INTERNSHIP REPORT

User activity and content of online sexual health conversations

Word count: 8656

Kirill Palenov - 2727021
Host organization: Amsterdam School of Communication Research – Universiteit van Amsterdam
University: Vrije Universiteit Amsterdam
Host organization supervisor: Dr. Johanna. M. F. van Oosten
University supervisor: Dr. Maike Tietschert, MSc

Contents

1. Introduction

Nowadays, social media has been used a lot by adolescents. Social networks provide a quick, anonymous toolbox to gain comprehensible information regarding sensitive health issues including sexual health behavior (Sunkara, 2021). At the same time, it is vital to realize that social media mostly delivers user-generated content (UGC) (Nikkelen et al., 2019). UGC stands for the content that users share online and that is created by themselves. As long as users mostly are not experts in the field, the content they produce cannot be verified for trustworthiness and accuracy (Chou et al., 2013). However, personal user experiences posted online, which is part of UGC, are often recognized as a credible source of knowledge (Sunkara, 2021). This encourages society to pay attention to the quality of information adolescents receive on the internet. Young people are more vulnerable to health misinformation spread online due to sophisticated techniques of misinformation, which include, for example, focusing on negative consequences, or propagation of conspiracy theories (Sunkara, 2021).

Zhao et al. (2021) describe misinformation on social networks as “messages that are posted to persuade other users”. Lewandowsky et al. (2012) claim that correction may be an effective way to combat health misinformation. The nature of social media requires correction to be done immediately before the spread of falsified information cannot be undone. This can be achieved through the moderation of the content, which involves “making a decision about the checking and verifying the adequacy of the detected content according to the rules and policies as defined by a particular SM [social media] platform” (Gongane et al., 2022). For example, Wadden et al. (2021) claim that moderation of online mental health communication engages users in the discussion, increases the level of civility, and facilitates positive changes in users' psychological perspective. However, to our knowledge, there is no research that has been carried out before, which could describe the impact of moderation of online sexual health communities. To identify this impact, we aim to compare a moderated by experts Q&A platform designed for adolescents where they can anonymously share their concerns and look for credible advice and a non-moderated forum, where young people are able to gain information from peer users.

To our knowledge, there is no research to describe how users discuss sexual health conversations in online Q&A platforms. As a first step to fill this gap, we aim to inductively investigate and describe online Q&A platforms – online spaces where adolescents can anonymously share their concerns and look for credible advice from peers – and descriptively compare a platform that is moderated by experts with a non-moderated platform.

More specifically, we aim to explore the behavioral features of key users of each platform, in terms of discussion initiation, interaction engagement, influential scope, relational mediation, and informational independence (cf. Zhao et al., 2021). In addition, for these key

users, we aim to investigate the content that they share, in terms of topic, sentiment, emotion, and civility (cf. Wadden et al. 2021; Zhao et al., 2021). This will allow us to gain first insights into the persuasiveness of users sharing sexual health information (departing from the ELM framework on persuasion, Petty & Cacioppo, 1986), and the quality of the information that they share, in moderated and non-moderated online platforms. This can then inform later research and interventions to reduce the spreading of misinformation on such platforms.

2. Theoretical framework

2.1. Framework development

The actual research is based on the two Master theses defended at the University of Utrecht in July 2022. Lai (2022) extensively studied the sentiment of sexual health information spread online on moderated and non-moderated platforms. For the purposes of the current study, moderation stands for the verification and factual correction of the UGC in relation to sexual health and behavior. Schreurs (2022) performed research on the credibility detection of sexual health content. The initial idea for further development was to compare the level of credibility of content shared on moderated and non-moderated platforms. The previous research on the credibility of information spread online (including health information), as well as Schreurs (2022) thesis, is mostly based on linguistic markers, such as the presence of personal perspective in the text, group reference, lexical diversity (number of unique words), pausality (punctuation density), typo ratio, etc. Schreurs (2022) claims some limitations for her thesis, such as the possible insufficiency of linguistic markers for deception detection (Castillo et al., 2011). Indeed, the misinformation identification based on the textual cues can be biased towards those with a lower level of education or the Dutch language. This way, the other features can be used to assess the credibility of information in social media, such as user-based, content-based, and social context-based features (Zhang & Ghorbani, 2020, as cited in Schreurs, 2022). For instance, some community-based approaches for deception detection were proposed by Zhao et al. (2021), hypothesizing that users who spread misinformation produce significantly more content.

Developing the research framework of the study based on the theses by Lai (2022) and Schreurs (2022), the goal of the project is to compare the moderated and non-moderated platforms for the purpose of measuring the credibility of the content circulating on these platforms. However, expanding the scope of the thesis, we encountered the problem of the application of the credibility concept. We departed from an ambitious goal to assess the truthfulness of the content shared on the resources. When elaborating on this concept, it was found that the nature of the content spread on the fora is highly subjective, while being composed mostly of the personal user experience, and thus not verifiable. For this reason,

it was decided to move away from the idea of credibility detection towards a more specific task, namely the identification of quality information (in terms of topics, sentiment, and toxicity) posted on the platform. Schreurs (2022) grounds her research in the misinformation detection field, which contains “a lot of overlapping markers in the detection of credible information”.

At the same time, the theses did not create any room for posing any specific hypotheses. Elaborating on the aforementioned limitations, we directed our framework toward cognitive psychological cues (Kumar & Geethakumari, 2014). Particularly, information perception may vary towards the source credibility level (p. 7). Since the objects of the research are two platforms, which spread UGC, the major disseminators of content are the other users, which do not have any persuasive power by default. However, there is little known about the role and behavior of the major disseminators (hereinafter, key users) in terms of content moderation. Therefore, we address the role of the key users in online sexual health discussions in terms of the content they produce and deliver. Specifically, we adopted the framework proposed by Zhao et al. (2021), which is also mentioned in Schreurs (2022). The framework stands for the complex study of health misinformation detection, which is spread online, and combines behavioral (user-based characteristics, such as profiling (Schreurs, 2022)), content (semantics), and social context-based (network-based (Schreurs, 2022)) traits.

Given the lack of previous research, no hypotheses could be posed. The outcomes of the Lai (2022) and Schreurs (2022) papers do not let us propose any integral metrics, which could be used for the platform comparison and validated through statistical tests. Therefore, all the results should be interpreted as purely descriptive.

2.2. Final framework

The purpose of this study is to describe the behavioral features and content of platforms for sexual health information sharing, and to descriptively compare moderated and non-moderated platforms. We formulate the following research question: “**What are the discourses of online sexual health discussions in terms of users’ activity and content, which is shared through the online platforms, in a moderated and non-moderated platform?**”

Following the research design by Zhao et al. (2021), we aim to start our research with the network analysis of both platforms. Behavioral peculiarities existing in online discussions can be derived from analyzing the ways the users communicate with each other. For this purpose, we intend to consider users as a network and identify the relations they are in, what roles users follow and what are their positions within the discussion (Borgatti et al., 2018). Specifically, we are going to build a network of interactions between users and

analyze their interpersonal communication through Social Network Analysis (SNA). SNA may uncover the mechanics of misinformation and fake news dissemination in social media, as well as the development of public opinion, media trust, and digital content circulation (Ognyanova, 2021). In line with descriptive user statistics, Zhao et al. (2021) introduce some quantitative metrics, such as the N of posts, created by users (discussion initiation), N of comments (interaction engagement), as well as various measures of centrality as a key determinant of users' behavior (influential scope, relational meditation, and informational independence), which are the main variables of SNA. Centrality is a property of the user, which describes their position towards other users. It lets us identify the importance of the user with the network, their influence on the other users, the degree of the leadership of a user in the network, etc. (Borgatti et al., 2018). After creating the network of users, we aim to zoom into the key actors to assess the content they produce for the quality of information and the types of routing they use.

Developing the central-level route of processing information, we aim to extract semantics from the key users' comments using topic modeling. Contributing to the peripheral-level route of processing information, along with behavioral features, we will perform the sentiment analysis of the key users' comments (including emotion detection). We also employ the metrics elaborated by Wadden et al. (2021), which are developed to evaluate the influence of moderation on discourse quality on communication platforms. Particularly, we are going to measure civility based on the toxicity of the comments.

The final framework is presented in Figure 1.

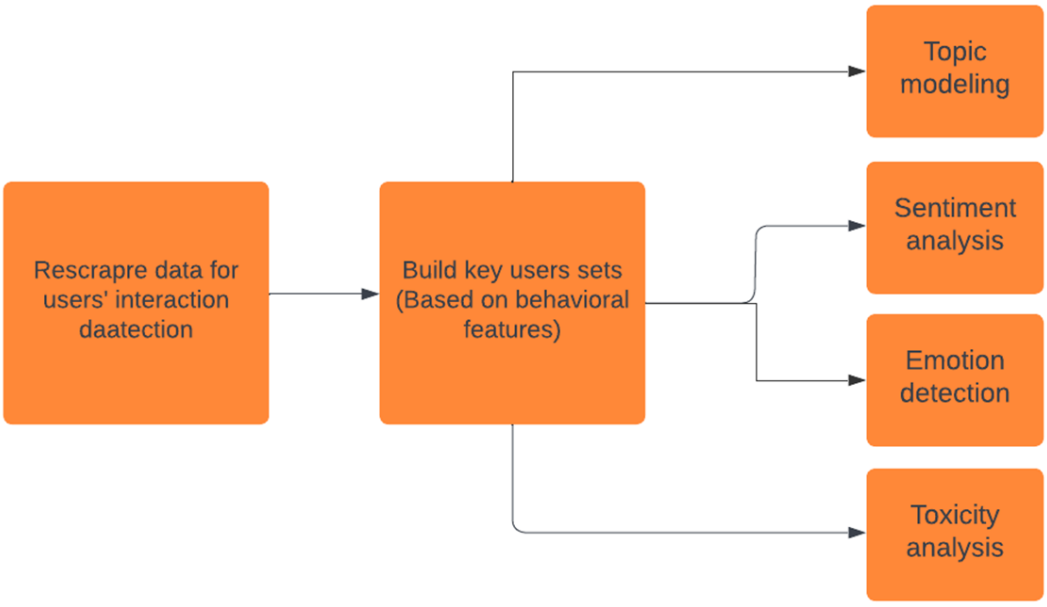

Figure 1. Final framework

3. Data

3.1. Data description

De Kindertelefoon and the FOK! forum are the platforms to be compared during the study.

De Kindertelefoon is a helpline platform where children can ask for advice that they do not intend to obtain in their own environment. The platform also runs a moderated forum where children can anonymously share their questions.

FOK! Is one of the largest non-moderated online communities in the Netherlands with a wide selection of topics to discuss including topics of sexuality. The platform offers more than 25000 threads, which refer to ‘sexuality’.

For both of the platforms, data consists of two datasets: topics dataset (dataset, where all the threads collected from the platforms are stored) and comments dataset (dataset, which contains all the comments from the threads posted within the specific timeframe (see below))

Threads and comments datasets for the both platforms were collected for the other research. Before the pre-registration, no analysis has been conducted related to the research plan (including calculation of summary statistics).

FOK!

The topics dataset for the FOK! Forum consists of the following fields: ID (of the thread), Title, TopicCat (topic category), LastResponse (timestamp of the last post in the thread), NComments (number of comments), NViews (number of views) and URL.

The comments dataset contains the following columns: ID (of the thread), CommentID (ID of the comment), Content (of the comment), and Time (the date and time when the comment was posted).

De Kindertelefoon

The topics dataset for De Kindertelefoon includes the following fields: ID (of the thread), Title, TopicCat (topic category), CreationTime (timestamp of creation of the thread),

LastReplyTime (timestamp of the last post in the thread), NReplies (number of replies), NViews (number of views), NLikes (number of likes) and URL.

The comments dataset is composed of TopicID (of the thread), Content (of the comment), CreateTime (the date and time when the comment was posted), and FirstPost (a flag, indicating whether the post is the first post in the thread)

The language of the content is Dutch. The number of comments for de Kindertelefoon is 87706, and for FOK! is 116962. Comments were posted between 2014 and 2021. The number of threads collected from de Kindertelefoon is 10947, while from FOK! 2053 threads were collected.

Data has been pulled from the Web, is publicly available, and can be obtained by any person to reproduce the analysis to the highest degree.

The comments have been collected from FOK! and De Kindertelefoon using scraping tools (BeautifulSoup Python package). They were extracted directly from the web pages as HTML code and parsed.

The data collection procedure usually requires permission from the participants/data owners to use the data for the research. In the case of research based on online communication, it may be difficult to obtain a permit from each of the members of the community. For this reason, requests to platforms have been made. De Kindertelefoon has granted permission to use their data, while FOK! did not respond. Terms and Conditions of the FOK! Forum does not limit data usage for scientific purposes, thus the permission is considered to be granted.

4. Methods

4.1. Data collection

For the current research, we will rescrape the topic web pages to gain additional information such as the nicknames of the commenters and nicknames mentioned in the quotes within these comments. Such a procedure will be executed based on the direct identifiers of the comments stated in the existing datasets. Therefore, no new comments will be added to the dataset. If the comment was deleted, then its content will be ignored for the purposes of the current research

To create a user's network, the comments were rescraped to gain authorship and the presence of addressing the other users. An author is considered to be a user on the platform. To study the users, their profile data were collected as well.

A traditional way on the fora to reply to someone is to quote the text of the comment, which is being responded to. Each quote is equipped with a link to the author of the original comment. Therefore, the presence of the quote in the comment represents the connection between the comment author and quoted user, which is expressed as an edge of the graph of the users' network. All the quotes referred to one user in one comment are considered to be a single edge in the graph.

Data from users' profiles were collected into the users' dataset. For FOK! and De Kindertelefoon platforms the sets of attributes are different depending on the personal information availability. For the FOK! we collected the nickname, platform's personal identifier, sex, place, date of birth, education, profession, hobbies, timestamp of the account registration, timestamp of the last visit to the platform, the number of posts posted by the user (at the moment of scraping), the number of posts per day (at the moment of scraping). De Kindertelefoon is less informative in the sense of personal information, which may be implied by a more critical attitude towards anonymity. However, there are more statistical data contained. The list of the collected features is the following: nickname, platform's personal identifier, sex, year of birth, interests, and timestamp of the account registration. Also for each of the profiles, some statistics are provided, such as the number of topics created per person, the number of reactions he/she put, and the number of best answers posted by the user. Moreover, there are also the counts of followers and followees.

For each of the users on both platforms, additional statistics have been calculated based on the existing 'comments dataset'. Specifically, the number of topic initiations will be calculated. The quantity of comments that occurred in the "comments dataset" per user has been counted and referred to as the 'sex-related' comments. In order to anonymize the users, all the nicknames were masked by the integer code, which is going to be used for further analysis.

the igraph package (Csardi & Nepusz, 2006) is used to create a network graph. While scraping, all the users were converted into nodes and quotes into edges. Each node contains a set of attributes collected from the profiles.

4. 2. Data analysis design

4.2.1. Network analysis

We depart from network analysis, aiming to measure key features of user networks on both platforms. Zhao et al. (2021) base their users' behavioral study on metrics proposed by Wu

& Liu (2019), where they describe the concept of discourse network analysis (DNA). Wu & Liu (2019) ground their research in the following metrics:

- Discussion initiation – the number of threads the user created
- Interaction engagement – the number of replies the user created
- Influential scope – degree of centrality of the user

Degree centrality is a characteristic of the node within the network, which stands for the number of other nodes connected to the node directly (with one edge). As well as the directed graph is going to be used, the in- and out-degree of centrality should be distinguished. Following Wu & Liu (2019), for our research in-degree centrality is going to be used. In-degree centrality represents the interaction activity of a user (Wu & Liu, 2019). It is calculated as follows:

$$C_{D_I}(b_i) = \sum_{j=1}^{N} l_{ji}$$

where l_{ji} is the number of edges from j to i.

- Relational mediation – betweenness centrality of the user

The betweenness centrality of a user describes a degree to which he is located between other nodes in the network. Betweenness centrality measures the extent to which the node lies between other nodes and, in the context of the DNA, to which degree the user[1] controls the discourse shared by other users (Wu & Liu, 2019). Betweenness centrality is computed as follows:

$$C_B(b_i) = \frac{\sum_{b_i \neq b_j \neq b_L} g_{jl}(b_i)}{g_{jl}}$$

where $g_{jl}(b_i)$ represents the amount of shortest ways connecting the discourse behaviors b_j and b_l , which contain the discourse behavior b_i.

- Informational independence – closeness centrality of the user

Closeness centrality stands for the sum of the lengths of the routes from a particular vertex to all the other vertices in the network. Users with a high degree of closeness centrality are

[1] Wu & Liu (2019) use the term ‘discourse behaviour’ instead of a ‘user’, representing a node as a part of the temporal network. Since a static network is used in our research, the node of the network stands for the user of the platform.

able to reach other users through a small number of contacts and thus are independent while being out of communication control (Wu & Liu, 2019). It is defined as follows:

$$C_C(b_i) = \frac{1}{\sum_{j=1}^{N} d(b_j, b_i)}$$

where $d(b_j, b_i)$ is the length of the shortest way between b_j and b_i.

Having the users' network built, the content produced by the main actors will be screened by topic modeling, sentiment analysis, emotion detection, and toxicity.

4.2.2 Topic modeling

Topic modeling is a semantics extraction technique based on the frequency of words occurring in the comments. For the topic modeling, the LDA algorithm based on the Gibbs method (Griffiths & Steyvers, 2004) will be used. Topic modeling will be performed based on the framework used by Schreurs (2022), which includes the following steps:

1. Dutch stop words from the NLTK corpus (*NLTK :: Natural Language Toolkit*, n.d.) will be united with the English stop words list.
2. Comments left by the particular group of key users will be grouped by threads for faster lemmatization.
3. Text will be converted to lowercase, and the punctuation will be removed.
4. LDA model from the scikit-learn package will be trained.
5. LDA models with different N topics will be tested for coherency and the most optimal one will be selected based on the u_mass and c_v coherency scores.

4.2.3. Sentiment analysis and emotion detection

For sentiment analysis, Zhao et al. (2021) evaluate intensity and polarity. Intensity is calculated based on the Naive Bayesian classification and represents the score from 0 to 1, representing the sentiment 'strength' of the content, while polarity assessment is based on the corpus-based sentiment analysis implemented in the LIWC technique (Tausczik & Pennebaker, 2009). Currently, there is no robust model of the Dutch language, trained to perform sentiment analysis on short texts such as fora comments. Dutch BERT models for sentiment analysis were trained on book reviews, which represent large pieces of text. Besides that, the language of the fora comments differs from the well-structured book reviews. Lai (2022) compared the state-of-the-art transformer models RobBERT and

BERTje with the rule-based sentiment analyzer Pattern.nl. It was concluded that the transformers did not fit the domain of sexual health information and are not able to classify the neutral comments while being trained on the book reviews, which were labeled only as positive or negative. Besides that, the transformers sometimes classified the negative information as positive. Based on these outcomes, Pattern.nl library was found suitable for performing sentiment analysis on sexual health comments (Lai, 2022). Based on these outcomes, we find it expedient to measure the sentiment based on the dictionary-based approach and employ other corpus-based methods. The sentiment is to be measured on a continuous scale from -1 to 1. The emotion detection will be also performed based on the corpus-based models since no robust transformers models for emotion identification of the Dutch texts exist (see more in p. 4.3). The following corpus-based models are considered to be used for the sentiment:

- Cornetto (Vossen et al., 2012)
    - textblob-nl implementation (gvisniuc, n.d.; Sarkar, 2019)
- NRC Word-Emotion Association Lexicon (Mohammad & Turney, n.d.)
- Pattern.nl (Smedt & Daelemans, 2012)
- RBEM
    - Polarity (Tromp & Pechenizkiy, 2013)
    - Emotion detection (Tromp & Pechenizkiy, 2014)

Particular models to be determined later.

4.2.4. Toxicity

Wadden et al. (2021) propose quantitative metrics to evaluate the effect of moderation on online mental health platforms. One of the parameters they are assessing is civility, which has increased thanks to moderation. We aim to measure the civility on both of the platforms using the toxicity (hate speech) detectors model based on the various BERT Dutch language model variations**.** We plan to use these models:

- ml6team/distilbert-base-dutch-cased-toxic-comments model
- ml6team/robbert-dutch-base-toxic-comments
- IMSyPP/hate_speech_nl (Ljubešić et al., 2020)

Particular model to be determined later.

4.3. Methodological choices

4.3.1. Scraping

FOK! Forum

Along with the identification of the authorship of the comments, it was decided to collect the profile data of the authors. For the purposes of the current internship, the study of profile features is out of scope, however, they may be used for further research. Initially, we attempted to scrape the profiles by downloading the HTML tree of profile web pages, however, we encountered the necessity to accept cookies settings. We could not resolve this issue by sending an HTTP request, so it was decided to accept the cookies by emulating a user's acceptance in the browser. For this purpose, the Selenium package (*Selenium*, n.d.) for web browsing automation was adopted. Thereby, comments were rescraped by HTTP requests, while profile information was collected through Selenium.

De Kindertelefoon

Unlike the FOK! Forum comments dataset, the dataset of De Kindertelefoon comments did not contain unique identifiers of the comments. This fact significantly hindered rescraping. In order to match the comments in the dataset and on the website some heuristics were used. Particularly, comments texts were compared after being trimmed. If the comments' texts did not match then a manual assignment was performed (the algorithm chose all the comments in the thread posted the same as the matching comment day and proposed to a user to select the right one). This strategy of comment identification was inefficient since the number of posts, which were not matched by text automatically, was rather high. Around 20 comments were labeled manually before it was decided to give up manual labeling. After text matching, 4879 comments were left unlabeled (3691 initializing topic comments (first post) and 1188 non-initializing comments). Considering that the chance of significant changes in the first post is relatively low, it was attempted to assign the comment ID and the authorship to the initializing posts without text matching. Thus, 2917 first posts were additionally labeled. The remaining comments could not be related to the samples in the dataset for different reasons, the main one being the unavailability of either a comment or the whole thread online. After all, 85744 posts (~97,8% of the whole dataset), whose authorship was determined, proceeded for further analysis.

4.3.2. Key users detection

Based on the user behavior metrics, proposed by Zhao et al. (2021), we intend to split our metrics into four groups: Initiators, Repliers, Mediating Interactors, and Independent Interactors.

Initiators

In order to identify key topic initiators, we aim to select the users whose Discussion initiation (Zhao et al., 2021) score is higher than Mean+1SD.

Repliers

The group of the most active repliers is to be detected in accordance with the variable “Interaction engagement” (Zhao et al., 2021). This variable includes the replies (comments) of all the users, including those comments that may not contain a quote and thus do not represent a piece of interaction. In other words, posting many comments may not necessarily result in a meaningful interaction with the other users. Therefore, this variable is renamed to “Posting activity” in this amendment. For this group of key users, we aim to select the most active posters, by selecting users whose score on Posting activity is Mean+1SD.

Interactors

In terms of interaction activity, we aim to detect two subgroups, based on the levels of mediation and independence of the key users. This was done as mediating interactors are not necessarily independent from others, independent interactors are not necessarily influential, and both groups may influence other users differently.

In order to do so, we first select the most active users in terms of their influential scope (Zhao et al., 2021). The most influential users are considered to be engaged interactors, meaning that they post a significant number of replies to other comments (i.e., quoting other users) and their comments are replied to (i.e., quoted) as well. Influential scope (Zhao et al., 2021) is assessed based on the intersection of in-degree centrality (authoritative power degree (Wu & Liu, 2019)) and out-degree centrality (the measure of “gregariousness” (Borgatti et al., 2018), i.e. replying activity). We select users with a degree centrality higher than Mean + 1SD for both these degrees of centrality, which will form the “influential users set” (IUS) that further group formation is based on.

Mediating interactors: Next, having captured the most influential users, we gain to measure their ability to control the discourse, which is expressed in the level of relational mediation (Wu & Liu, 2019). To form this group, betweenness centrality is computed for the users in the IUS. We consider users with a high level of relational mediation to have a betweenness centrality score of higher than Mean + 1SD.

Independent interactors: We aim to assess the level of informational independence (Wu & Liu, 2019) of the key users from the IUS by measuring their closeness centrality. The users with a high level of information independence are considered to be those IUS users, whose closeness centrality score is higher than Mean + 1SD.

For all four groups of key users, we will run topic modeling, sentiment analysis, and toxicity detection separately.

It should be noted that Mean + 1SD is an arbitrarily selected threshold. Initially, the idea was to identify key users with a ‘high’ degree of centrality. The literature lacks a clear explanation of what high centrality means. The only explanation is provided by Powell et al.

claiming that “a high degree centrality score simply means that a node has a larger than the average number of connections for that graph” (2015). However, since the degrees of centralities are not normally distributed, they may not represent the whole set of actual key users. Some other techniques were tested to obtain the key user set, such as saturation or K-core measure. The idea of saturation was based on the assumption that after a certain threshold newly added key users will not bring new users they influence on to the network. The saturation was tested on the out-degree centrality. An empirically defined stop could be set at the level of 127 top out-degree key users, who cover ~49.5% of the network with an average amount of users brought to the network by an incoming key user to be 20 and a standard deviation to be 56, while the next 128 top out-degree users covered ~5% more (54.52% in total) of the whole network and brought only 2 new users to the network in average with a standard deviation to be equal 1.7. Thus, 128 users for the FOK! Forum could be set as a saturation threshold. However, the same strategy for De Kindertelefoon did not produce any significant results because of the very low saturation speed. This could be a result of the extremely low density of the De Kindertelefoon users' network.

The other technique to be tested was the K-core. It should be noted that K-core can produce affordable results in key users cluster detection, i.e. discovering the subgraphs in which key users are interrelated. However, the K-core network with a relatively high K does not capture the nodes with a high degree of centrality for the whole network. For this reason, the K-core algorithm for key users detection was also rejected.

Considering that saturation for the FOK! Forum produced a set with a similar cardinality as the mean + 1 standard deviation threshold, it was decided to leave Mean+1SD as a cut-off for all the behavioral measurements to build the key users set.

4.3.3. Sentiment analysis

It was decided to use the Pattern.nl package since it was already used by Lai (2022) and therefore the results of the sentiment analysis for the key users will be better compatible and comparable to the overall platform sentiment analysis performed by Lai (2022).

4.3.4. Emotion detection

Taking into account the limitations claimed by Schreurs (2022), the decision was made to conduct a multilevel platform analysis including the features, which were not studied before. The directions for further research mentioned above offer a vast variety of tools, which could be used for automated content and context analysis. Therefore, some machine learning models for solving natural language processing tasks (NLP) were observed on the HuggingFace repository, which was one of the most used repositories for machine learning

solutions. The main limitation that occurred at this point was the number of models, which were relevant to the Dutch language. Particularly, we had the intention to enrich the sentiment analysis performed by Lai (2022) and extract emotion from the comments. After extensive research, it should be acknowledged that no valid emotion detection models were found on the HuggingFace, which would be available to use from scratch. Moreover, De Bruyne et al. (2021) analyzed state-of-the-art Dutch language models and concluded that there were no robust results on emotion detection for the Dutch language by now. They trained RoBERTa and BERTje models based on the TV show captions and tweets corpora (which could be a suitable dataset for the purpose of the current research), and the results were quite modest (De Bruyne et al., 2021) and do not let us count on the machine learning language models. Specifically, the F1-measure for both of the models is lower than 0.4, while the highest cost-corrected accuracy (cost-correction is the metric to assess the cost of error for each emotion) is 0.65. To simplify, existing transformer models predict the emotion correctly in lower than 40% of cases. Since the goal of the research is not to experiment on machine-learning sentiment detection, we find these scores insufficient for the purposes of our research and therefore don't intend to use them, substituting them for corpus-based models, classifying emotions based on the frequency occurrence of specific emotion-labeled words in the comments.

Up to the current point, two corpus-based models for emotion detection were found: NRC Word-Emotion Association Lexicon (Mohammad & Turney, n.d.) (further NRCWEAL) and RBEM model (Tromp & Pechenizkiy, 2014). To support their paper, Tromp & Pechenizkiy uploaded the code they used for their model on Github (ErikTromp, n.d.). The code is not equipped with guidance and is provided only in PHP and Java programming languages, while the code for the current research was written in Python. These circumstances hindered the application of the model, so it was decided to choose the NRCWEAL and LeXmo python package (Bavli, 2021) as an implementation tool. The NRCWEAL contains 14155 English words, which are labeled with 8 emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust). Besides that, words are equipped with categorical polarity values (either positive or negative). The NRCWEAL is translated into 108 languages including Dutch.

The LeXmo package was optimized: particularly, the LeXmo function was converted to the class, which uploaded the file with the lexicon only once when being initialized instead of reading the lexicon every time the new comment was passed to calculate emotions. The Dutch lexicon file was downloaded from the NRC Lexicon website and put into the directory of the package. The emotion of the comment is calculated as a relation of the sum of emotion labels of the words to the number of words. Considering the fact that the lexicon is primarily composed of English words, while the Dutch words are just translations from English, some Dutch words were duplicated for different original English words. To normalize the weight of each word in the comment, it was decided to count an emotion of the word as an average emotion value of all the English translations of the Dutch word mentioned in the corpus. So far, the formula is the following:

$$e(w) = \frac{\sum_{i=1}^{n} val(e,w\_eng\_translation_i)}{n},$$

where *e* is the function of emotion calculation, *w* is the Dutch word, *n* is the number of English translations of the *w*, $w\_eng\_translation_i$ is the English translation of the *w*, *val* is the value of particular emotion *e* of $w\_eng\_translation_i$.

Overall the particular emotion value for a comment is calculated as follows:

$$E(c,e) = \frac{\sum_{i=1}^{n} e(w_i)}{n},$$

where *E(c,e)* is the function of calculation emotion *e* of the comment *c*, w_i is the ith word of the comment *c*, and *n* is the number of words in the comment *c*.

4.3.5. Toxicity

IMSyPP/hate_speech_nl model (Ljubešić et al., 2020) was chosen, since it provides a deeper insight, classifying the toxicity of the comments on the 4-level scale (0 - acceptable, 1 - inappropriate, 2 - offensive, 3 - violent). The other models only verify comments as toxic or non-toxic,

5. RESULTS

5.1. Network Description

Initially, the networks of quotes created based on the already existing comments dataset consisted of 5182 users for the FOK! Forum with 76279 quotes detected. Some comment authors were unidentifiable since the attempts of accessing their profiles resulted in errors because of different reasons (e.g. the profile was deleted or hidden in accordance with the privacy restrictions). We consider all these users as a single anonymous user, who includes the links to other users by all the anonymous users, who were quoted by anonymous profiles. It is clear that such an anonymous user is an outlying node in the graph, which does not bring any significant information to the network. Therefore, it was removed from

the network. Eventually, the anonymous user took away 9360 comments (which constitutes 8% of the comments dataset).

De Kindertelefoon network includes 6122 users with 10220 links between them. The comments dataset also contained comments from anonymous users as well, however, the nature of the Q&A platform suggests that all the users can post to the platform completely anonymously, without revealing their nicknames or any other information. This feature was widely used by the participants. All the anonymous users left 55102 comments (62,8% of the comments dataset). These comments were excluded from the network building process. The final size of the networks can be found in Table 1.

	Users		Quotes	
	With anon	Without anon	With anon	Without anon
FOK!	5182	5181	76279	69826
De Kindertelefoon	6122	6121	10220	5546

Table 1. Whole platforms network characteristics

Given the number of the edges in De Kindertelefoon graph, the number of interactions per user is considered to be very small: on average, each user quotes someone less than once. It meant that the network is considered to be rather sparse. To test this hypothesis, it was decided to measure the density of the networks. For De Kindertelefoon, the density is $\sim 1.48 \cdot 10^{-4}$, meaning that the network contains only 0.148% of all the possible edges. The FOK! Forum has a significantly higher density, which equals to $\sim 2.5 \cdot 10^{-3}$ (2.53% of all the possible connections), which is roughly 17 times higher than De Kindertelefoon does. There also should be noted that the average amount of comments per thread on the FOK! Forum and De Kindertelefoon are 57 and 8 respectively. Such a difference can be explained by the nature of the platforms: FOK! Forum is an online community, which proposes a durable interaction, while De Kindertelefoon represents the Q&A platform, which involves a situative model of communication when the content is shared on the form or questions, which are closed when the answer is found (Agichtein et al., 2008).

5.2. Key users

For both of the platforms 4 key user sets were constructed. The sizes of the sets, as well as the mean and standard deviation values, which were used as a threshold for the composition of the sets, are represented in Table 2. The number of influential Discussion initiators is 5 times less for the FOK! Forum than for De Kindertelefoon. At the same time, discussion initiators produced 105 topics (~5% of all the topics) on the FOK! Forum and 1531 topics (~14% of all the topics) on De Kindertelefoon. This again can be explained by

the constructions of the platforms (see above the difference between the forum and the Q&A platform). The Repliers posted 9801 comments on the FOK! Forum and 12390 comments on De Kindertelefoon (8,3% and 14% of all the comments in the datasets on the two platforms respectively).

	FOK!	De Kindertelefoon
Discussion Initiators	51 (M=1.74, SD=3.01)	246 (M=1.84, sd=1.95)
Repliers	212 (M=21.7, SD=83.16)	174 (M=6.79, SD=20.45)
IUS	183 (In-degree: M=13.47, SD=51.65) (Out-degree: M=13.47, SD=59.67)	91 (In-degree: M=0.9, SD=3.38) (Out-degree: M=0.9, SD=4,77)
Mediating interactors	108 (M=4527.71, SD=34336.46)	70 (M=446.55, SD=4272.56)
Independent interactors	73 (M=0.324, SD=0.085)	0 (M=0.298, SD=0.225)

Table 2. Key users sets

Mediators and Independent Interactors, which are composed based on the IUS (see section 4.3). Interestingly, no independent users were found in the IUS of De Kindertelefoon. It may be the result of the short threads on De Kindertelefoon, while users have no ground to stay away from the communication network because they are always bound within a particular thread.

5.3. Topic modeling

Initiators:

For De Kindertelefoon (n_topics=6, *u_mass* score=-1.195) (see Appendix, Fig. 10), there seem to be more distinct topics, such as size/shape of genitals (topic 4 and 6), advice on sexual activities (topic 1 and 3), and normative statements such as ‘good’ or ‘weird’ (topic 5).

On the FOK! Forum the topics (see Appendix, Fig. 6) are less distinct (n_topic=4, *u_mass* score=-0.65), but do also seem to indicate some normative statements such as ‘good’ and ‘nice’ (topic 1), or are about porn and homosexuality (topic 4).

Repliers:

On De Kindertelefoon, the topics (n_topic=5, u_mass score=-1.356) (see Appendix, Fig. 11) for the Repliers are similar to the Initiators, there is quite some talk about the size/shape of

genitals (topic 1, 4 and 5) and sexual activities (topic 4 and 5), as well as more normative statements such as ‘good’ or ‘nice’ (topic 3 and 4).

On the FOK! Forum (n_topics=3, u_mass score=-0.335) (see Appendix, Fig. 7), all topics seem to mostly be about opinions and normative statements, such as ‘good’, ‘nice’, ‘beautiful’, and ‘thinking/finding’, ‘normal’ (which can also be seen as the stop-word ‘just’), ‘have to’. Some indication of contraception (topic 1) and porn (topic 4) is present as well.

Mediators:

On De Kindertelefoon (n_topics=4, u_mass score=-1.4) (see Appendix, Fig. 12), topic 1 is mostly an ‘opinion’ topic (containing the words such as ‘good’, ‘normal’, ‘have to’, ‘finding’, ‘thinking’, ‘knowing’). Topic 2 and 3 are about genitals, sexual activities, and porn. Topic 4 seems to be a combination of the other 3 topics.

On the FOK! Forum (n_topics=3, u_mass score=-0.38) (see Appendix, Fig. 8), topic 1 includes talk about contraception, and topic 2 and 3 seem to be opinion topics, just like topic 1 of the Kindertelefoon.

Independents:

On the FOK! Forum (n_topics=4, u_mass_score=-0.4 (see Appendix, Fig. 9), the four topics of the independents are very similar. Interesting to note that the topics do not include many keywords about sex, contraceptives, etc., and seem to be mostly neutral words such as ‘thinking’, ‘finding’, ‘thinking’, ‘saying’ etc.

To sum up, De Kindertelefoon seems to have more specific topics about genital (size) and sexual activities, for all of the key user groups. Interestingly, contraceptives are not mentioned that much on De Kindertelefoon, and are mostly mentioned on the FOK! Forum by Repliers and Mediators. FOK! Forum also includes some more talk about porn, and seems to be largely a ‘reactive’ forum, where people voice opinions, which look rather supportive.

5.4. Sentiment analysis

Lai (2022) performed a sentiment analysis on the whole datasets of both platforms. The results showed almost neutral sentiment on both of the platforms (FOK!: M=0.065, SD=0.26; De Kindertelefoon: M=0.088, SD=0.25) (Lai, 2022). After cleaning the data (particularly, the comments, which “were not understood by the Pattern.nl” (i.e., very short or empty comments), the average sentiment increased (FOK!: M=0.1, SD=0.32; De Kindertelefoon: 0.11, SD=0.28). Cleaning includes the removal of the comments that “were not understood by Pattern” (Lai, 2022). Lai (2022) does not provide the full algorithm of cleaning, however, she noted that Pattern.nl “does not handle short sentences well”, while

the comments, which were not “understood” were mostly empty, too abstract, or contained a lot of Dutch slang.

The sentiment analysis was performed both on the uncleaned and cleaned data. However, in the current research cleaning included only empty comments removal, which resulted in 113526 analyzed comments for the FOK! Forum and 84050 comments for De Kindertelefoon. More elaborated sentiment analysis is to be conducted outside of the current internship. Table 3 represents the results of the sentiment analysis on the uncleaned data. All the subsets of the comments posted by the key users on the FOK! Forum do not demonstrate any difference in polarity. However, the key actors of De Kindertelefoon post fewer positive comments than all the users together, although the sentiment for all the key users increased after cleaning (Table 4). Interestingly, the polarity of the FOK! Forum did not change significantly after the removal of empty comments and maintains the level of uncleaned overall polarity. Cleaning replication by Lai (2022) for the key users' comments to be processed out of this internship.

	Total (Lai, 2022)	Discussion initiators	Repliers	Relational Mediators	Independents
FOK!	M=0.065 SD=0.26	M=0.064 SD=0.265	M=0.066 SD=0.26	M=0.065 SD=0.258	M=0.066 SD=0.258
De Kindertelefoon	M=0.088 SD=0.25	M=0.071 SD=0.25	M=0.078 SD=0.25	M=0.065 sd=0.25	

Table 3. Uncleaned comments sentiment

	Total (Lai, 2022)	Discussion initiators	Repliers	Relational Mediators	Independents
FOK!	M=0.1 SD=0.32	M=0.064 SD=0.265	M=0.066 SD=0.26	M=0.065 SD=0.258	M=0.066 SD=0.258
De Kindertelefoon	M=0.11 SD=0.28	M=0.077 SD=0.26	M=0.085 SD=0.26	M=0.072 sd=0.266	

Table 4. Cleaned comments polarity

Subjectivity of the comments posted by the key users is generally lower than the comments posted by the whole communities (Table 5 & 6). It may be proposed that the key actors tend to talk more objectively, using more rational language. However, no robust results can be presented in the current report since more elaborated analysis is needed.

	Total (Lai, 2022)	Discussion initiators	Repliers	Relational Mediators	Independents

FOK!	M=0.43 SD=0.34	M=0.431 SD=0.343	M=0.43 SD=0.345	M=0.435 SD=0.341	M=0.442 SD=0.339
De Kindertelefoon	M=0.52 SD=0.31	M=0.454 SD=0.338	M=0.462 SD=0.334	M=0.432 sd=0.346	

Table 5. Uncleaned comments subjectivity

	Total (Lai, 2022)	Discussion initiators	Repliers	Relational Mediators	Independents
FOK!	M=0.65 SD=0.19	M=0.431 SD=0.343	M=0.43 SD=0.345	M=0.435 SD=0.341	M=0.442 SD=0.339
De Kindertelefoon	M=0.65 SD=0.19	M=0.494 SD=0.325	M=0.5 SD=0.318	M=0.488 SD=0.329	

Table 6. Cleaned comments subjectivity

5.5. Emotion detection

Emotion detection was carried out for the whole datasets of De Kindertelefoon and the FOK! Forum without splitting the comments across key users. Overall, the words are very neutral and do not contain any emotion: for instance, trust, the most present emotion in the comments, occurred only in 1.5%-1.6% of the comments on both platforms (Fig. 2). FOK! Forum contains a bit more emotional content across all the emotions except fear. Emotion detection for the key user groups will be conducted outside of this research.

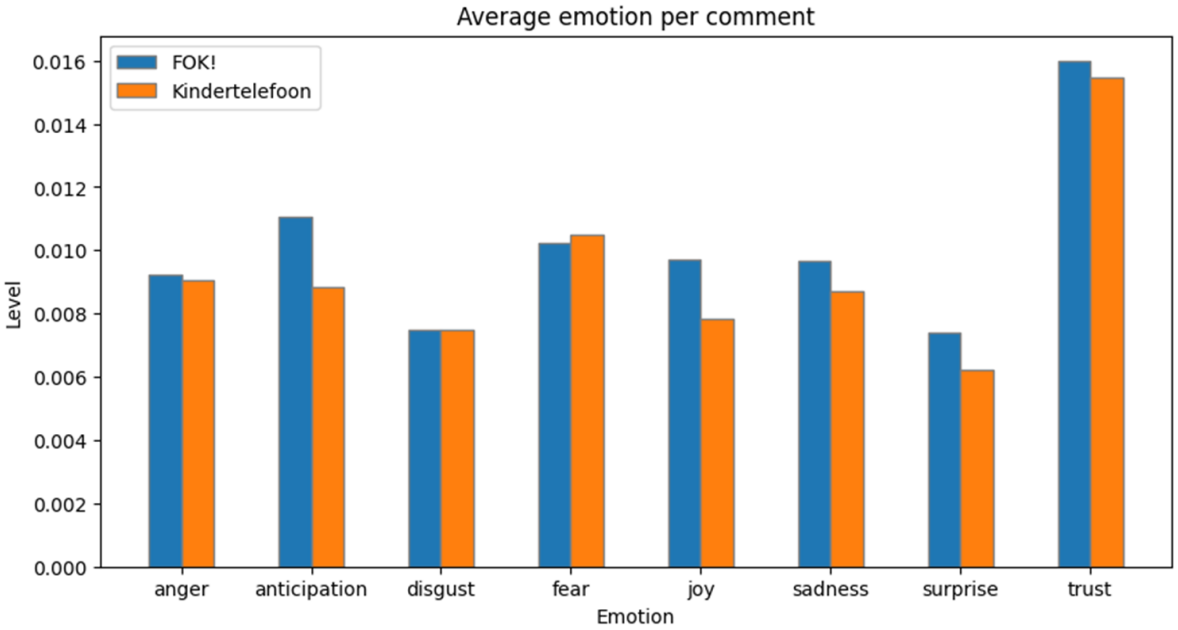

Figure 2. Emotions distribution across the platforms.

5.6. Toxicity

Toxicity analysis shows that on average the key users of De Kindertelefoon post more non-toxic comments than the key users of the FOK! Forum (Fig. 3). The amount of violent comments for De Kindertelefoon is lower than 0.1% for all the key user groups, while almost absent in the Relational Mediators group. For the FOK! Forum the level of violent comments does not exceed 0.27%. However, the level of offensive comments is rather high (higher than the number of inappropriate comments) on both of the platforms (28.2% on the FOK! Forum vs. 19.9% on De Kindertelefoon for the whole platforms) Interestingly, the level of offensive comments on De Kindertelefoon across the key users' groups is lower than across the entire platform (12.6%-13.4% vs. 19.9%), while for the FOK! Forum it remains on the same level (28%-30%). The proportion of offensive and inappropriate comments on De Kindertelefoon is similar – around 12-15% across all the key users groups. The number of appropriate comments across all the key user groups for De Kindertelefoon is a bit higher than across the whole dataset (72–74% vs. 67%), however, for the FOK! Forum these proportions are quite equal (63-66% vs. 66%).

Overall, the level of toxicity is surprisingly high, which goes against the conclusions of the topic modeling results, specifically for the FOK! Forum, where the sharing of points of view is considered to be rather positive.

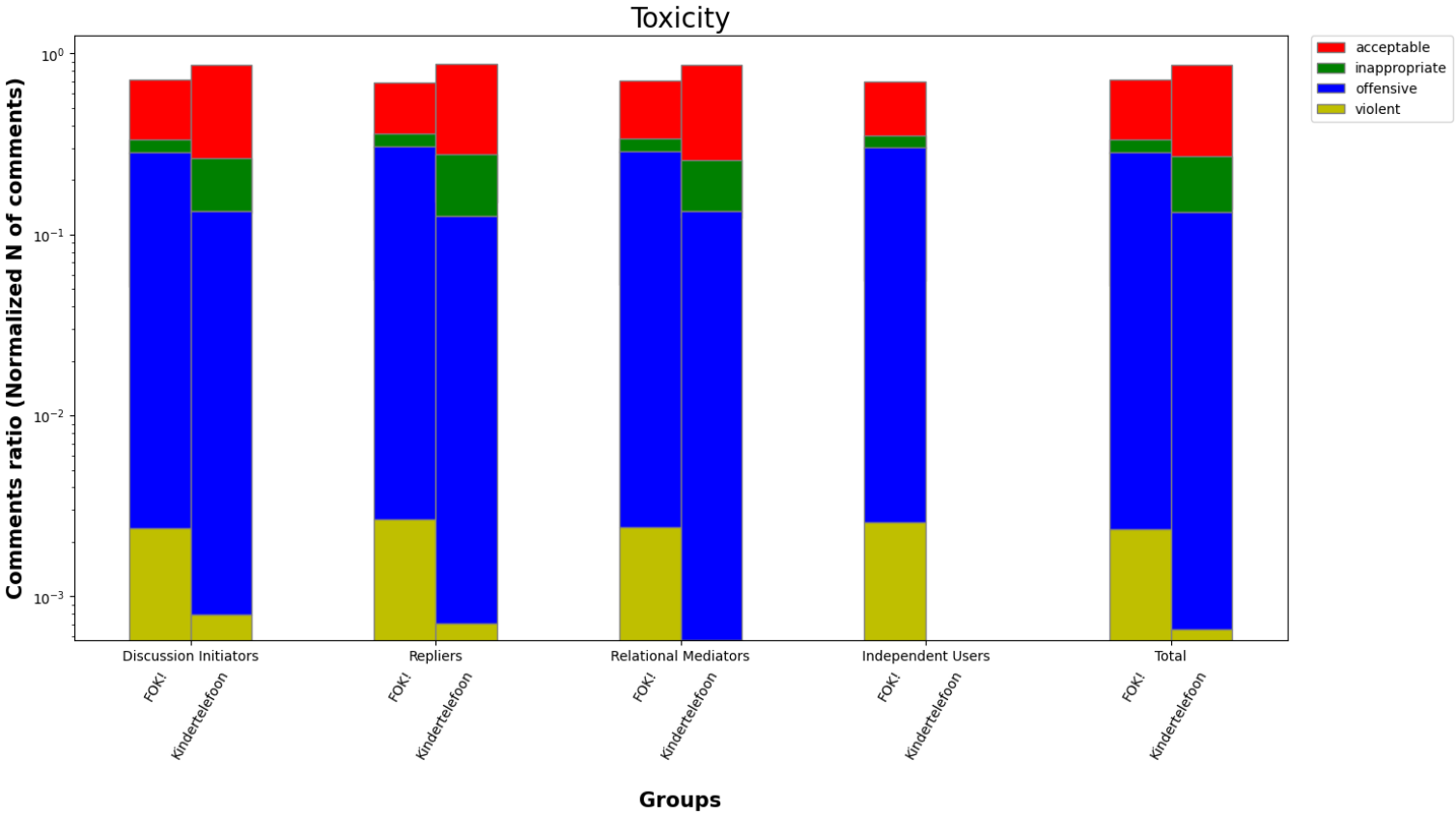

Figure 3. Toxicity distribution across the platforms and key user groups.

6. DISCUSSION

This project is one more step in the sequence of research, which is designed to establish the quality of the information on sexual health shared online. Previous work has been carried out based on the idea of credibility detection through linguistic markers and sentiment analysis (Lai, 2022; Schreurs, 2022). Our idea was to expand this analysis to the features, which could grasp the context of the platforms, while not focusing only on the text. The extension of the scope in terms of the methods led to difficulties with the definition of credibility because this concept was not sufficiently elaborated in the literature. Because of this, it was decided to stick to the better-defined concept such as misinformation. Misinformation, including one related to health issues, is a widely studied phenomenon, which proposes a reasonable selection of methods.

Since the current research has a clear inductive nature and is based on the data collected during the previous work, we are developing our framework based on the methods available to analyze the data. Indeed, the majority of the work related to credibility and misinformation detection is mostly based on content features. So far, the extension of the scope to the

context-based features encounters a lack of literature, needed for establishing a solid theoretical framework.

We justify our approach based on the work of Zhao et al. (2021), who built the classification model for misinformation detection, which was grounded in the features observed in our research. Although more literature review is needed to set up the framework on the context-based features, we see our work as an initial step to go beyond the text markers towards more complex studies of misinformation including user behavior, which could be a better predictor of misinformation identification (Zhao et al., 2021).

The network analysis revealed some difficulties when comparing platforms of different types: traditional forums and Q&A sources. Defining an edge as a quote may limit the quality of the network, since the graph of the quotes mostly captures the users, who communicate mainly within the boundaries of a certain specific thread, and thus it may be complicated to capture the network of the platforms overall. This may result in the low density of the networks, which was discussed above. For further analysis of key user detection, it could be beneficial to incorporate other types of bonds (the data on profiles has been collected in the course of the current research but not used). Another idea for improving the quality of the graph based on the quotes is to capture the clusters of the key users, which are formed based on the aggregation of semantically close threads.

The first results of our research propose that key users mostly follow the same discourses as the majority of users do. They discuss the topics, which are discussed by all the users on the platforms with a bit more neutral sentiment. The level of toxicity is a bit higher across the key users of the FOK! Forum, however, on De Kindertelefoon key users on the contrary tend to be more civil, which goes in line with Wadden et al. (2021), who postulate the positive effect of moderation on civility.

Emotion detection results are supported by the sentiment analysis, which suggests an emotionally neutral nature of the content. As long as fear is the only emotion received a higher score for De Kindertelefoon than for the FOK! Forum among all the emotions, some light may be shed on the nature of the platforms. More precisely, De Kindertelefoon is a helpline moderated by professionals, it may be the point of contact in case of emergency situations when children are frightened and seek competent advice. The prevalence of the other emotions on the FOK! Forum goes in line with the results of topic modeling, which suggests that the FOK! Forum is used as a platform for supportive opinion sharing. At the same time, more variance in emotions for the FOK! Forum can also be the result of non-moderation, which lets users express their attitudes more freely.

7. CONCLUSION

The aim of the current research is to identify the effect of moderation on the discourses, spread online by key users. Descriptively, it can be concluded that the moderated platform

delivers less toxic content, which has more positive sentiment. It also stimulates less emotionally diverse content. However, as long as the key users are the main actors responsible for content sharing, which includes misinformation, a more elaborated analysis of the content they produce is needed to address the quality of the argument.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). Finding high-quality content in social media. *Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08*. http://dx.doi.org/10.1145/1341531.1341557

Bavli, D. (2021, November 17). Unlocking emotions in text using python. *Better Programming*. https://betterprogramming.pub/unlocking-emotions-in-text-using-python-6d062b48d71f

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619–638.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2018). *Analyzing social networks*. SAGE Publications Limited.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th International Conference on World Wide Web - WWW '11*. http://dx.doi.org/10.1145/1963405.1963500

Chou, W. S., Prestin, A., Lyons, C., & Wen, K. (2013). Web 2.0 For health promotion: Reviewing the current evidence. *American Journal of Public Health*, *103*(1), e9–e18.

Crocamo, C., Viviani, M., Famiglini, L., Bartoli, F., Pasi, G., & Carrà, G. (2021). Surveilling COVID-19 emotional contagion on twitter by sentiment analysis. *European Psychiatry*, *64*(1).

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695.

De Bruyne, L., De Clercq, O., & Hoste, V. (2021). Prospects for Dutch emotion detection: Insights from the new emotioNL dataset. *Computational Linguistics in the Netherlands Journal*, *11*, 231–255.

ErikTromp. (n.d.). *GitHub - ErikTromp/RBEM: Multilingual sentiment analysis*. GitHub. Retrieved December 22, 2022, from https://github.com/ErikTromp/RBEM

GitHub - Gvisniuc/textblob-nl: Dutch language support for TextBlob. (2019, November 7). GitHub. https://github.com/gvisniuc/textblob-nl

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, *12*(1). https://doi.org/10.1007/s13278-022-00951-3

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl1), 5228–5235.

Kumar, K. P. K., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1).

Lai, J. (2022). *The extent of sentiment on sexual health topics between moderated and non-moderated websites* [Master Thesis, Utrecht Universiteit]. https://doi.org/20.500.12932/42715

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Ljubešić, N., Markov, I., Fišer, D., & Daelemans, W. (2020). The LILAH emotion lexicon of Croatian, Dutch and Slovene. *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, 153–157.

Mohammad, S. M., & Turney, P. (n.d.). *NRC emotion lexicon*.

Nikkelen, S. W. C., van Oosten, J. M. F., & van den Borne, M. M. J. J. (2019). Sexuality education in the digital era: Intrinsic and extrinsic predictors of online sexual information seeking among youth. *The Journal of Sex Research*, *57*(2), 189–199.

NLTK :: Natural language toolkit. (n.d.). Retrieved December 21, 2022, from https://www.nltk.org/

Ognyanova, K. (2021). Network approaches to misinformation evaluation and correction. In M. S. Weber & I. Yanovitzky (Eds.), *Networks, Knowledge Brokers, and the Public Policymaking Process* (pp. 351–373). Springer International Publishing.

Petty, R., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (pp. 1–24). Springer.

Powell, J., Hopkins, M., Powell, J., & Hopkins, M. (2015). 19 - Graph analytics techniques. In *Chandos Information Professional Series* (pp. 167–174). Chandos Publishing. https://www.sciencedirect.com/science/article/pii/B9781843347538000191

Saha, K., Ernala, S. K., Dutta, S., Sharma, E., & De Choudhury, M. (2020). Understanding moderation in online mental health communities. *Lecture Notes in Computer Science*, 87–107.

Sarkar, D. (2019). *Text analytics with python: A practitioner's guide to natural language processing*. Apress.

Schreurs, J. (2022). *Sexual health information: Credible or not?* [Master Thesis, Utrecht Universiteit]. https://doi.org/20.500.12932/42760

Selenium. (n.d.). Selenium. Retrieved December 22, 2022, from https://www.selenium.dev/

Smedt, T. D., & Daelemans, W. (n.d.). ``Vreselijk mooi!'' (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 3568–3572.

Sunkara, J. (2021). Sexual health misinformation and potential interventions among youth on social media. *The Cardinal Edge*, *1*(1).

Tausczik, Y. R., & Pennebaker, J. W. (2009). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Tromp, E., & Pechenizkiy, M. (2013). RBEM. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '13*.

Tromp, E., & Pechenizkiy, M. (2014). *Rule-based emotion detection on social media: Putting tweets on plutchik's wheel*.

Vossen, P., Maks, I., Segers, R., Vliet, H. van der, Moens, M.-F., Hofmann, K., Sang, E. T. K., & de Rijke, M. (2012). Cornetto: A combinatorial lexical semantic database for Dutch. In *Essential Speech and Language Technology for Dutch* (pp. 165–184). Springer Berlin Heidelberg.

Wadden, D., August, T., Li, Q., & Althoff, T. (2021). The effect of moderation on online mental health conversations. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 751–763.

Wu, J., & Liu, Y. (2019). Deception detection methods incorporating discourse network metrics in synchronous computer-mediated communication. *Journal of Information Science*, *46*(1), 64–81.

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhao, Y., Da, J., & Yan, J. (2021). Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management*, *58*(1), 102390.

Appendix

Key users networks

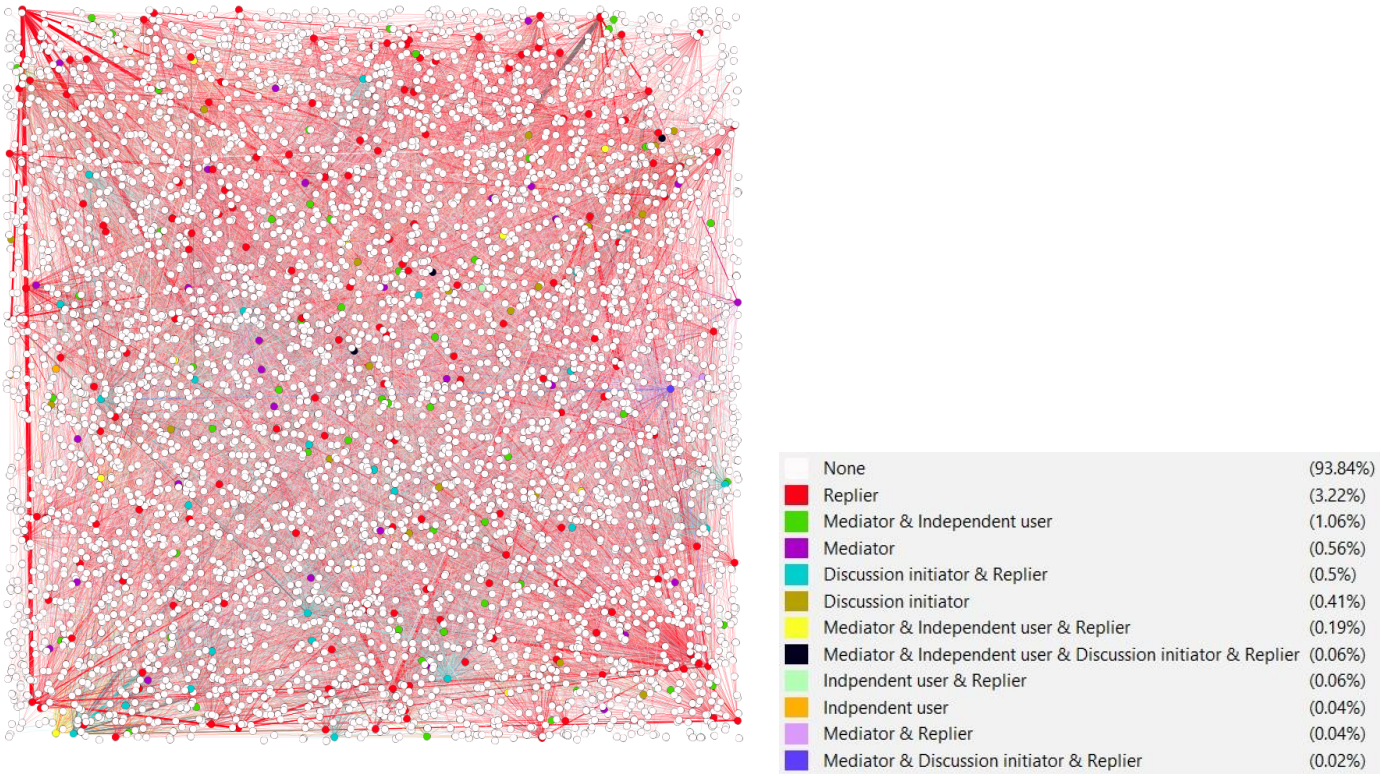

Figure 4. FOK! Forum key users network.

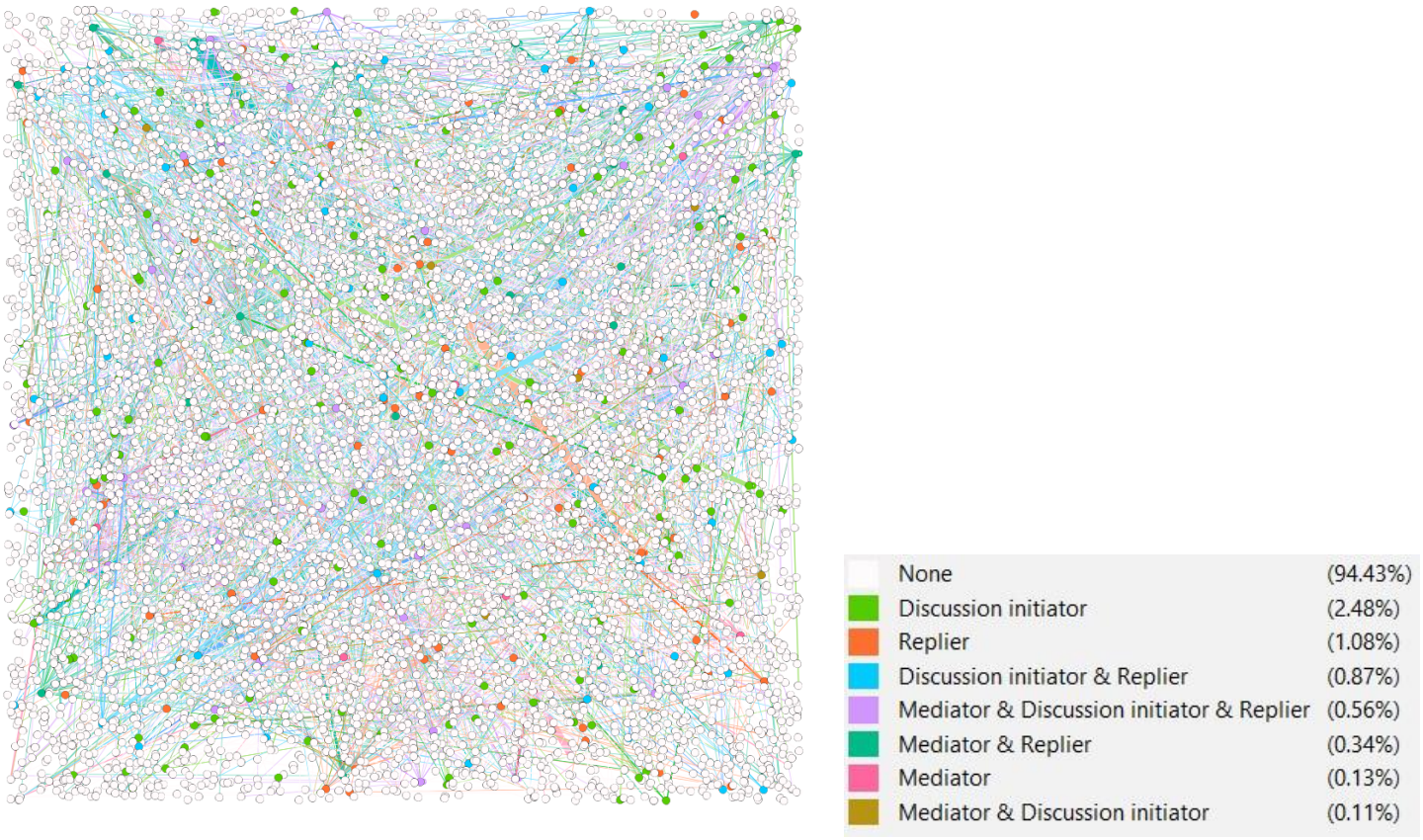

Figure 5. De Kindertelefoon users network

Topic modeling diagrams

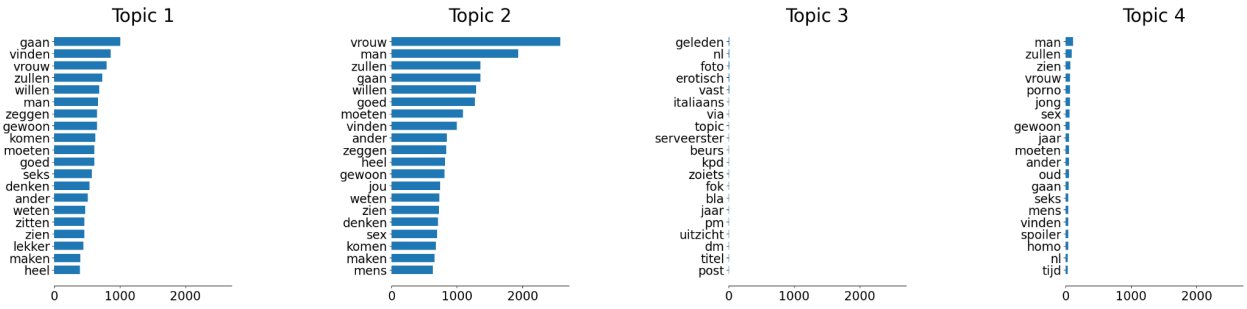

Figure 6. FOK! Forum Discussion Initiations topics

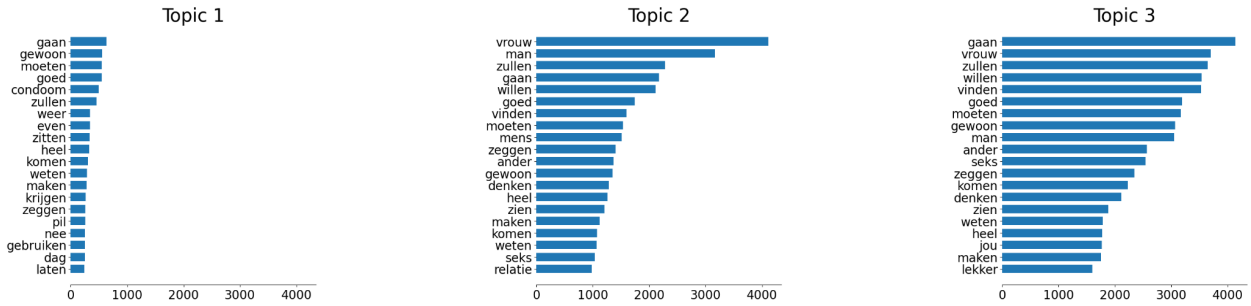

Figure 7. FOK! Forum Repliers topics

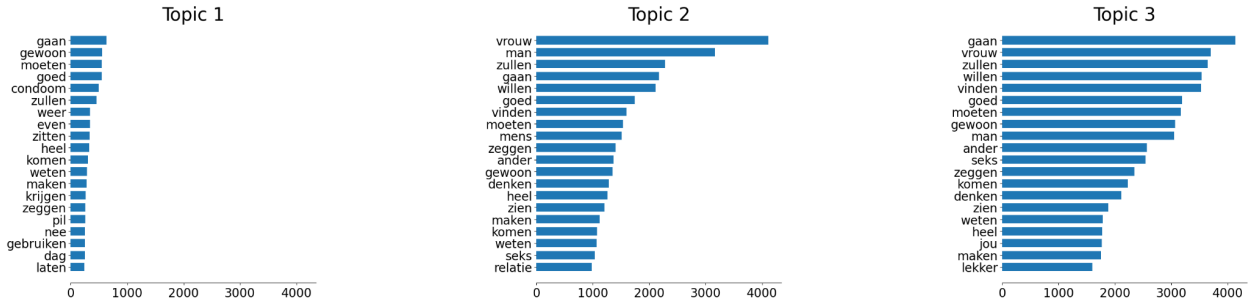

Figure 8. FOK! Forum Relational Mediators topics

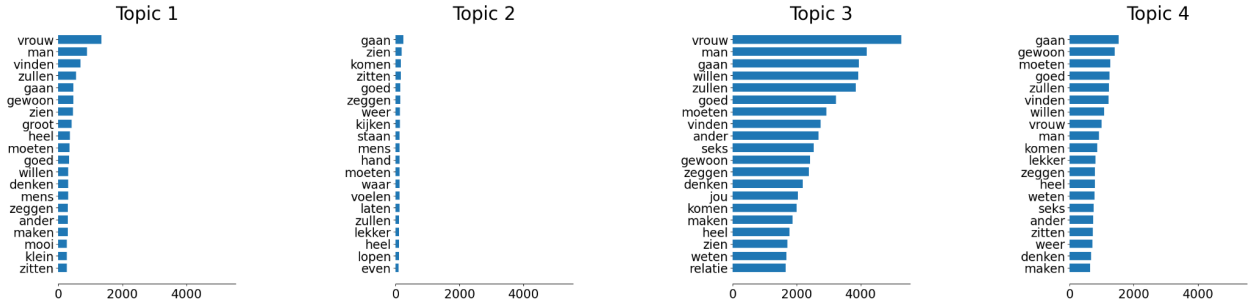

Figure 9. FOK! Forum Independents topics

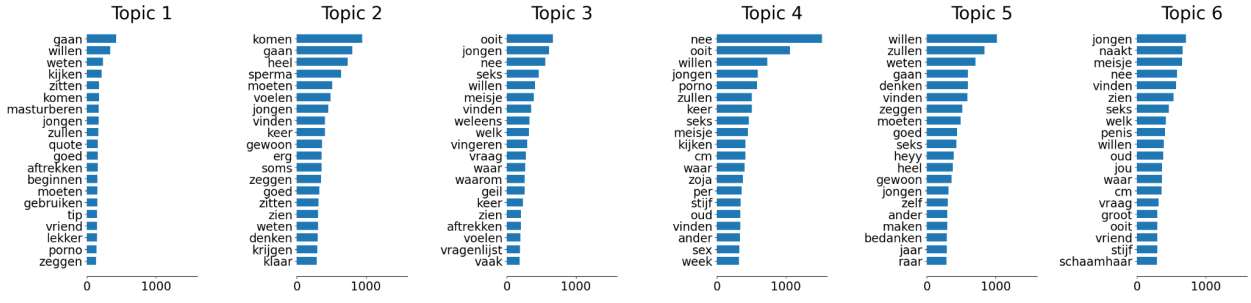

Figure 10. De Kindertelefoon Discussion Initiations topics

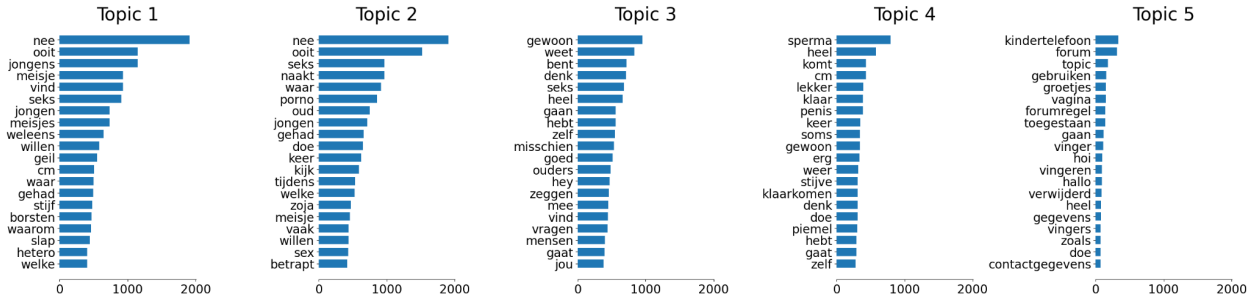

Figure 11. De Kindertelefoon Repliers topics

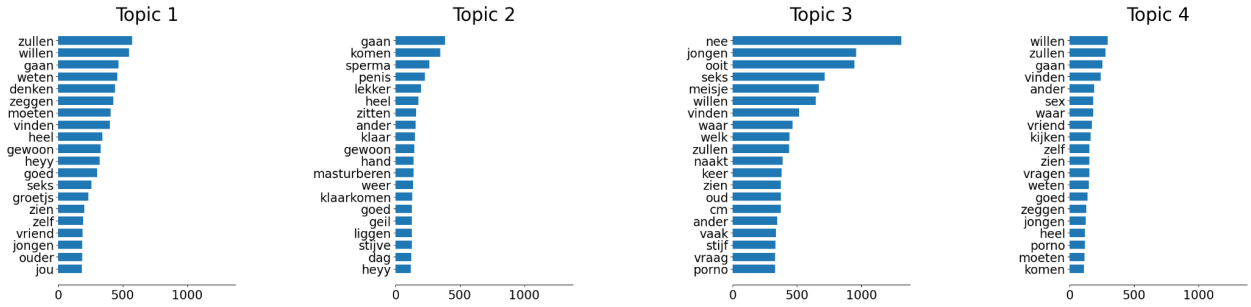

Figure 12. De Kindertelefoon Relational Mediators topics