the context of participant’s experiences and language use intact (Carter and Kondor, 2020).

In this article, we set out a discussion of how methods drawn from these two disciplines can augment a qualitatively-driven analysis of textual data. In doing so, we argue that the application of these methods affords a number of benefits for researchers working at the intersection of psychology, qualitative methods, and language use, including improved data coverage, depth, and scope.

2. Premise

We present two approaches for comparison. The first is a ‘traditional’ approach to the data collection, coupled with a thematic analysis of data drawn from a sex work forum where punters talk to each other and sex workers. While this approach utilises a large dataset by thematic analysis standards (~300,000 words), it still only represented less than 1% of the data available within the forum at the time.

The second approach utilises techniques from computer science and corpus linguistics, including topic modelling, keyword analysis, and concordance lines. Using this approach, we were able to access a much larger data set (16,472,006 words), which represents almost 100% of the data available within the forum at the time of collection. The corpus analysis approaches the data initially from a quantitative perspective before engaging in further qualitative interpretation and interrogation of the results. Taking these approaches together allows us to interrogate the data holistically, providing a more detailed mixed methods analysis.

3. Approach 1: traditional manual thematic analysis

3.1. Data retrieval and ethical considerations

A highly active UK-hosted web forum that allowed discussion between punters and sex workers was selected for analysis. In line with the BPS guidance (Kaye et al., 2021), the community is not named (see §3.1.2). This community was selected due to its previous inclusion in related research and the high level of activity which generated a rich source of textual data for analysis. 2631 posts across 60 threads were collected from the general discussion board of an online forum for the discussion of sex work (approximately 300,000 words) for a period of one week (July 2015). Considering the active nature of the forum, a data collection window of one week was suggested to ensure the amount of data collected was of a manageable volume. If a thread was active during that one week period (through the addition or editing of a post), it was included in the sample. 619 individual users were identified, with approximately 100 identifiable as sex workers.

The BPS internet-mediated guidelines for research (Kaye et al., 2021) focus on the protection of privacy and dignity for both communities and individuals, integrity, social responsibility, and minimising harm. Of particular relevance for our research are the issues of consent, privacy, and responsibility in the context of reducing harm to the online community and members. Because the research area is sensitive, appropriate steps have been taken to address ethical concerns around consent, privacy, and harm, as outlined in Section 3.1.1 and 3.1.2 below. In doing so, this present work is consistent with aspects of published research in other online community regarding sensitive topics, such as Parkinson’s support forums (Attard and Coulson, 2012), adolescent self-harm forums (Whitlock et al., 2006), and most relevantly, online sex work discussion (Pettinger, 2011).

3.1.1. Privacy and consent

The BPS guidelines suggest that observation of public behaviour should only take place in a public space where there is not a reasonable expectation of privacy. Observation of public spaces would not require active consent to be sought by the researchers if the participants could expect to be observed by strangers. This is a difficult concept to extend to online research and spaces, but in this case the community forum was a fully indexed site on Google, viewable without membership. No sign up was required and no password protected sub-forums existed. While the content might be considered to be of a sensitive nature, it was freely shared in a readily accessible and searchable space. Further to this, the members of the community often took steps to self-anonymise through the use of usernames or self-censorship, indicating some awareness of potential observation.

3.1.2. Responsibility and harm reduction

The BPS guidance is mindful of how traceable information can be online. This can compromise anonymity and confidentiality in research, causing harm to individuals as well as the larger communities of which they are part. This is especially true of research in sensitive areas where there may be limited communities that individuals can join. The ‘outing’ of both individuals and communities can be harmful. Though users frequently employed nicknames or anonymous usernames, further anonymisation occurred with the use of pseudonyms during the analysis. Any personally identifiable information from the forum posts was removed, as were usernames in case they were used across different platforms. In line with the BPS suggestions (Kaye et al., 2021, p. 19), the community has also not been named.

Though the full quotes and context were used for analysis, in line with both the BPS (Kaye et al., 2021, p. 19) and Markham (2012), truncated quotes were used where necessary to reduce identifiable information present. As Markham (2012) suggests, fabrication for the purpose of protecting participants can be ethical practice if the original meaning is maintained. Therefore, this truncation can be considered ethical practice as, although the presented quotes are not verbatim, such truncated quotes still convey the appropriate original meaning and protect individuals from the community from unwanted scrutiny via reverse quotation searches online. As mentioned above, this is also consistent with the practice of paraphrasing quotations founds in research areas such as self-harm online (Whitlock et al., 2006).

3.2. Data analysis

Thematic Analysis was carried out according to the guidelines set out in Braun and Clarke (2006) and Clarke and Braun (2014). For the thematic analysis, a largely inductive approach was taken to ensure the coding and final themes had a strong foundation in the data. This also allowed for flexibility in the potential final themes due to not being aligned to one particular theoretical stance. Rather than group the data by user, the analysis focused on each discussion thread in turn (as a data item roughly equivalent to a traditional transcript) and considered the themes across the dataset as a whole. The intention was to approach the themes at the semantic level. Due to the size of the data sample, NVivo 11 was used throughout to manage the analysis. A physical copy of the data was not used and instead the electronic PDF versions were read and then coded via NVivo 11.

During phase one (reading and familiarisation), the PDFs were read outside of NVivo for ease. During the second phrase (initial coding), coding was conducted line by line (where appropriate) and post by post for each forum thread using the NVivo coding function. NVivo 11 was also later used to check for the coverage of codes across the whole data set when considering potential themes, though the percentage of coverage was not used as a sole consideration. During phase three (searching for themes), an initial thematic map was developed to help identify candidate themes for review. At this stage, four candidate themes were highlighted, with potential room for a sub-theme within one (regarding sharing advice and experiences). During the fourth phase (reviewing themes), the review took place on the extract/code and theme level, with a re-examination of the thematic map for the finalised candidate themes. At this stage the theme of practical advice (theme 3) was separated from experience sharing (theme 2), as it was judged that two distinct functions were occurring here rather than one being a sub-element of the other. For the fifth phase (definition and naming), the themes were checked against the data for coherence and consistency.

3.3. Results

Analysis revealed several themes relating to: Preferences, Sharing Experiences, Practical Advice, Meta/Philosophical Issues, and Shared Community and Cohesion Devices.

3.3.1. Theme 1: preferences

Both punters and sex worker users on the forum often expressed (and interacted on the basis of) their personal preferences, and often by extension the personal reasons or experiences that had led to them. While the most often discussed preference was for the physical characteristics of their partner, other preferences were considered, such as the type of session, location, ethnicity of individuals, price, and even length of booking.

> *"For a holy grail-type punt, ideally she would have that face/eyes, be in her early twenties, about 5′6", medium build, perky natural C-cups, good hip-waist ratio, smooth tum, round bum, shapely thighs and calves, neat feet, so not fussy at all really ….LOL"* [User 1 – Punter]

The sharing of preferences was often linked to sharing of experiences or advice.

> *"I started off fussy, only wanting to go for fairly young (under 35) white women with big breasts who didn't smoke. This lasted all of one punt, after which my curiosity about a French lady in Glasgow overruled my desire for big breasts (and, it turned out, my age limit, as she was nearer 40 than 30.) So I became far less fussy"* [User 2 – Punter]

While the supporting excerpts within this theme suggest a strong overlap with other themes, there was often sharing without elaboration. The purpose of sharing preferences could fill a similar role to that of sharing experiences. For some, it may have led to a reduction in social isolation as others sharing preferences act as an echo chamber of sorts, normalising their own wants and preferences (Merry, 2016; Turetsky and Riddle, 2018). Sharing may also form part of increasing cohesiveness, although the variety of preferences across the two user groups may have been better at reinforcing the need for acceptance of different wants and social norms within what is essentially a unique fringe online community.

3.3.2. Theme 2: sharing experiences

Many users were using the forum to share their own experiences with punting, or were sharing their experiences to reciprocate sharing from another individual. On some occasions the sharing of experiences served other purposes, through the form of a 'me too' narrative. Rather than providing additional information or starting a new thread, individuals would often state that they had similar experiences to other posters within the thread and then confirm the similarity with additional information about their personal experience.

> *"I know exactly what you mean about feeling like people in the street on the way know what you're up to for some reason. I sometimes have a similar feeling on the tube on the way back from an appointment"* [User 3 – Punter]

The purpose of this is potentially two-fold: in the first instance, it may help validate their experiences and actions through shared experiences, a normalisation process of behaviour that is common in other research contexts (Whitlock et al., 2006). As the shared interest of the online community is arguably considered socially deviant, and therefore socially abnormal, the confirmation process of sharing similar experiences can serve to normalise the behaviour within the community, reduce social isolation (McKenna and Bargh, 1998), and lead to the community, in part, acting as an echo chamber (Merry, 2016; Turetsky and Riddle, 2018).

Second, by establishing a common experience that can be attributed to genuine group members (i.e. those members 'in the know'), greater cohesion can be generated within the community. This could be an attempt to establish legitimacy within an online community (Horne and Wiggins, 2009; Stommel and Koole, 2010; Armstrong et al., 2012). By sharing their experience and putting it to the scrutiny of the community, they can justify their inclusion within the in-group, separate to that of the lower status categories of 'lurkers', 'newbies', and 'virgins'.

3.3.3. Theme 3: practical advice

A common theme to emerge from the data was the request and production of practical advice. This advice was shared between and within the two main user groups within the forum and in some cases took the form of offering alternative considerations that might not be readily apparent to the members of one of the sub-groups.

> *"Christ! You don't ask someone if they have herpes!!! Oh my goodness!"* [User 4 – Sex Worker]

Advice was also about the more rudimentary aspects of interactions between punters and sex workers, such as places to stay, appropriate and safe methods of contact, and how not to draw attention or be caught.

> *"There's an increasing trend amongst the budget hotels - whereby the card key for the room door also operates the lift. For a girl in residence, this can*

> *be a problem … the trick is to time your entry to the lift/stairs with others."* [User 5 – Punter]

Some of the advice was broader and related to matters of disclosure and discussion with family members, again with both user groups offering advice of their own, based on both opinion and experience.

> *"My family found out in the most hurtful and appalling way. i would die first before let them find out again … You should always try to protect your parents because they are precious, but at the same time don't treat them like fools."* [User 6 – Sex Worker]

This theme shares a commonality with many other forums analysed in the literature (Attard and Coulson, 2012; Sinha et al., 2018; Pendry and Salvatore, 2015) and highlights the functional aspect of the forum. While there were users who were clearly there to be part of the community, socialise, and share experiences, there were also others who used the forum as a source of information and expertise. Again, this aligns with research such as Whitlock et al. (2006) when considering the dual role of online communities for both normalising behaviour and acting as a source of practical information.

3.3.4. Theme 4: philosophical and meta discussion

Users would discuss matters that were not directly related to sex work experiences or the practicalities of such interactions. There were many instances of sex workers and punters considering issues that were much wider and almost philosophical in nature.

> *"Prostitution is a trade, within the personal-services sector of the economy. Not too different from hairdressing. People who engage in such trades are by definition professionals as distinct from amateurs, because their talents are for hire. They may - or may not - also show qualities that are called professional, such as pride in their work, or setting themselves high standards of service."* [User 7 – Punter]

This covered issues such as whether sex working was a profession, whether paying for sex while in a supposed monogamous relationship constituted infidelity, and issues such as sex as a product.

> *"Some people are content to live with each other because they still love their partner, but there is no sexual attraction for one or either of them. If it is the man, then he can punt (should he feel inclined)"* [User 8 – Punter]

Sex as a consumerist interaction was often referred to, either explicitly or as a personal view within other thread topics.

> *"My advice would be like when 'buying' anything - decide first what you want, in terms of service, location, price range and looks"* [User 9 – Punter]

This was a theme covered by both user groups and aligns closely to previous findings of Pettinger (2013) of punters as 'deserving customers'. Sometimes there would be agreement across the user groups (positive or negatively towards consumerism and sex), suggestive of a 'prosumer' community (i.e., a community with a more dynamic and collaborative view of sex, economics, and consumerism), such as that described by Lahav-Raz (2019). The only major conflicts that occurred between users concerned the name of a paid sex act and a disagreement over a punter's behaviour towards someone they were 'paying' for.

> *"She's not his personal property. Potential abusive mindset developing there imho."* [User 10 – Punter]

Much of the meta-discussion considered etiquette, what was acceptable behaviour, and the boundaries between on and offline users, or indirectly referred to unacceptable behaviour.

3.3.5. Theme 5: Shared Community and Cohesion Devices

Much interaction and content within the forum served the purpose of improving cohesion within the community, establishing norms and limits of acceptable behaviour, and encouraging the involvement and protection of the community. At times this involved the use of in-jokes that relied on either knowledge of the broader offline community and lifestyles (drawing on information more readily available and relevant to actual punters and sex workers), or relied on knowledge of the online community (drawing on historical events or previous interactions with users).

> *"I always give my real name, but as its Jon nobody believes me!"* [User 11 – Punter]

This mirrors findings from Reddit communities discussing illicit drug use (Costello et al., 2017), where humour was used in the maintenance of the community, even when such humour was considered off-topic.

The forum content was also peppered with references to certain outgroups, both online and offline. Another online community was often singled out (due to historical clashes and perceived transgression) and linked to behaviour or examples deemed unacceptable by the community.

> *"That place is just too bizarre for words, I'm constantly amazed at the double standards they exhibit. If it wasn't so sadly pervasive, it would be funny."* [User 12 – Sex Worker]

The offline out-group referred to was that of street girls and punters who frequent street girls. The majority of the forums, sex workers identified as off-street or private/independent escorts and many men shared experiences of frequenting independent sex workers rather than those who worked in parlours or as street girls. Weitzer (2009) highlights fundamental differences in the experiences of different types of sex worker, as well the perceived risks. The polymorphous model of sex work outlined by Weitzer (2009) also suggests differing levels of agency, exploitation, and risk across different types of sex workers. Relating this back to the theme of shared community, it may be that there is a community awareness or acceptance of the differences in agency and risk, leading to a preference for sex workers and situations of reduced risk and increased agency (off street and independent).

Both the links to offline and online outgroups were also often framed negatively in the context of community specific issues that carried weight, such as sexual health and safety. The outgroups were characterised as promoting or accepting unsafe behaviours that ultimately threatened the health of all other user groups and communities.

> *"I will probably get slated for this but £30 a punt whats it coming to. For me thats street prices, plus the fact I would imagine they will get all sorts of undesirables too*." [User 13 – Punter]

Within the context of the in-group community, the forum users, risking of sexual health, and the well-being of those in either sub-group was deemed as being transgressive and deviant by their own established norms. Costello et al. (2017) also report a similar theme of establishing community norms and acceptable behaviour in the face of potential interactions or misinformation deemed to be potentially harmful to the health of individuals in regard to illicit drug use.

4. Approach 2: automated corpus analysis

4.1. Topic modelling

The second analytical approach adopted in this article draws on methods from computer science and corpus linguistics. First, the forum data was collected using technology developed as part of WebCorpLSE (Kehoe and Gee, 2012), a web search engine for linguistic study which

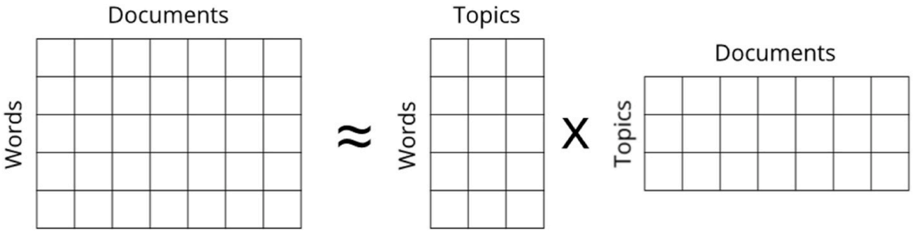

Fig. 1. The NMF approach to topic modelling attempts to find two matrices incorporating topics which, when multiplied, best approximate the distribution of words across documents in the corpus.

incorporates web crawling, text processing, linguistic annotation, search functions, and statistical analysis tools. Limiting the data collection to the general discussion forum sections, the threads and posts were downloaded and stored in structured JSON format data files, including the author, text, date, reply status (indicating whether or not the post was a direct reply), and post count of the posting user. Where replies contained quotes duplicating previous content, the quotes were removed. The resulting corpus contained 14,232 forum threads, 255, 891 forum posts and 16,472,006 words covering the period 2009–2015.

As one of our goals was to find computational methods that paralleled the thematic analysis above, we turned to topic modelling. While this is a relatively novel approach in text-based psychological research, it has been used in a handful of recent studies (e.g. Santilli et al., 2017; Carron-Arthur et al., 2016), as well as being tested in a similar capacity in linguistic studies (e.g. Törnberg and Törnberg, 2016; Murakami et al., 2017; Brookes and McEnery, 2019). A topic model is characterised as a latent organising structure within a corpus. More specifically, each document within the corpus is considered to be a mixture of topics and, where a topic occurs, it is assumed that so does a subset of words related to the topic. Thus, words which are indicative of a topic will co-occur within documents on the topic. Given that the distribution of words across the documents of the corpus is known, the topics can be inferred. In practical terms, a topic modelling algorithm will extract clusters of words and documents from the corpus, where each cluster represents a topic.

The topic modelling approach used in this article is Non-negative Matrix Factorisation (NMF; Lee and Seung, 1999). Describing the algorithm in detail is beyond the scope of the present discussion, but it is important to note that NMF starts with a representation of the corpus in matrix format (which records the frequency of each word in each document) and attempts to find a good alternative representation consisting of two matrices which incorporate topic (see Fig. 1, modified from Kuang et al., 2017: 3). More precisely, we follow the method described in Greene et al. (2014), who have developed a useful process for choosing the number of topics to extract (often the most important parameter to decide on when constructing the topic model).[2] Through this process, we determined 21 topics to be appropriate for our corpus. The words most strongly associated with each topic are shown in Table 1, alongside labels that have been applied to each topic following manual inspection.

The summary labels shown in Table 1 were first established based on the list of top ranked words for each topic. To ensure accuracy and aid in the interpretation of the topics, we extracted examples of the top ranked words from the corpus in the form of concordances (a list of examples of a word presented with the words to the left and right of each occurrence). As it is possible to extract both the words and documents most strongly associated with each topic, we extracted concordances for the top 20 ranked words in the top 100 ranked documents for each topic. This resulted in an average of 7,700 concordance lines per topic – more than would be feasible to analyse manually. Consequently, these concordance lines were placed in random order with the qualitative analysis beginning from the top of the list, inspecting at least 100 concordance lines to determine the validity of the summary labels, and continuing to inspect more concordance lines where the topic was unclear. Through this process, we could verify and/or correct the summary labels associated with each topic.

The topics shown in Table 1 can be seen as representing the most frequently discussed concerns of the community. We describe them in more detail here. It is worth noting that the online community includes both sex workers (mostly female) and punters (mostly male). More detail on how this was determined can found in §4.2, where we investigate differences between community members.

The EMOTIONS/INTERCONNECTION/INTERACTION/OTHER topic was the most problematic. The words most strongly associated with this topic do not appear to be topical, but rather are words found with high frequency in most corpora and are concerned with indefinite and general aspects of everyday life (e.g. *think, know, see, get, time, say, want, said, feel*). Table 2 shows randomly selected concordance lines from this cluster (organised alphabetically by node word), and while no overarching topic emerges, themes relating to internal feelings (5, 6, 12, 16), interpersonal relationships (2, 7, 8, 9, 14, 15, 20) and aspects of offline and online interaction (such as politeness; 1, 3, 4) can be identified. These themes are also present in the remaining topic clusters (discussed below), which we found more clearly topical in their nature and easier to verify using concordances.

Three topics relate to the maintenance of the community and website. The NEW MEMBERS topic consists of forum threads dedicated to welcoming new members and allowing them to introduce themselves to the rest of the community. The other two (FIELD REPORTS and REVIEWS/FEEDBACK) concern the reviews (also called *field report(s)*; abbreviation *fr (s)*) punters submit about their encounters with sex workers. Community members discuss how to write and interpret the reports, as well as the practical aspects of maintaining the field report database. We also find community members discussing and making comparisons to another sex work website that also includes reviews.

Five of the topics relate to preferences punters have in the sex workers they visit. This includes PROFILE PHOTOS/LOOKS, AGE, ETHNICITY/DISCRIMINATION, PUBIC HAIR, and CLOTHING. The punters comment on the influence the profile photos have when choosing a sex worker, issues arising where profile photos are thought to be fake, and their preferences in terms of body shape, breast size, and natural/fake appearance. Preferences regarding age are often discussed in relation to the punters' own age, with a preference for sex workers younger than themselves being common. In the discussion of ethnicity, we see that some punters have their own racial preferences, but concerns over discrimination and racism are also raised. This includes instances where it is claimed a sex

[2] In terms of technical details, the NMF implementation used is from the SciKit Learn Python package (Pedregosa et al., 2011) with the deterministic Non-negative Double Singular Value Decomposition initialiser (Boutsidis and Gallopoulos, 2008). As is typical for NMF, TF-IDF and document length normalisation were applied to the document-term matrix. All words were converted to lowercase and those occurring in a stopword list were excluded. The stopword list used includes common grammatical words (e.g. *the, of, to, a, and*), with the goal of retaining only content words (adjectives, adverbs, nouns and verbs).

worker will reject punters based on race, as well as community members calling each other out for making racist statements. Discussions regarding pubic hair and clothing are again centred around physical preferences and the effect they have on the punters' experience, although, in addition, punters also discuss their own hair grooming regimes.

In the topics PRICE, LOCATION/HOTELS, LOGISTICS/DURATION, CONTACTING/BOOKING, NAMES/ANONYMITY, and AGENCY/PARLOUR/INDEPENDENT, we observe community members seeking advice on the practical elements of soliciting sex workers. In discussions regarding location and the use of real names, for example, punters show their concern with separating soliciting sex from other aspects of their lives, with many showing a preference for using hotels and only divulging their first name. Specific terminology is used (by both sex workers and punters) to distinguish between *outcalls* (where the escort visits a location determined by the punter) and *incalls* (where the punter goes to a location determined by the escort), whether these be homes or hotels. The discussion of names also concerns the names of sex workers, with some punters sharing their reasons why knowing the real name of a sex worker is preferred and/or what they like in a fake name, as well as escorts sharing some of the practical reasons why they might use fake names. The forum members also discuss the (dis)advantages of agencies and parlours in comparison to visiting independent sex workers, as well as sharing their experiences visiting (or working in) both. In general, the practical considerations revealed in these topics show aspects of the commoditisation of sex work, especially in the discussion of the price, which is framed in relation to the quality or variety of the services offered and the "market" as a whole.

The LEGALITY/CRIME topic reveals the different levels of certainty between forum members concerning what is or is not legal regarding sex work, as well as some members arguing why or how it should be safely legalised. Punters also draw a number of comparisons between individual sex workers and brothels, with concerns regarding the links between brothels and human trafficking being raised. References to the police are associated with trust, with both sides of the argument (can/cannot trust the police) represented in the discussion.

SEXUAL HEALTH and HYGIENE are also discussed by the community. Concerns over hygiene are focussed mainly (but not exclusively) on the sex workers, in particular where a sex worker may see multiple clients in one day. This ties into discussions on sexual health, which centre around the perceived risk of a variety of sexual acts. In this topic, we see two highly frequent abbreviations - *bb* ('bareback', i.e. intercourse without a condom) and *owo* ('oral without', i.e. oral sex without a condom). Unprotected sex is the main sexual health concern of punters, with multiple attitudes to risk being represented. A number of members take the position that any kind of unprotected sex is too risky to consider, but others are willing to take the risk to enhance their experience. A hierarchy of risks emerge in some discussions, with unprotected oral sex categorised as less risky, intercourse as risky, and anal sex as the riskiest practice. The punters also express concerns with seeing escorts who actively advertise bareback as an option, judging such escorts negatively, but also noting that this has become increasingly widespread and normalised, with the demand for unprotected sex causing more escorts to take this risk.

Connected to sexual health and risk, we also found a topic relating to SEXUAL ACTS/DESIRES. Under this topic, we see punters (and some sex workers) discussing the kinds of sex acts in which they like to engage. Unprotected oral sex is often mentioned as desired or required by punters, as represented by the terms *owo* and *cim* ('cum in mouth') and we noted above how the risk involved in these practices may be normalised by the community. In addition, we see that many punters desire additional experiences, such as *kissing*, *dfk* ('deep French kissing'), and *gfe* (the 'girlfriend experience'). The girlfriend experience is defined in various ways by the punters, often with an emphasis placed on intimacy and claimed authenticity. The expectations of the girlfriend experience may include French kissing and cuddling, or that it also includes oral sex. A number of punters also express a desire to engage in cunnilingus, typically referred to as *ro* ('reverse oral') on the forum. These desires can be seen as an attempt to capture experiences that may be missing from the punters' romantic lives, as can be seen in the following topics.

In the RELATIONSHIPS/MARRIAGE topic, punters discuss soliciting sex in the context of their other relationships. This includes their own justifications for soliciting sex workers while in a long-term relationship, such as a lack of sex in the relationship, wanting to engage in sexual practices that they feel they can't do with their partner, or looking for a wide range of sexual experiences with more people. Some comments draw a distinction between male and female libido, arguing that the male sex drive is stronger and that men have urges that need to fulfilled. This is particularly relevant where it is perceived that sex is lacking in a relationship or that sex is being used by the female partner to exert control, furthering dominant ideologies regarding toxic masculinity, 'aggrieved manhood', and even some views shared within online incel communities (Ging, 2019; Menzie, 2020). However, we do not wish to falsely paint a picture of widespread toxicity in this regard. The punters draw on their own, sometimes very different, experiences in these discussions.

The FREQUENCY/HISTORY with which punters solicit sex is also discussed and framed in terms of starting age, number of years, number of experiences, long-term cost, and how often might be considered typical (e.g. times per week, month, or year). Connected with these discussions is the extent to which punters regularly see the same sex worker (a *regular*) or prefer to see a variety sex workers. Furthermore, as can be seen in the EMOTIONS/INTERCONNECTION/INTERACTION/OTHER topic, some discussions centre around friendships or romantic feelings that have developed between punters and sex workers, although these can be problematic when not reciprocated.

4.2. Keyness comparison

The topic modelling analysis also revealed a number of differences in the level to which sex workers and punters engaged in each topic. To increase our understanding of these differences, we turned to a tool widely used within corpus linguistics called 'keyness' (or 'keywords'), in which corpora are compared to find substantial differences in word frequency.

To compare the forum posts made by sex workers with those by punters, the relevant posts needed to be identified. Thus, the forum user data in the corpus was enriched with additional parameters through a process of manual coding. A pilot stage of the coding revealed it was possible to reliably capture the role (sex worker, punter, other), gender, and age (in low granularity categories of 0–19, 20–39, 40–59 and 60+). This information could only be determined from the content of user comments, and so we acknowledge that it represents no more than the members' own presentation of their online identity. Two researchers coded all forum members with more than 100 comments (445 users), showing high inter-coder agreement: 100% for role, 100% for gender and 90% for age. In this article, we focus on the **role** aspect of the data.

Two sub-corpora were constructed based on the coded forum posts, one containing posts made by sex workers (57,541 posts; 4,137,245 words) and one containing posts made by punters (134,863 posts; 7,999,941 words). The keyness analysis applies a statistical measure to compare the word frequencies extracted from the two sub-corpora, accounting for differences in sub-corpus size. While various statistical measures have been proposed to achieve this (see Kilgarriff, 2001, 2005 and Gabrielatos, 2018 for discussions), we employ the measure suggested in Kilgarriff (2009), which is the ratio of the normalised frequencies adjusted by a smoothing factor. This measure indicates how many more times frequent a word is in the corpus under investigation

Table 1
Topic model produced from the forum corpus.

Top ranked words per topic	Summary label
punt punting year years punts girl good regular time girls week now last back day first new month days just	frequency/history
sex women wife men relationship married partner life woman marriage love punting sexual think just want man get relationships feel	relationships/marriage
bb bareback condom condoms risk hiv sex owo unprotected offer health risks offering gum clinic offered safe use wgs tested	sexual health
photos pics pictures face looks size site look photo natural breasts boobs picture girl slim website profile fake gallery pic	profile photos/looks
price prices pay hour charge rate money rates service girls market services paying charging extra offer paid business fee afford	price
kissing cum oral orgasm cim mouth owo tongue girl massage just kiss anal gfe enjoy ro lady cock dfk porn	sexual acts/desires
fr frs report reports review reviews write field negative fr's bad read lady punters writing written list positive good post	field reports
police prostitution brothel law illegal people trafficking work article trafficked uk legal news industry street brothels prostitute workers working prostitutes	legality/crime
just think know see people get time wg say want way client post really good said feel clients money things	emotions/interconnection/interaction/other
welcome hi forum hello new site thanks board hope post forward x enjoy newbie punting london posts looking just fun	new members
age older younger old young mature years ladies year women men ages girls think see early lady late look mid	age
shower clean smell bath wash water soap bathroom showered showers showering hygiene minutes fresh time towel just take wet gel	hygiene
hotel room hotels reception door rooms outcall card home outcalls place lift night bed staff car flat incall house bar	location/hotels
agency agencies girls girl parlour indie parlours indies work london working independent clients reviews know worked good escort escorts experience	agency/parlour/independent
hour time booking hours minutes day bookings mins book overnight booked longer appointment minute half get lady advance early two	logistics/duration
black race white racist asian men girls women see english indian colour racism thai british ee people size guys oriental	ethnicity/discrimination
phone text call number email booking texts calls emails contact get numbers mobile reply answer address ring sim confirm just	contacting/booking
feedback profile site reviews profiles booking girls negative system positive review good number email genuine ladies search see website leave	reviews/feedback (comparison with other websites)
hair shaved shave hairy trimmed shaven bush shaving pubes pubic natural trim smooth pussy stubble prefer down balls skin look	pubic hair
name real names use called surname first know used address using facebook remember details john working number fake give now	names/anonymity
wear stockings heels dress wearing sexy lingerie suspenders clothes knickers shoes dressed outfit lady look love skirt jeans naked pair	clothing

Table 2
Randomly selected concordance lines from the emotions/interconnection/interaction/other topic shown in KeyWord in Context (KWIC) format, organised alphabetically by node word. The node word is displayed in the centre, with a 16 word context shown on either side.

No.	Left Context	Node	Right Context
1	statements is taken to be, by definition, a personal attack. Having said that, I do not	**feel**	restricted in what I post and do not give castigation a single thought
2	someone a lot (In the ‘real’ world, I equate this with work colleagues, you like them,	**feel**	fond of them, but when they leave you very quickly drift apart). A fondness for someone
3	was thinking of booking posted something spiteful, I may well not book her because I would	**feel**	that when I revealed my posting identity, she may feel “oh no, not him”. But I
4	this will be by accident and sometimes by design. I suppose it is fairly easy to	**feel**	a little hurt if someone blasts away at a post you have just made, but that's
5	don't think there was a second when i wasn't the centre of her attention, made me	**feel**	very special indeed. Can't wait to see her again
6	sexual partner. If I am sexually appealing to her then that is absolutely fantastic, but I	**know**	it's not real as soon as I hand over the envelope. If I really like her
7	It might or might not be an act, without punting with the same WGs i wouldnt	**know**	, but i dont concern myself with a WGs private business myself. It often happens if i
8	make you see sense. She didn't engage with you but how do you know she doesnt	**know**	who you are? Now think about yourself for a minute. How do you know she won't
9	a girl a few times and really like her in terms of her personality etc. I	**know**	her real name, we text each other occasionally (on our personal numbers) and get along just
10	her legs up over her head on the bed so she was stuck and could not	**get**	up or move away. She said many times to him please, you are hurting me but
11	kid. i do what i need to do to survive and be good at this job.	**just**	as you do
12	an escort is (which often will be genuine) and now loving they might appear it is	**just**	because they are very good at what they do. I have met some really lovely lasses
13	knows too much about you who may try to expose you if they feel spurned. You	**really**	have to get a hard head to this and be very very careful who you trust
14	anywhere near me again speaks volumes about how insensitive and unaware he is. He'd asked to	**see**	me outside of work 3 times and each time I said no. Why did he even
15	, you said that you felt it was fate to see only her as all plans to	**see**	other girls fell through so why not visit a parlour a couple of times for shorter
16	started to pay for a school trip the second did it through financial desperation at the	**time**	. I do not feel for the first one, I very much do for the second one
17	duty - their life choice means that they are what they are 24/7. There are, I	**think**	, others who have decided that they need the money, for reasons good or bad, and the
18	All very nice but I have to ask. What makes them	**think**	that working girls want to be saved from this life, or that its even a life
19	again? Ok Example: Do we really need to have a war over age? I get you	**want**	truth, and you prefer a 21 yr old to a 40 yr old. But for half
20	wife for example and ex girlfriends - had this way of getting close to me - and the	**way**	they kissed, the way they fucked, looked into my eyes, held me tight - was all inspired

compared to the reference corpus, while the smoothing factor helps to reduce the less useful (but sometimes large) ratios that may be observed for rare words.[3]

The results of the keyness analysis are provided in Tables 3 and 4, which show the top 50 keywords based on the keyness measure described above. The results show predominately heteronormative discourses and many of the differences can be characterised as polarised and stereotypical representations of femininity and masculinity in the posts of sex workers and punters, respectively (see Lawson et al. *in progress* for further discussion of these points).

We also see variation in the terms of address that sex workers and punters use to refer to one another. More specifically, sex workers use *client(s), guy(s), gent(s), gentleman/gentlemen, man/men, chap(s)*, and

[3] A smoothing factor of 100 per million words was used. All words were converted to lowercase and those occurring in the same stopword list as the topic modelling analysis were excluded.

Table 3
Words used more frequently by sex workers than punters. Sorted by keyness. Top 50 shown. Norm. freq. shows frequency per million words in the sex workers sub-corpus.

Word	Norm. freq.	Keyness	Word	Norm. freq.	Keyness
X	952.6	5.17	gentleman	117.5	1.60
Clients	1507.0	4.01	appointments	138.3	1.59
Client	1268.2	3.48	industry	300.7	1.57
Xx	242.2	2.85	job	660.1	1.55
Kisses	264.7	2.84	oh	518.5	1.55
Guys	1747.8	2.72	gentlemen	94.0	1.55
Guy	1260.0	2.48	boys	98.9	1.53
Xxx	177.4	2.36	texts	128.8	1.52
Men	1473.2	2.24	text	345.9	1.51
Lol	855.6	2.20	call	651.6	1.50
bookings	550.1	2.15	silly	155.7	1.50
emails	197.0	1.88	booking	1171.8	1.47
appointment	391.6	1.85	asking	412.4	1.47
gents	116.5	1.79	chap	94.3	1.46
ha	220.4	1.78	willy	74.9	1.46
email	363.0	1.76	person	578.9	1.45
gent	99.6	1.74	worked	262.0	1.43
haha	127.1	1.72	work	1379.9	1.42
calls	204.0	1.66	day	1271.1	1.42
lots	357.5	1.65	wash	112.4	1.42
man	780.7	1.65	thankyou	49.8	1.42
website	419.6	1.64	fanny	82.4	1.42
escorting	206.7	1.62	ask	839.9	1.41
rude	158.1	1.61	hair	285.9	1.41
chaps	97.4	1.61	children	143.3	1.40

Table 4
Words used more frequently by punters than sex workers. Sorted by keyness. Top 50 shown. Norm. freq. shows frequency per million words in the punters sub-corpus.

Word	Norm. Freq.	Keyness	Word	Norm. Freq.	Keyness
punt	1671.4	5.22	pussy	139.8	1.56
punting	1387.4	4.00	quality	152.6	1.54
punts	497.8	3.90	beautiful	174.1	1.53
wgs	1037.8	2.80	attractive	209.3	1.51
punted	199.9	2.68	porn	247.1	1.51
wg	1658.3	2.42	parlours	212.9	1.50
frs	250.3	2.37	girl	2038.3	1.50
punters	1040.4	2.22	recent	128.5	1.49
fr	318.5	2.20	recall	87.8	1.49
punter	810.4	2.07	soho	74.0	1.49
view	455.1	1.96	oriental	70.8	1.48
lady	1749.1	1.79	parties	175.5	1.48
london	424.5	1.78	hr	76.5	1.48
thai	109.4	1.70	overall	66.3	1.47
certainly	466.3	1.67	party	195.0	1.47
bb	197.4	1.65	breasts	75.6	1.47
imo	148.0	1.62	ro	95.6	1.46
ee	121.0	1.62	plan	147.8	1.46
hobby	118.1	1.59	price	333.9	1.45
experience	820.8	1.59	reports	190.4	1.45
dfk	96.4	1.59	positive	154.5	1.45
great	786.8	1.56	excellent	142.8	1.45
seems	486.1	1.56	sps	81.1	1.45
experiences	223.5	1.56	owo	205.4	1.44
interest	226.4	1.56	saw	399.5	1.43

boys, whereas punters use *wg(s), lady*, and *girl.* There is some parallel usage across both groups, including polite terms of address (*gent* and *lady*) and youthful terms of address (*boys* and *girl*). However, the preferred terms used by punters are slang terms, with references to sex workers being abbreviations of *working girl* and referring to themselves as *punter(s)*, as well as the act of soliciting sex as *punting.* This more informal language use can be viewed as in-group lexis which, through its consistent use, helps facilitate and consolidate community ties.

The keyness analysis also shows that sex workers more frequently engage in interactional politeness strategies, as shown by terms relating to laughter (*lol, ha, haha*) and terms that show affection (*kisses, x, xx, xxx*). These terms could be representative of gender differences (see Hall, 1995; Lakoff, 1975, 1990), generational differences (sex workers tend to be younger than the punters; see McSweeny, 2018), or may be designed to add an extra element of intimacy (see Grover, 2015: 49). Further research in this dimension is required to uncover the nature of inter sex worker and punter communication within the community.

In comparison, punters tend to use a large number of abbreviations, especially in relation to sexual acts/experiences (*bb* for bareback, *dfk* for deep French kiss, *ro* for reverse oral, *owo* for oral without), ethnicity (*ee* for eastern european), and practical elements (*fr* and *frs* for field report (s), *hr* for hour). Similar to processes of lexicalisation, this is indicative of a need the punters have to enable discussion of these aspects of sex work on a regular basis. The abbreviations may also serve to normalise the discussion of topics which would otherwise be considered taboo. In addition, a difference can be observed regarding the words used for sexual anatomy, with punters using the terms *pussy* and *breasts*, while sex workers use more informal – even childlike – terms, such as *willy* and *fanny*. Again, this may help sex workers trivialise taboo subjects in a light-hearted manner when interacting with other forum members.

Relating the results back to topics, we see punters using terms concerning sexual and partner preference noted in the topic modelling results above, including *bb, dfk, ro, owo, thai, ee*, and *oriental.* One additional aspect of this is *porn*, the discussions of which involve comparisons between sex workers and porn stars regarding looks and the sexual acts that punters prefer to experience during sex opposed to those they prefer in pornography. In addition, the punters frequently refer to soliciting sex as a *hobby*, although the accuracy of this is debated by the punters in some threads. In this respect, it is sometimes compared to other hobbies – normalising the practice to some extent – with aspects such as time, cost, and legality also being discussed in this context. Connected to this are terms like *price* and *hr*, noted in the topic modelling analysis as relating to discussions of value for money and market rates. In contrast to a hobby, the topics revealed by the sex workers' keywords relate to practical and business focussed matters, using terms such as *industry*, *work*, and *job*, and posting about aspects such as *booking(s), appointment(s), email(s)*, and *call(s)*.

5. Discussion

The goals of this article have been threefold: first, to investigate the priorities of an online community where sex work was discussed openly by those who engaged in it; second, to test the suitability of large-scale text analysis methods drawn from computer science (i.e. topic modelling) and corpus linguistics (i.e. keyness and concordancing) as complimentary approaches to the study of textual data to manually conducted thematic analyses; and third, to highlight how these innovations in textual analysis can uncover a level of detail that is difficult to achieve with manual approaches. In this final section of the article, we briefly summarise the two approaches adopted in our analysis, before ending with a discussion of the potential limitations and opportunities afforded by corpus-assisted methods.

In terms of approach 1, our initial manual data collection and resulting thematic analysis produced five themes, contributing further knowledge around community preferences, norms, and out-group perceptions. Both user groups contributed to discussions around these core themes. While the themes revealed by the traditional thematic analysis were largely reproduced in the topic modelling, additional themes were revealed that were overlooked in the manual approach. This includes topics relating to sexual hygiene and desire, relationships and marriage, crime and legality, and ethnicity and discrimination. Ultimately, the topic modelling produced a more detailed representation of the community's priorities and lived experiences. Moreover, the keyness analysis helped to establish how these differed for sex workers and punters, as well as revealing additional topical elements, further contributing a wider range of themes and considerations not typically found in past research into sex work online.

Taken together, the automated methods outlined in this article provide an objective and quantitative basis upon which to build an analysis of textual data, allowing more data to be processed than is feasible in a manual thematic analysis. Indeed, a common concern when increasing the sample size in such research is that, for practical reasons, some of the fine-grained detail and context would be lost. We would argue that a corpus-assisted approach allows researchers to engage with a wider range of data, enhancing the level of detail and coverage, all while maintaining the original context. And although there is certainly a learning curve in adopting some of the methodologies outlined in this article, this is offset to some extent by the lower effort needed at the data collection stage and any subsequent first-pass analyses.

It should also be noted that the process of checking the topic modelling and keyword results through the inspection of concordances, although time consuming, turned out to be an essential step in ensuring the accuracy of our interpretation of the initial results. One illustrative example of this was the emotions/interconnection/interaction/Other topic, where it was impossible to achieve any interpretation without inspecting the contexts of the words selected. The process of inspecting concordances following a keyness analysis is typical, arguably even essential, in corpus linguistic studies. Taking a similar approach to the results of topic modelling is also of merit.

While our findings can be considered a preliminary exploration of a community discussing sex work, the experiences of, and stances taken by, the members of the community regarding these concerns warrant further attention. For example, the formulation and normalisation of risk in relation to sexual health, punters' desires, and pressures exerted on sex workers to offer risky services in a competitive 'market' has implications for public health, while further investigation of discussions about personal relationships outside of sex work and justifications for engaging in it can contribute to the understanding of sexual desire, heteronormativity, and masculine identity. It is also important to note that a critical perspective concerning agency, inequality, exploitation, and misogyny is lacking in the present discussion. Finally, we are aware that our own individual positionalities and personal backgrounds will also have an impact on the perspectives explored in this article. As a recent report by the Sex Workers' Rights Advocacy Network notes, "sex work research often reflects the biased views of health and social workers, researchers, and policy makers" (SWAN, 2019, p. 23). While we have tried to let the data speak for itself and explored methods to enhance objectivity, textual interpretation will always be influenced by the lived experiences and histories of those doing the interpretation. It is our intention to address all of the issues identified here more fully in future work (see, for example, McIlhone et al. forthcoming).

Taking all of this together, this article has suggested some productive synergies between qualitative research methods and corpus-based approaches. By utilising a blended approach in our analysis, we were able to explore a much larger body of data and uncover more comprehensive patterns in usage, drawing on topic modelling, concordancing, and keywords. More broadly, in demonstrating some of the ways in which corpus methods can be used to interrogate online forum data from both linguistic and psychological perspectives, we hope that other researchers will see the utility of such methods as robust ways of tackling large-scale textual analysis while retaining the benefits of fine-grained qualitative approaches.

Credit author statement

Pelham Carter: Conceptualization, Methodology, Formal analysis, Writing – original draft, **Matt Gee**: Software, Methodology, Formal analysis, Writing – original draft. **Hollie McIlhone**: Data curation, Formal analysis. **Harkeeret Lally**: Data curation, **Robert Lawson**: Conceptualization, Funding acquisition, Project administration, Writing – review & editing.

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 article.

Acknowledgements

We would like to thank audiences at the 22nd Annual CyberPsychology conference, the 26th Lavender Languages and Linguistics conference, and the 2nd Approaches to Digital Discourse Analysis conference for their comments, questions, and suggestions, all of which helped improve the final article. We would also like to thank the anonymous reviewers for their careful feedback on earlier versions Andrew Kehoe for proof reading, and the editors for their support throughout. The writing of this article was supported by funding from Birmingham City University's STEAM grant.

References

Armstrong, N., Koteyko, N., Powell, J., 2012. 'Oh dear, should I really be saying that on here?': issues of identity and authority in an online diabetes community. Health 16 (4), 347–365.

Attard, A., Coulson, N.S., 2012. A thematic analysis of patient communication in Parkinson's disease online support group discussion forums. Comput. Hum. Behav. 28 (2), 500–506.

Baker, P., 2018. Language, sexuality and corpus linguistics. J. Lang. Sexuality 7 (2), 263–279.

Baker, P., Gabrielatos, C., McEnery, T., 2012. Sketching Muslims: a corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. Appl. Linguist. 34 (3), 255–278.

Boutsidis, C., Gallopoulos, E., 2008. SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recogn. 41 (4), 1350–1362. https://doi.org/10.1016/j.patcog.2007.09.010.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101.

Brookes, G., McEnery, T., 2019. The utility of topic modelling for discourse studies: a critical evaluation. Discourse Stud. 21, 3–21. https://doi.org/10.1177/1461445618814032.

Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., Griffiths, K.M., 2016. What's all the talk about? Topic modelling in a mental health Internet support group. BMC Psychiatr. 16, 367. https://doi.org/10.1186/s12888-016-1073-5.

Carter, P., Kondor, K., 2020. Researching the radical right: making use of the digital space and its challenges. In: Littler, M., Lee, B. (Eds.), Digital Extremisms: Palgrave Studies in Cybercrime and Cybersecurity. Palgrave Macmillan, Basingstoke, pp. 223–252.

Castle, T., Lee, J., 2008. Ordering sex in cyberspace: a content analysis of escort websites. Int. J. Cult. Stud. 11 (1), 107–121.

Clarke, V., Braun, V., 2014. Thematic analysis. In: Encyclopedia of Critical Psychology. Springer, New York, pp. 1947–1952.

Costello, K.L., Martin III, J.D., Edwards Brinegar, A., 2017. Online disclosure of illicit information: information behaviors in two drug forums. J. Ass. Info. Sci. Technol. 68 (10), 2439–2448.

Gabrielatos, C., 2018. Keyness analysis: nature, metrics and techniques. In: Taylor, C. (Ed.), Corpus Approaches to Discourse. Routledge, Oxon, pp. 225–258.

Ging, D., 2019. Alphas, betas, and incels: theorizing the masculinities of the manosphere. Men Masculinities 22 (4), 638–657. https://doi.org/10.1177/1097184X17706401.

Greene, D., O'Callaghan, D., Cunningham, P., 2014. How many topics? Stability analysis for topic models. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer, New York, pp. 498–513. https://doi.org/10.1007/978-3-662-44848-9_32.

Grover, C., 2015. Intimacy, Power and Pleasure: the Linguistic Construction of Identity in Online Personal-Ads for Casual Sex. Unpublished PhD Thesis. University of Sydney.

Hall, K., 1995. Lip service on the fantasy lines. In: Hall, K., Bucholtz, M. (Eds.), Gender Articulated: Language and the Socially Constructed Self. Routledge, pp. 183–216.

Holt, T.J., Blevins, K.R., 2007. Examining sex work from the client's perspective: assessing johns using on-line data. Deviant Behav. 28 (4), 333–354.

Horne, J., Wiggins, S., 2009. Doing being 'on the edge': managing the dilemma of being authentically suicidal in an online forum. Sociol. Health Illness 31 (2), 170–184.

Kaye, L., Hewson, C., Buchanan, T., Coulson, N., Branley-Bell, D., Fullwood, C., Devlin, L., 2021. Ethics Guideliens for Internet-Mediated research. The British Psychological Society, Leicester.

Kehoe, A., Gee, M., 2012. Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. In: Oksefjell Ebeling, S., Ebeling, J., Hasselgård, H. (Eds.), Studies in Variation, Contacts and Change in English Volume 12: Aspects of Corpus Linguistics: Compilation, Annotation, Analysis.

Kilgarriff, A., 2001. Comparing corpora. Int. J. Corpus Linguist. 6, 97–133.

Kilgarriff, A., 2005. Language is never, ever, ever, random. Corpus Linguist. Linguistic Theory 1 (2), 263–276. https://doi.org/10.1515/cllt.2005.1.2.263.

Kilgarriff, A., 2009. Simple maths for keywords. In: Mahlberg, M., González-Díaz, V., Smith, C. (Eds.), Proceedings of Corpus Linguistics Conference CL2009. University of Liverpool, UK.

Kuang, D., Brantingham, P.J., Bertozzi, A.L., 2017. Crime topic modeling. Crime Sci. 6 (1), 1–20.

Lahav-Raz, Y., 2019. The prosumer economy and the sex industry: the creation of an online community of sex prosumers. J. Cultural Econ. 12 (6), 539–551.

Lakoff, R., 1975. Language and Women's Place. Harper & Row.

Lakoff, R., 1990. Talking Power: the Politics of Language. Basic Books.

Lawson, R. Gee, M., Carter, P., McIlhone, H., & Lally, H. (in progress). Masculinities and heteronormativity within a taboo community of practice. In Lawson, R. (Ed.), Discourses of Digital Masculinities. Cambridge: Cambridge University Press.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788–791.

Markham, A., 2012. Fabrication as ethical practice: qualitative inquiry in ambiguous internet contexts. Inf. Commun. Soc. 15 (3), 334–353. In press.

McEnery, T., Hardie, A., 2011. Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.

McIlhone, H., Lawson, R., Carter, P., Gee, M. & Lally, H. (forthcoming). "If you find any nice local men remember to pass them on to me when you're finished with them": discourses of female sexual desire, partner preference, and motivation in an online sex-work forum. In Borba, R. & Rowlett, B. (Eds.), The Language of Sex Work. Oxon: Routledge.

McKenna, K.Y., Bargh, J.A., 1998. Coming out in the age of the Internet: identity "demarginalization" through virtual group participation. J. Pers. Soc. Psychol. 75 (3), 681.

McSweeny, M.A., 2018. The Pragmatics of Text Messaging: Making Meaning in Messages. Routledge, Oxon.

Menzie, L., 2020. Stacys, Beckys, and Chads: the construction of femininity and hegemonic masculinity within incel rhetoric. Psychol. Sexuality 1–17.

Merry, M., 2016. Making friends and enemies on social media: the case of gun policy organizations. Online Inf. Rev. 40 (5), 624–642.

Murakami, A., Thompson, P., Hunston, S., Vajn, D., 2017. 'What is this corpus about?': using topic modelling to explore a specialised corpus. Corpora 12, 243–277. https://doi.org/10.3366/cor.2017.0118.

Orpin, D., 2005. Corpus linguistics and critical discourse analysis: examining the ideology of sleaze. Int. J. Corpus Linguist. 10 (1), 37–61.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12 (85), 2825–2830.

Pendry, L.F., Salvatore, J., 2015. Individual and social benefits of online discussion forums. Comput. Hum. Behav. 50, 211–220.

Pettinger, L., 2011. 'Knows how to please a man': studying customers to understand service work. Socio. Rev. 59 (2), 223–241.

Pettinger, L., 2013. Market moralities in the field of commercial sex. J. Cultural Econ. 6 (2), 184–199.

Pettinger, L., 2015. The judgement machine: markets, internet technologies and policies in commercial sex. Soc. Pol. Soc. 14 (1), 135–143.

Rodriquez, J., 2013. Narrating dementia: self and community in an online forum. Qual. Health Res. 23 (9), 1215–1227.

Santilli, S., Nota, L., Pilato, G., 2017. A comparison on the use of LSA and LDA in psychology analysis on "courage" definitions. Int. J. Semantic Comput. (IJSC) 11 (3), 373–389. https://doi.org/10.1142/S1793351X17400153.

Sinha, A., Porter, T., Wilson, A., 2018. The use of online health forums by patients with chronic cough: qualitative study. J. Med. Internet Res. 20 (1).

Stella Montreal, 2013. Language matters: talking about sex work. Available from. https://www.nswp.org/sites/nswp.org/files/StellaInfoSheetLanguageMatters.pdf.

Stommel, W., Koole, T., 2010. The online support group as a community: a micro-analysis of the interaction with a new member. Discourse Stud. 12 (3), 357–378.

Sex Workers' Rights Advocacy Network (SWAN), 2019. Nothing about us without us: a brief guide on meaningful involvement of sex workers and their organisations in central-Eastern europe and central asia. Available from. https://www.swannet.org/files/swannet/NothingAboutUsWithoutUs_ENG_web.pdf.

Törnberg, A., Törnberg, P., 2016. Combining CDA and topic modeling: analyzing discursive connections between Islamophobia and anti-feminism on an online forum. Discourse Soc. 27, 401–422. https://doi.org/10.1177/0957926516634546.

Turetsky, K.M., Riddle, T.A., 2018. Porous chambers, echoes of valence and stereotypes: a Network analysis of online news coverage interconnectedness following a nationally polarizing race-related event. Soc. Psychol. Personality Sci. 9 (2), 163–175.

Weitzer, R., 2009. Sociology of sex work. Annu. Rev. Sociol. 35, 213–234.

Whitlock, J.L., Powers, J.L., Eckenrode, J., 2006. The virtual cutting edge: the internet and adolescent self-injury. Dev. Psychol. 42 (3), 407–417.