FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
«NATIONAL RESEARCH UNIVERSITY
«HIGHER SCHOOL OF ECONOMICS»
Faculty of Social Sciences
Department of Methods for Collection and Analysis of Sociological Information

Lapshina Ksenia Mikhailovna

FRAMING THE DISCUSSION ON SEXUAL VIOLENCE IN RUSSIAN MASS MEDIA

Final qualifying work – BACHELOR THESIS
Educational program 39.03.01 «Sociology»

Advisor:
Associate Researcher
Center for the Sociology of Higher Education
Institute of Education
Zhuchkova Svetlana Vasilyevna

MOSCOW
2024

CONTENTS

This thesis examines the framing of sexual violence against women in Russian news media from 2016 to 2023. The study employed topic modeling with the Pachinko allocation model and sentiment analysis using the FastText social network model to identify thematic structures and explore the emotional tones of the discussion. The results revealed that Russian news media coverage of sexual violence against women from 2016 to 2023 featured 37 dominant themes, with a significant focus on criminal cases, legislative regulation, and sensationalist narratives. Contrary to initial assumptions, victims were often portrayed with agency, and responsibility was attributed to both individuals and institutions. Additionally, while sensationalist rhetoric was prevalent, sentiment analysis indicated weak correlations between emotional tones and specific themes, with coverage remaining consistent over time. The interpretation of the thematic structure of the discussion also reveals the use of problem definition, cause diagnosis, moral judgment, and remedy suggestion framing functions in the coverage of sexual violence by the Russia news mesia, highlighting how media narratives shape public perception by focusing on individual responsibility, sensationalism, and the roles of government and societal factors. The study's limitations include reliance on Medialogia resources, which may bias article selection, and the inherent constraints of topic modeling and sentiment analysis, which can overlook nuanced themes and misrepresent sentiment.

INDTRODUCTION

Gender-based violence is the subject of active public discussions in post-Soviet Russia, where the formation of a modern gender order is relatively recent: in the media, a large number of cases related to various forms of violence against women by spouses, cohabitants or close relatives have been discussed in recent years.

In general, the discussion of gender-based violence in the media largely remains focused on private problems and specific cases, framing situations of violence as unexpected and situational, although in English-language media in recent years there has been a tendency to frame violence as a systemic problem requiring public discussion (Cullen et al., 2019; Ryan et al., 2006). It is worth noting that studies that draw conclusions about this trend have been conducted in countries where legislative instruments to protect women from violence have been in place for a long time, while in Russia gender-sensitive legislation is only at the discussion stage. Parallel to the discussion of gender-based violence in the media, there were two iterations - in 2016 and 2019 - of the discussion of the draft law "On the Prevention of Domestic Violence in the Russian Federation",[1] which in both versions was never adopted; however, in 2017, a federal law decriminalizing domestic violence was adopted.[2] Moreover, the legislation regulating sexual harassment was discussed in the Duma, where the potential bills also froze at the discussion stage and were never adopted.[3]

Meanwhile, the media plays an important role in practices, policies and public perceptions of gender-based and, in particular, sexual violence; public opinion is assumed to influence policy (Aroustamian, 2020). Public opinion, in its turn, is influenced by the media which has the ability to shape public consciousness: for

[1] Bill № 1183390-6 “On the Prevention of Domestic Violence in the Russian Federation” (in Russian). SOZD State Duma website. URL: https://sozd.duma.gov.ru/bill/1183390-6

[2] Federal Law "On Amendments to Article 116 of the Criminal Code of the Russian Federation" dated 02/07/2017 N 8-FZ URL: https://www.consultant.ru/document/cons_doc_LAW_212385/

[3] For example, see
Оксана Пушкина впишет харассмент в закон на фоне скандала со Слуцким. РБК, 27.02.2018. URL: https://www.rbc.ru/politics/27/02/2018/5a942e6a9a79471333d3128a
Депутат Госдумы разрабатывает законопроект, предусматривающий лишение свободы за харассмент. Комменсантъ, 18.06.2019. URL: https://www.kommersant.ru/doc/4004333

example, in regard to domestic violence, even regular access to radio and television, regardless of the content, leads to less acceptance of such violence among women (Bhushan, Singh, 2014). In communication and journalism studies, the social constructivism perspective is adopted that grants the media the power of constructing social reality through framing, and on the other hand, taking into account the fact that the audience has preliminary political and social attitudes (Baysha, 2004; Scheufele, 1999).

Therefore, the very content of media articles and the interpretive schemes used by different institutional actors in delivering information can be a way to influence public discourse or the course of political discussions (Boydstun et al., 2013). As for sexual violence, research suggests that since most people have no or too little personal experience with these issues and do not encounter a large number of crimes on an everyday level, they tend to rely on other sources for information about crime and violence with media being the primary one (Baranauskas, & Drakulich, 2018; Kort-Butler & Habecker, 2018; Kort-Butler & Hartshorn, 2011; O'Hear, 2020). However, the coverage of sexually violent incidents is traditionally considered to be quite distorted, relying on the "rape myths", cultural stereotypes and high levels of sensationalist rhetoric that limits the ability of interpreting sexual violence as a systemic problem that requires social or governmental action (Alcoff & Gray, 1993, Benedict, 1993; Boranijašević, 2018; Block, 2002; Hindes & Fileborn, 2020; Hollander & Rodgers, 2014; Lemish, 2004; O'Hara, 2012; Serisier, 2017).

The use of different framing strategies in covering the debate on sexual violence is particularly important in the context of societies that do not have developed gender-sensitive legislation that regulates gender-based violence, as the public debate unfolds, in particular, towards the adoption or non-adoption of such legislation.

Therefore, the ***research question*** of this thesis may be put as: how do Russian news media frame the cases of sexual violence against women, as well the sexual violence against women in general?

The ***theoretical object*** of the study is framing of the cases of sexual violence against women, as well as of the sexual violence against women in general, by the Russian news media. The ***subject*** of the study is ways of covering the cases of sexual violence against women and of the sexual violence against women in general used by Russian news media, as well as the interpretive frameworks, or frames, through which such coverage is carried out. The ***purpose of the study***, therefore, is to identify the ways through which the cases of sexual violence against women and of the discussion on sexual violence against women in general are covered by the Russian news media, as well as the interpretive frameworks, or frames, through which such coverage is carried out.

The ***theoretical objectives*** of the study that were set to develop a theoretical framework were (1) to explore the concept of sexual violence within the wider context of gender-based violence and describe the common characteristics of sexual violence based on the available research; (2) to identify and explore the concepts of frame and framing, as well as the characteristics of the media framing process as a way of construction the social reality, particularly the reality of violence and crime, based on the available research; (3) to describe the framing patterns that are used in media coverage of sexual violence against women, as well as the other characteristics of media coverage of sexual violence, such as victim portrayals, based on the available research.

The ***empirical research object*** of this thesis is the texts of Russian news media articles that describe cases of sexual violence against women or contribute to the discussion on sexual violence against women in other ways. The final study sample consisted of 15 342 texts of the articles on sexual violence against women published by 65 Russian news media outlets throughout the period of 2016-2023, automatically collected from the newspaper websites with the usage of custom scraping and parsing functions.

To identify the ways in which Russian media frames sexual violence and therefore to achieve the purpose of the study, several ***empirical research objectives*** were set:

RO1. To identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media.

RO2. To explore the relationship between the themes present in the articles on sexual violence against women and the tone that is used by Russian news media to cover such violence.

RO3. To explore the temporal differences in the thematic coverage of sexual violence against women in Russian news media.

The RO1 is essentially a descriptive task, though it was carried out through topic modeling that itself requires high levels of interpretative work. In relation to this fact, as well as with some patterns of media coverage of sexual violence identified in the literature, I made several ***assumptions*** regarding the thematic structure of the discussion on sexual violence in Russian mass media:

A1. There will be themes within the thematic structure of the discussion on sexual violence against women that indicate the absence of agency of the victims of sexual violence.

A2. Within the thematic structure of the discussion on sexual violence against women, newspapers will tend to attribute responsibility for the violence to individuals rather than the government or society.

A3. There will be a sensationalist rhetoric present in the Russian news media articles that cover sexual violence.

For the other research objectives, the ***hypotheses*** are:

H1. The themes that explore the cases of sexual violence will be correlated to either positive or negative sentiment in the coverage of sexual violence.

H2. The themes that explore sexual violence as a social problem will be correlated to the neutral sentiment in the coverage of sexual violence.

H3. There will be no correlation between the other themes and level of sentiment in the articles covering sexual violence against women.

H4. The general themes within the discussion on sexual violence in Russian news media will tend to be fading over time in terms of their presence in the articles.

To address the assumptions and test the hypotheses, several methods were carried out: topic modeling with the selected Pachinko allocation topic model, sentiment analysis with the usage of FastText social network model for the embeddings and the methods for testing linear correlation and comparing means to discover the patterns of how Russian news media cover sexual violence. The results include the outcomes of hypotheses testing, as well as substantive interpretations of the thematic structure of the discussion on sexual violence.

This thesis has several limitations. It relies on Medialogia resources, which may bias article selection and exclude independent Russian news media, potentially affecting the framing of sexual violence. Topic modeling and sentiment analysis, while useful, can overlook nuanced themes and misrepresent sentiment. Future research should explore non-probabilistic methods like BERT modeling and more accurate emotion detection methods, including manual coding. The diverse thematic structure also limits the model's ability to identify framing patterns, suggesting the need for more focused analyses and supplementary manual coding to enrich findings.

This thesis is structured in a following way. The first chapter is the literature review where the theoretical objectives of the study are met and key concepts, research results within the field and the theoretical framework are outlined. Then, in the second chapter, I outline the methods for research, the process of data collection and database building, as well as the procedures and analyses through which the key methods were carried out. In the third chapter, I address all the assumptions and hypotheses based on the analyses results. Then follows a discussion chapter where I address the possible interpretation of the results in terms of framing functions. In the conclusion, I make a summary of the study results and discussion, with also addressing the limitations of the study and the opportunities for further research.

CHAPTER I.
SEXUAL VIOLENCE IN MASS MEDIA

This thesis analyzes the ways that Russian news media portrayed cases of sexual violence and framed the discussion on sexual violence in general through the years 2016 – 2023. In order to achieve the study goals, the exploration of the main concepts of the study – sexual violence and framing – as well as the patterns that are key to the media coverage of violence, has to be made. In this chapter, I explore the concept of sexual violence as the element of gender-based violence and then the concept of framing as a complex process through which media constructs the reality of crime and violence. I further explore the intersection of these concepts, overviewing the research on media portrayals of sexual violence as well as the interpretative schemas that media use to address the problem. As a result, I build the research objectives and hypotheses for the further analyses conducted in my thesis.

Sexual violence as the element of gender-based violence

Sexual violence as a concept, as well as the empirical study subject of social sciences and criminology research, lies within a broader scope of gender-based violence. In this subchapter, I define sexual violence within academic research, as well as in the human rights field, then connect this concept to the broader concept of gender-based violence. Further, I overview the common features of various cases of sexual violence and their relation to the popular narratives on the subject.

Defining sexual violence as a social problem

The World Health Organization (WHO) defines sexual violence as “any sexual act, attempt to obtain a sexual act, unwanted sexual comments or advances, or acts to traffic, or otherwise directed, against a person’s sexuality using coercion, by any person regardless of their relationship to the victim, in any setting, including but not limited to home and work” (Krug et al., 2002). In scholarly texts, sexual violence is closely related to the concept of “consent”, regarding its violation as a social contract and ways of determining whether the act of violence has occurred, and to the concept of “victim”,

which is often criticized in different fields of Western research tradition and activism for overshadowing agency, being replaced by wordings of "surviving" violence (Alcoff, 2009). Sexually violent acts perpetrated against women, as well as those perpetrated against men and children, are characterized by a high diversity of circumstances and settings, as well as by diverse relations between the victim and the perpetrator and forms in which such violence occurs. Dartnall & Jewkes (2013) differentiate between such forms of sexual violence as: "rape in marriage or dating relationships; rape of non-romantic acquaintances; sexual abuse by those in positions of trust, such as clergy, medical practitioners or teachers; rape by strangers; multiple perpetrator rape; sexual contact involving trickery, deception, blackmail or of persons who are incapacitated or are too drugged, drunk or intoxicated to consent; rape during armed conflict; sexual harassment" (p. 4) and others, with a common denominator of consent to engage in a sexual act not being given of not being given freely.

Sexual violence is recognized as an important social problem on the international level. Consensus on defining sexual violence as a problem is reflected in the results of the United Nations (UN) conference in China in 1995, where violence against women, including its sexual component, was identified as a problem within the social, health and sustainable development optics (United Nations, 1995), which led to a growing interest in research on violence against women in the social sciences in the following decade (McMurray, 2005). Research suggests that sexual violence against women is often a result of unequal power equations, both real and perceived, between men and women (Kalra & Bhugra, 2013) and also has severe consequences for those experienced such violence, therefore stating the need for adoption of major prevention programs on a societal level (Schwartz, 1999). Moreover, some scholars state that the consequences of male violence lie beyond just the victims of violence, and there are significant costs of intimate violence on a national level, including the government expenses on medical and mental health service delivery (Russo & Pirlott, 2006).

Gender-based violence

Sexual violence is usually recognized as a part of a broader form of violence, gender-based violence, the main feature of which is the perpetration of violence in a strong connection with gender relations or gender inequality – for example, violence against a partner (however, partner violence is not synonymous with gender-based violence and, as sexual violence, is a separate term within the gender-based violence umbrella). It is worth noting that both women and men can be perpetrators and victims of gender-based violence, although research is predominantly aimed at violence against women and identifies it as a separate concept of *gender-based violence against women*.

Despite the existence of studies, mainly based on the United States materials, that show the absence of significant differences in the scales of partner violence committed either by men or women (Archer, 2002; Frieze, 2005), most researchers agree on the fact that women are more likely to become victims of sexual violence, and, moreover, the typical perpetrator of such violence is a male acquaintance (Abrahams et al, 2004; Hamby, 2014; Dartnall & Jewkes, 2013; Ricardo & Barker, 2008). At the same time, the existence and high prevalence of cases of sexual violence against men is also recognized in the field of social sciences, however, researchers note that this form of violence has been understudied from both theoretical and practical points of view, and it is very difficult to assess its real scales (Russell, 2007; Zalewski et al. 2018). In particular, the difficulties in estimating the prevalence of sexual violence against men come from the fact that the male victim does not fit the conventional narratives about the victim of sexual violence (Armstrong et al, 2018).

The differentiation of gender-based and sexual violence into separate concepts of violence against women and violence against men is based on the general idea of violence as associated with power and control, as well as social hierarchies, where women are a vulnerable group in terms of gender inequality. Gender-based violence is also associated with other social statuses: for example, some research suggests that in countries where women have very low economic status compared to men, the

prevalence of physical and sexual forms of partner violence is lower because men are able to exercise economic and resource control over their wives (McMurray, 2005).

In addition, researchers note that the predictors, meanings and consequences of gender-based violence vary depending on who commits it and to whom it is committed: the gender of the perpetrator and the victim of violence is involved in the social perception of violence, for example in assessing the severity of the acts committed and the degree of seriousness of the consequences (Marshall, 1992(a), 1992(b)). Russo and Pirlott note that, in addition to situational and cultural differences in perceptions of violence, gender relations also play a significant role since "gendered inequalities at home and at work create gender differences in perceived entitlements and give different meanings to the resources women and men bring to their relationships" (Russo & Pirlott, 2006). Finally, the very image of the victim of sexual violence is culturally defined as feminine (Hollander, 2001; McMurray, 2005).

The common features of sexual violence

Typically, sexual violence occurs in a form that is far from "archetypal rape" (Armstrong et al, 2018). First of all, most victims of sexual violence personally know their attackers (Abrahams et al, 2004; Hamby, 2014; Dartnall & Jewkes, 2013; Planty, 2013; Ricardo & Barker, 2008). Also, some research suggests while most women experience unwanted sexual contact and sexual harassment in public places, they typically do not physically resist violence, with indirect refusals and facial expressions being the most common refusal strategies in the situations of sexual harassment (Graham et al., 2017). However, other research suggests that the lack of refusal isn't present in the cases of more "serious" types of sexual assault (Hollander & Rodgers, 2014).

Another pattern of sexual violence is the large scale of unreported cases, with sexually violent incidents missing in the official statistics and being available to study only through victimization survey studies (Fisher et al., 2003). The research suggests that one of main reasons of such high prevalence of unreported cases comes from the

fact that many victims of sexual violence, even in the cases of rape and sexual assault, do not acknowledge their cases as such forms of violence, even when the case has all the features of a corresponding crime (Cleere & Lynn, 2013). Furthermore, as Armstrong and colleagues (2018) note, the victims of sexual violence do not conform to social stereotypes in terms of race, class, age and sexuality.

Therefore, there is a significant distortion between the real characteristics of sexual violence cases and those in public cognition as well as in popular narrative representations, including media portrayals of such crimes.

Media framing and its role in the social construction of violence

The media is considered to be a primary source, or frame of reference, for public's attitudes and opinions about violence, crime and justice system. Since most people have no or too little personal experience with these issues and do not encounter a large number of crimes on an everyday level, they tend to rely on other sources for information about crime and violence with media being the primary one (Baranauskas, & Drakulich, 2018; Kort-Butler & Habecker, 2018; Kort-Butler & Hartshorn, 2011; O'Hear, 2020). Moreover, some scholars note that the majority of the public directly reports the information in the news to be their main way to form an opinion about crime (Pollak & Kubrin, 2007). Even when considering other possible sources for people's knowledge about crime and justice, such as personal victimization and network contacts, the media effects are outstanding in terms of their ability to shape attitudes towards crime (Kort-Butler & Habecker, 2018).

At the same time, the representation of crime and violence in the news and mass media is characterized by a certain degree of distortion of the real prevalence of crime, adding up to the distortion already present in the official crime statistics. The said distortion refers to a great degree of latent, or unreported, cases which are claimed to be "a more significant proportion of all crime at any given time" (Čvorović, 2022, p. 250). One way to overcome the problem of data distortion is to study crime through the survey

method which is commonly used in Western criminology[4]. Russian academia seems to follow this line of the crime research: with an academically recognized call to study latent crime (Иншаков, 2008), a victimization survey is conducted regularly by the Institute for the Rule of Law at the European University at Saint-Petersburg to create a possibility of estimating the number of crimes that are not included in official statistics for various reasons (Серебренников & Титаев, 2022).

Additionally to the distorted media representation of the prevalence of various types of crime, the characteristics of crime cases, victims and perpetrators portrayed in the media are often inverse to those reflected in official crime statistics, following the "law of opposites" (Pollak & Kubrin, 2007), and also convey many cultural stereotypes (Gruenewald, Chermak & Pizarro, 2013). Altogether, the news media is claimed to be "central to the process of constructing the social reality of crime" as, in the process of defining and framing problems, it "acts as a primer on crime and justice by providing tools that consumers can use to interpret information and events" (Kort-Butler & Habecker, 2018).

Therefore, *the process of framing* is one of the key aspects of how media portrays sexual violence and other social problems, and how audience, affected by this framing, shapes its attitudes toward it. In this part of literature review, I make an overview of framing research, starting with its psychological and sociological origins, and then moving to journalism studies where framing research got most of its modern theoretical and empirical development. Further I address framing as a complex process and research program, dive into the operational concepts of framing, and then move to the problem of frame detection within diverse methodological traditions.

[4] See, for example,
a) National Crime Victimization Survey (NCVS). URL: https://bjs.ojp.gov/data-collection/ncvs and
b) Crime Survey for England and Wales (CSEW). URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingseptember2023#related-links

Frames and framing: where do they come from

The research on frames and framing in communication texts is marked by the lack of an established conceptualization for both these terms. Not sharing a common theoretical model, framing studies show a wide variety of definitions and means for measurement for both *frame* and the *framing process* concepts, which, as many scholars point out, leads to a great vagueness of the research tradition, both in theoretical and empirical sense (Entman, 1993; Entman, Matthes & Pellicano, 2009; Guenther et al., 2023; Matthes & Kohring, 2008; Matthes, 2009; Saldaña Villa, 2017; Scheufele, 1999; Semetko & Valkenburg, 2000; Shoemaker & Reese, 2014). In overall, though, frames may be viewed as a mean to create a consistent meaning from different pieces of information, as some scholars, for example, see them either as "a central organizing idea or story line that provides meaning to an unfolding strip of events" (Gamson & Modigliani, 1987, p.143) or as a way of making sense of a "seemingly disconnected list of facts" (Shoemaker & Reese, 2014, p. 176).

In my thesis, I use Entman's (1993; also further Entman, Matthes & Pellicano, 2009) definition of a framing since, one the one hand, it is probably the most used way of conceptualizing frames in empirical studies (Saldaña Villa, 2017), and, on the other hand, as I will show in the following sections, it provides an opportunity for operationalization and measurement that no other attempt to define the concept of frame and the framing process offers. According to Entman (1993), *to frame* is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and / or treatment recommendation for the item described" (p. 52). In a more recent work, Entman and colleagues (Entman, Matthes & Pellicano, 2009) argue that the term "frame" comes from traditional sociology, where the concept was first introduced by Bateson (1972) and developed into a separate theory of frame analysis by Goffman (1974), though several scholars point out that framing can be traced back both to sociology and psychology (Guenther et al., 2023; Saldaña Villa, 2017; Scheufele & Tewksbury, 2007).

The psychological tradition mostly offers the link between ways of presentation of particular issues and events in media or discourse and ways in which people make sense of perceived information, furthermore being able to make moral judgements about said events and issues. Iyengar's (1996) study on attribution of responsibility for political issues, meaning who (or what, if responsibility is attributed not to a person but to a political or social entity), how, and why people hold responsible for these issues, may be somewhat of a textbook example of a psychological approach to framing. There, framing is understood as the "effects of presentation on judgment and choice" (Iyengar, 1996, p. 61), highlighting the influence that, as defined in sociologically-oriented communication studies, media frames have on cognitive characteristics of people themselves and the process of forming an opinion about an issue depending on ways of how the issue is presented. For Entman and colleagues (2009), whose definition of framing refers primarily to communication texts, Iyengar's definition (or any similar one) would not represent the framing itself but mainly the effects of framing and "psychological processes that underlie such effect(s)" (p.183). Putting aside conceptual inconsistency, it is important to note that psychological studies of framing traditionally use experiments as methods for discovering the effects of news frames. For example, Iyengar's (1996) participants were divided into two groups which were exposed to television news programs characterized by two distinct types of news framing, making it possible to reveal differences in attributions of responsibility depending on the narratives which develop in people's minds after receiving information about several political issues, such as poverty or crime, presented in a certain way.

The other significant part of research on framing utilized in the experimental method is represented by a much earlier work by Kahneman and Tversky (1984) which focuses on cognitive and psychophysical determinants of the decision-making processes in contexts where there is a certain degree of risk presented to people who are making a decision. There, the scholars demonstrate two problems (which later became a classic example of equivalence framing as opposed to emphasis framing – for the clarification of these types of framing see below in this section) constituted by the need to make a

decision in the field of health policy. The problems are presented by Kahneman and Tversky (1984) in a following way:

> Problem 1.
>
> Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.
>
> Which of the two programs would you favor?
>
> <…> Problem 2.
>
> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. (p. 343)

In these problems, the proposed pairs of programs to fight the outbreak are identical in terms of actual numbers of people saved by them yet completely different in terms of how these numbers are verbally presented. Program A, for example, implies that 200 of 600 people will be saved and that the remaining 400 will die, which is also true for program C. Yet the first one is worded in terms of potential gains and the second one in terms of potential losses, or, in other words, these two programs *frame* the issue of the "unusual disease" differently. There, according to the study, framing the same outcome in terms of potential gains or potential losses is key for people who are being exposed to this framing to make a certain decision: the authors report that program A was preferred by 72% of respondents and program C was preferred by only 22% of respondents in respective iterations of the experimental study (Kahneman & Tversky, 1984, p. 343).

Therefore, the origins of psychological approach are an important background for framing research in terms of empirically validating the idea that patterns of presentation

of information are causally connected to people's opinions on the matter and also their judgements and choices. Nevertheless, psychological work on framing shows very little theoretical foundation of the term "frame", though there were certainly some attempts to make such conceptualizations. For example, in the study on how the audience makes sense of issues based on news reports on the 1999 Kosovo crisis, Berinsky and Kinder (2006) distinguish two types of frames: frames in discourse and frames in cognition. While *frames in discourse* are those of a primary focus within sociologically oriented communication studies, marking, according to the authors, ways in which different political actors define essential issues and suggest how the audience should think about them, *frames in cognition* derive from cognitive psychology and indicate different models of understanding that people use in order to process events and issues, which results in a "coherent mental representation" (Berinsky & Kinder, 2006, p. 642) of received information, usually in a narrative form. Therefore, these frames in cognition are a primary focus of a psychological tradition of framing studies. This tradition is also divided, as Guenther and colleagues (Guenther et al., 2023) point out, into research on *equivalence framing*, or "based on prospect theory, dealing with different linguistic presentations of the same information" (p.4), which can be illustrated by Kahneman and Tversky's (1984) work discussed above, and *emphasis framing*, or "presenting a topic differently through selection and salience of content" (p.4), which is present in Iyengar's (1996) work also discussed at the beginning of this section, though some scholars attribute emphasis framing uniquely to sociological tradition.

The sociological tradition of framing research, as has been noted above, traces back to the classical work of Erving Goffman "Frame analysis. An essay on the organization of experience" (1974) where the concept of frame was originally taken from Bateson (1972) and developed into a separate theory within the micro-sociological tradition. As Sokolov (Соколов, 2022) summarizes, "Frame analysis" "is an eccentric attempt to understand how non-literalness of social life works – not in a way that certain events are possible to be differently interpreted, but more in a way of how the same event can have a certain meaning and at the same time not have this meaning" (p. 17).

Probably the most commonly used, by both scholars and university professors, illustration of Goffman's point concerning non-literalness of life and certain situations is *theatrical frame* (Goffman, 1974), which, to put is simply, is a situation of a non-literal interpretations of the actors' emotions to which the audience is exposed during acting due to the mutual understanding and interpretation of what constitutes the acting itself. This is closely connected to one of the main ideas of frame analysis that transferred from Goffman to the studies of framing in communication texts: people are involved in a constant process of interpretation of incoming information or events, where they apply "interpretative schemas, or primary frameworks" (Scheufele & Tewksbury, 2007, p. 12) to make sense of life events and, to put it more broadly, of the world around them. As scholars point out, frame theory is not a holistic theoretical construct, but rather is a set of different conceptual ideas, where the term "frame" has developed in many ways, very distinct from Goffman's original ideas, within such fields as linguistics or anthropology (Вахштайн, 2008; Соколов, 2022). Moreover, as Vakhstein (Вахштайн, 2008) argues, all of the concepts which constitute the scattered field of frame theory are "organized around the problem of contextualizing an action event, <…> [thus] the term "frame" is a blanket term for the term "context" (p. 78).

The notion of context, however, must be understood here with a certain degree of carefulness and be distinguished from contextualizing in terms of providing a background for social issues when we talk about the distinction between episodic framing (through which issues are "constructed around specific instances and individuals" (Entman, Matthes & Pellicano, 2009, p. 176)) and thematic framing (through which issues are put in a broader context, for example, are connected to corresponding trends) within the field of communication studies. It can be argued here, following Shoemaker and Reese's (2014) statement on how frames connect facts together in order to create meaning, that Vakhstein's (Вахштайн, 2008) notion of context refers to the same idea of meaning that is created by any type of framing. Therefore, for example, episodic framing, which attributes social problems only to certain people or other entities, create a context, though not in terms of a broader

background but in terms of contextualizing these problems within a space of individual responsibility (see Iyengar, 1996).

Despite appearing originally in both sociology and psychology, the process of framing received its most theoretical and empirical attention within journalism studies, which, in their modern stage, were influenced by namely sociological ideas. This influence marked a shift in the discipline addressed by scholars as the "sociological turn in the field" (Wahl-Jorgensen & Hanitzsch, 2009, p. 6) which brought critical perspective as well as new research problems and methodology into the studies of news media and journalism in general. Generally, within communication and journalism studies, the framing process is considered to be a part of research on *media effects*, or various outcomes of media use on both micro- and macro-levels (Valkenburg, Peter & Walther, 2016). There, framing may be viewed by scholars as either a theory of media effects (Scheufele, 1999) or one of media effects models (Scheufele & Tewksbury, 2007). The media effects research itself emerged in 1920s, as Valkenburg and colleagues (Valkenburg, Peter & Walther, 2016) point out, along with the concept of mass communication – though the focus on the media effects became prominent only 30 years later, boosted by the "introduction of television and the emergence of academic communication departments in Europe and the United States" (p. 317). Moreover, the whole field of mass communication studies is synonymous to the studying of media effects since the whole academic field is "based on the premise that the media have significant effects" (McQuail, 1994, as cited in Scheufele, 1999, p. 104).

It is common among scholars to define four major stages in studies of mass communication (Baysha, 2004; McQuail, 1994; Scheufele, 1999; Scheufele & Tewksbury, 2007; Wahl-Jorgensen & Hanitzsch, 2009). In the first stage (1920s – 1930s), mass communications studies focused on political propaganda during World War I with a premise that mass media has major influence on people's attitudes; the second stage (1940s – 1960s) was marked by an attention to people's personal attitudes which are essentially sustainable and can only be reinforced by the media, hence, moving forward from strong media effects to limited ones; the third stage (1970s –

1980s) introduced cognitive effects as a result of looking for new strong effects of mass media; finally, the fourth and present stage (since 1980s) is characterized by a turn to the idea of social constructivism (Scheufele, 1999). As Baysha (2004) puts it, "social constructivism explains the relationship between media and audiences by combining elements of both strong and limited effects of mass media" (p. 234), therefore, on the one hand, granting the media the power of constructing social reality through framing, and on the other hand, taking into account the fact that the audience has preliminary political and social attitudes, and also the fact that framing is highly influenced by the attitudes and ideologies of those who construct them – for example, journalists or organizations. Moreover, according to Wahl-Jorgensen and Hanitzsch (2009), this stage of research was also marked by the adoption of qualitative methodologies, "most notably ethnographic and discourse analytical strategies" (p. 6), which provided researchers with a more clear and detailed understanding of how the whole process of framing is structured.

Therefore, the combination of strong and limited media effects, the focus on different entities taking part in the framing process and the emergence of new methodological complexity created the modern tradition of research on media effects, where framing, among other media effects such as agenda setting and priming, is considered to be a complex process. In the next section, I will overview three attempts to develop integrated models of the framing process and framing research made by Scheufele (1999), by Entman, Matthes and Pellicano (2009) and by Matthes (2012), as well as other notions of various processes involved in framing.

The complexity of the framing process

The complexity of the framing process refers to the fact that frames are not just a mere attribute of a communication text, but are rather contained within strategic communications of political actors as well as in the minds of the audience, the latter addressed by Berinsky and Kinder (2006) as frames in cognition. Therefore, it is argued that the framing process happens within a cycle where news frames are influenced by

political actors and then compete with individual audience frames for a certain interpretation of an event or a problem (Saldaña Villa, 2017), which is also in line with Kinder and Sanders's (1990) notion that frames function as both "devices embedded in political discourse" and as "internal structures of the mind" (as cited in Scheufele, 1999, p. 106). On the notion that the news frame always has some primary influence, Shoemaker and Reese (2014) propose the hierarchical model, which consists of five levels of influence: "social systems, social institutions, organizations, routines, and individuals" (p. 8), stating that the shaping of news content happens on both macro- and micro-levels of social relations. Each of the levels may be further elaborated into separate lines of research, for example, the research on how journalists produce frames depending on individual, organizational and societal levels of influence, with such contextual factors as cultural availability of certain interpretation of events or interpretations offered by colleagues (Brüggemann, 2014). However, most of the research on framing is concentrated within single parts of the complex process and even single types of communicating actors, though frames in communication texts come from a set of arenas that compete for audience's opinions about certain issues (Saldaña Villa, 2017).

As I have pointed out at the beginning of the previous section, framing research is commonly considered to be built of many scattered concepts and is far from a consistent theoretical model. One of the most significant calls for a synthesis of communication research on framing in order to develop common conceptual schemas was made by Entman. In his program paper (Entman, 1993), the scholar states that, in order to gain a strong disciplinary status, communication studies have to produce new core knowledge: within framing research, there is a need to create "a general statement of framing theory" (p. 51), which would explain how exactly, for example, frames are embedded in texts or how they influence audience's thinking patterns. This statement was picked up and used as a theoretical premise by many scholars (see, for example, Guenther et al., 2023; Matthes & Kohring, 2008; Matthes, 2009; Saldaña Villa, 2017; Scheufele, 1999; Semetko & Valkenburg, 2000; Shoemaker & Reese, 2014), though the

idea of a single consistent theoretical paradigm as the right path for the framing research development was also criticized. The major critique of Entman's proposition was made by D'Angelo (2002) in his paper on the multiparadigmatic nature of framing research. There, following Lakatos (1974), he argues that there should not be a "mended" theoretical paradigm of framing, because the "knowledge about a phenomenon grows within an environment, called *a research program*, that both supports competition among different theories and provides criteria to evaluate individual theories in light of new data" (D'Angelo, 2002, pp. 870-871). Therefore, originating from the same problem of vagueness within a research field, the arguments about its further development are quite opposite to each other. Nevertheless, both D'Angelo and Entman agree on the complex nature of the framing process, with D'Angelo taking into account Entman's statement about framing occurring in several locations within the communication process – "the communicator, the text, the receiver, and the culture" (Entman, 1993, as cited in D'Angelo, p. 873).

Following Entman's call for a clear conceptual framework, there was an attempt made by Scheufele (1999) to combine all the fragmentated approaches to framing into a one comprehensive model. Within the proposed model, framing is divided into three elements – inputs, processes, and outcomes, – which combine together different elements of framing process, such as frame building and effects of framing, as well as different types of frames and actors within the framing process, understood as a cycle of influences (Scheufele, 1999; illustration in Fig. 1).

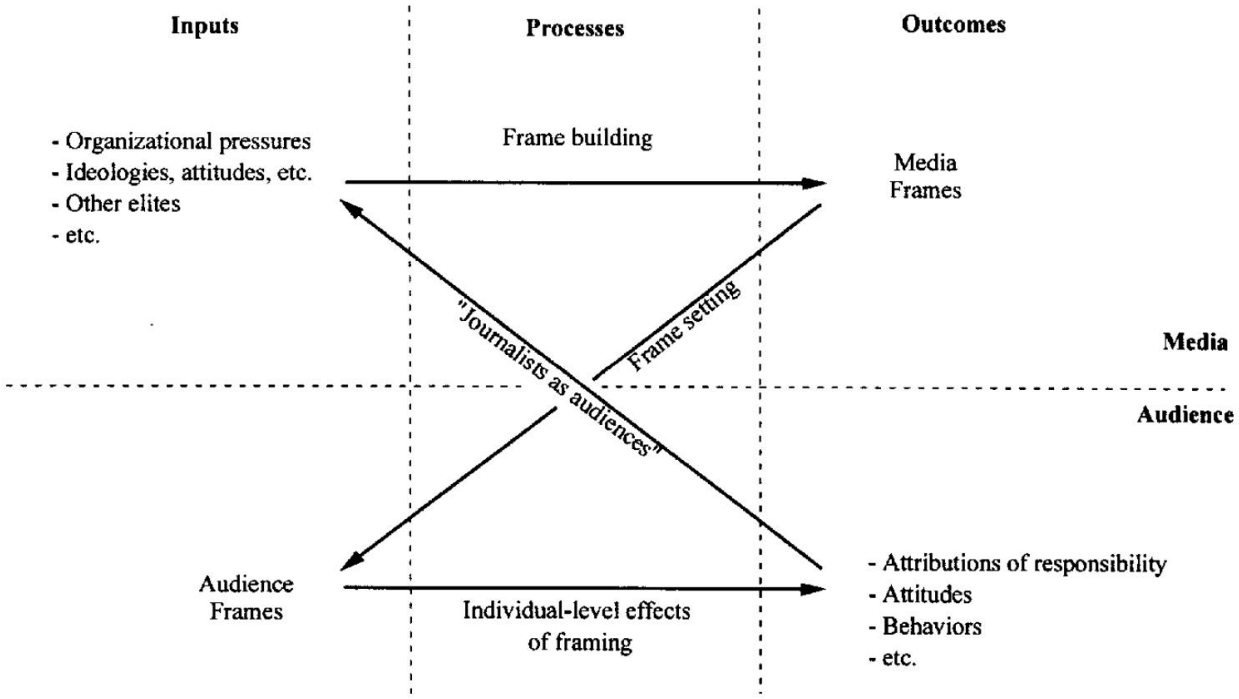

Figure 1. The process model of framing research (Scheufele, 1999)

The cyclic nature of the framing process was also acknowledged by Entman, Matthes and Pellicano in their collaborative paper (2009), although the main statement within the complex nature of framing was that the process is highly influenced by the time variable. On this matter, Entman and colleagues (2009) argue that the process of framing is *diachronic* in a way "that exposure during a given period is presumed to increase probabilities of particular responses during a future period, while diminishing the probability of thinking about other potentially relevant objects or traits" (p. 177; illustration in Fig. 2). The other important notion of the continuity of the framing process is that certain frames are finite, as once they appear in communication texts enough times to be internalized into personal cognitive schemas, the need of usage of these frames fades through time (Entman, Matthes & Pellicano, 2009). It is important to note here that the discussion on sexual violence against women present in the Russian news media has been going on for many years, and hence the notions on continuity of the

framing process may be true for the empirical object of this study. Therefore, I include the hypothesis based on the change of topics over time into my analysis.

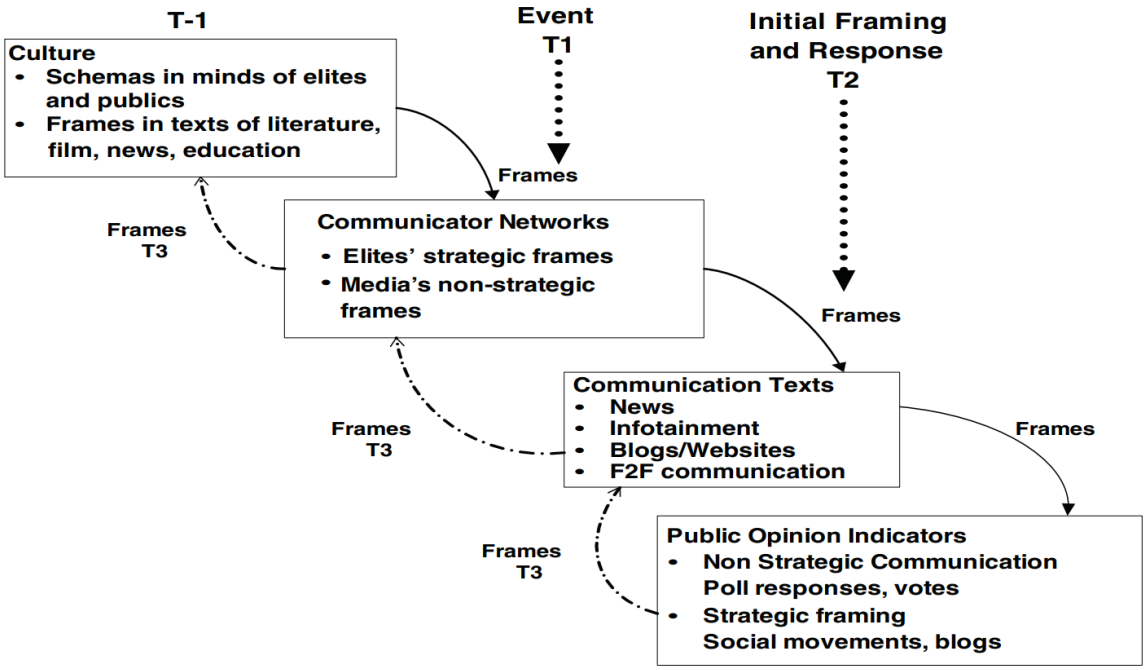

Figure 2. Diachronic framing process (Entman, Matthes & Pellicano, 2009)

The third and the last notable attempt to put together many scattered research traditions was made by Matthes (2012) who developed an integrated model for framing research. Following D'Angelo's notion of framing as a research program, Matthes (2012) proposes an approach which doesn't imply an existence of a single framing paradigm but rather proposes a way to empirically analyze framing through three perspectives: media actors, political actors and citizens as actors (illustration in Fig. 3).

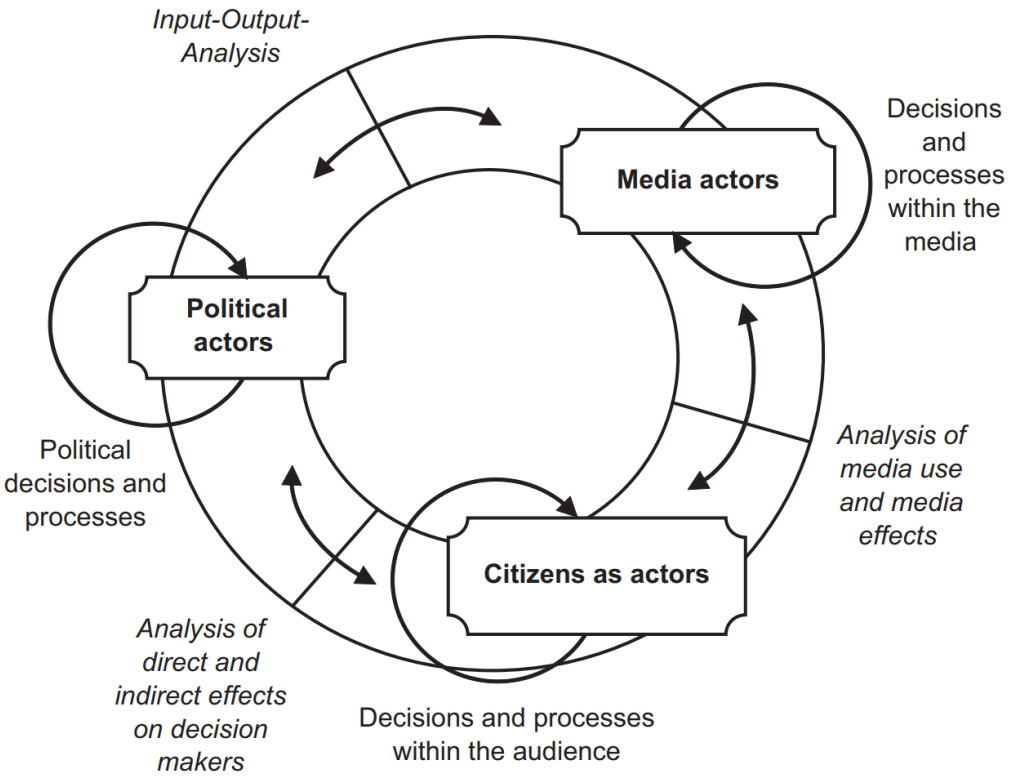

Figure 3. The illustration of the design of the interdisciplinary research on framing (Matthes, 2012)

Mattes' integrative approach was adopted by Saldaña Villa (2017) in her dissertation on identifying frames and frame functions within the media portrayal of the earthquake in Chile. There, she analyses how three "key actors of the framing process – the media, audience, and political elites, compete for their own frames to become salient and shape public opinion" (p. x), being all represented by three sets of data – news stories, online comments and government press releases respectively. In my thesis, I will concentrate on single-message frames as representing one of the key actors of the framing process – the media, – though taking into account that the research may be further broadened by inclusion of other framing process actors described by Matthes into the analysis.

So, what's in a frame? The need for clear operationalization

The conceptualizations of framing within the framing research literature usually fall into two genres of definition (Entman, Matthes & Pellicano, 2009; Matthes, 2009).

First genre may be described as *defining frames in general terms*, where all the conceptual definitions, as Matthes (2009) puts it, despite being useful, "leave the explicit operational understanding of the frame concept open" (p. 350). The textbook example is aforementioned Gamson and Modigliani's (1987) description of the frame as ''a central organizing idea or story line that provides meaning to an unfolding strip of events'' (p. 143). Within the same genre is Gitlin's (2003) definition where the frames are understood as ''principles of selection, emphasis, and presentation composed of little tacit theories about what exists, what happens, and what matters'' (p. 6). Some scholars define frames in more a more specific way in line with their disciplinary affiliations (see, for example, framing within social movements research – Fomin, 2022). There, framing may be described as "how elites compete to define issues their way" (Berinsky & Kinder, 2006, p. 640), marking the framing process more as a part of political strategy rather than viewing it more broadly as an attribute of how information is presented in media and discourse, yet still in line with the genre of general definition which is not applicable to the further method for detecting a frame within a specific discourse.

The other genre of definition is *defining framing in terms of framing functions*, or what "frames generally do" (Entman, Matthes & Pellicano, 2009, p. 176; Matthes, 2009, p. 350). This genre is mainly rooted in Entman's (1993) definition of framing, which is based on selecting some aspects of information and making them salient in a communication text. The specification of the concept "frame" by Entman and colleagues (Entman, Matthes & Pellicano, 2009), who also add the time perspective to its original definition, is following:

> A frame repeatedly invokes the same objects and traits, using identical or synonymous words and symbols in a series of similar communications that are concentrated in time. These frames function to promote an interpretation of a problematic situation or actor and (implicit or explicit) support of a desirable response, often along with a moral judgment that provides an emotional charge (p.177)

These definitions, though, are more of a combination of several framing characteristics, whereas the framing functions themselves were introduced by Entman's in his original work "Framing: Toward Clarification of a Fractured Paradigm" (Entman, 1993) in a following way:

> Frames, then, *define problems* – determine what a causal agent is doing with what costs and benefits, usually measured in terms of common cultural values; *diagnose causes* – identify the forces creating the problem; make *moral judgments* – evaluate causal agents and their effects; and *suggest remedies* – offer and justify treatments for the problems and predict their likely effects (p. 52).

Here, it can be argued that the second genre of definition, where the frames are conceptually explained though their functions, provides clear guidelines for operationalization. These operational definitions, therefore, "can be translated to frame indicators" (Matthes, 2009) and provide validity to framing research in terms of detecting frames in communication texts, where scholars can be sure that what they are detecting is truly a frame and not other textual features such as themes. It is important to note here that a frame does not necessarily contain all the functions, just as one text can contain several functions at the same time (Entman, Matthes & Pellicano, 2009).

Entman's framing functions were adopted in the aforementioned research by Saldana Villa (2017) in her study of framing the Chilean earthquake. There, the main body of research was based on topic modeling of the discussion on the earthquake with media, government and public as three main actors of the framing process: Saldana Villa modeled a 20-topic thematic structure of her textual data where topics were then the main units of analysis in terms of hypotheses testing. Then, after carefully analyzing the topics and the difference of topic usage between various actors and outlets, she generally interpreted the topics and their appearance in discourse in terms of framing functions. The strong side of this approach is that the researcher made no premise on defining frames themselves within the work but rather premised that some thematic features of the discussion may yield the same functions as frames (as in, framing functions). This approach minimizes the risk of interpreting themes as frames even if the themes are not

frames themselves, which was the core of Matthes' (2009) notion on clear operationalization. In my thesis, I will adopt the same approach with the inductive strategy that will be discussed below.

The very many types of frames

The literature on framing provides several typologies of frames which are useful to consider regarding operational validity. The initial distinction is the one between *issue-specific frames*, which are pertinent only to specific topics or events, and *generic frames*, which can be identified within different thematic scopes and are not limited by the time of occurrence or different cultural contexts (De Vreese, 2005). Issue-specific frames are mostly used in research of thematically narrow issues, such as coverages of certain wars or debates on short-term political scandals (Entman, Matthes & Pellicano, 2009). On the contrast, generic frames may be applied to many issues, especially broad ones like the long-lasting debate on sexual violence.

Within generic frames, there are several typological distinctions proposed in literature. One suggestion is to differentiate between procedural framing, where the emphasis is made on certain political strategies or struggle between the elites, and substantive framing which concentrates on the essential characteristics of the issues (Entman, 2004). A more general, not binded exclusively to politics but appliable to media framing, typology is Iyengar's (1996) differentiation between *episodic and thematic frames*. There, episodic frame stands for merely an illustration of some issue, depicting it though concrete events or people (for example, news may concentrate entirely on some poor person or a violent case), while thematic frame illustrates issues in a more abstract way yet puts them into a wider context (for example, containing this type of frame, a communication text on poverty is concentrated on statistical facts rather than a poor person). According to Iyengar (1996), while no news reports are purely thematic or episodic, one of the framing types is nevertheless significantly more prevalent than the other. With this premise, Iyengar experimentally analyzed the effects of television news framing on audience's patterns for attributing responsibility, with the

key conclusion that episodic coverage of news results in a more individualistic attributions of responsibility, meaning that, in regard of social issues, people exposed to thematic framing tend to blame the suffering individuals rather than the government or society.

Iyengar's (1996) conclusions were further challenged by Semetko and Valkenburg (2000) in their research on European politics displayed through press and television news. In regard of attribution of responsibility, their study showed that television news may contain the episodic frame and the same time hold the government responsible for the social issues. However, the scholars highlighted the potential influence of political culture on the effects of framing since the study was made in Holland, "where there is a strong social welfare state, [and] the government is expected to provide answers to social problems" (Semetko & Valkenburg, 2000, p. 106). Moving back to the typologies, Semetko and Valkenburg's (2000) paper is also notable for distinguishing five generic frames – attribution of responsibility, conflict, human interest, economic consequences and morality – which, following a deductive approach, were predefined by previous studies and coded for content analysis. One of the results of the study that is important to note is the media difference in framing which was found not between media types (press vs. television news) but between serious and sensationalist newspapers. Sensationalism itself is an important concept within this thesis since the media tends to sensationalize the problem of sexual violence (see next subchapter). As Semetko and Valkenburg (2000) state, serious newspapers are more likely to use attribution of responsibility and conflict frames, while sensationalist newspapers tend to use human interest frame. However, while providing a relevant result, the scholars didn't operationalize the seriousness of the newspaper and relied on their own expert understanding of Holland's media context.

One of the other possible distinctions in framing research that is present in literature is the distinction between equivalency framing and emphasis (issue) framing. The latter type of framing is concerned with an "emphasis on a subset of potentially relevant considerations [that] causes individuals to focus on these considerations when

constructing their opinions" (Druckman, 2001, as cited in Entman, Matthes & Pellicano, 2009, p. 182), while emphasis framing occurs when the same issue is differently formulated, as in Kahneman and Tversky's (1984) experiments on audience perceptions of health policy problems. However, as Entman and colleagues (2009) note, equivalency and emphasis framing are more of the *framing effects* types rather than frame types since equivalency framing presents the logically same pieces of information in different forms that alter audience's perceptions of them, while emphasis framing is constituted by different pieces of information with the highlight on the more important ones, being "concerned with increasing or decreasing the salience of an issue" (p. 182). Therefore, in terms of functionality, emphasis framing is essentially in line with the conceptual meaning of a frame adopted in my thesis, where I follow Entman's idea of *selection and salience* being the mechanisms that constitute framing as such.

Frame detection and measurement

Framing research is methodologically diverse and provides several general strategies for the measurement of media frames in communication messages. Entman and colleagues (2009) identify four methodological approaches for detecting frames: qualitative approach, manual-holistic approach, manual-clustering approach and computer-assisted approach. Matthes and Kohring (2008) differentiate between five approaches, though pointing out that they are not mutually exclusive: hermeneutic approach, linguistic approach, manual holistic approach, computer-assisted approach and deductive approach. Matthes (2009) stresses out that methodological steps in framing research depend on several overlapping features: "(a) whether the analysis is text-based or number-based, (b) whether frames are determined inductively or deductively (c) whether coding is manual or computer-assisted, and (d) whether data-reduction techniques are used to reveal frames or whole frames are coded as such" (p. 351). Moreover, the methodological strategy depends on a more general consideration of frames as either dependent or independent variables (Scheufele, 1999).

Probably the most general methodological distinction within frame detection strategies is the distinction between deductive and inductive approaches. *A deductive approach* is constituted of a literature-driven a priori definitions of the frames that are being measured in a text or other form of a communication message. The prime method used within this approach is content analysis where the researcher defines operational features of frames and then codes texts according to the prevalence of such features in order to detect frames or framing elements (see Matthes & Kohring, 2008; Semetko & Valkenburg, 2000), though some scholars see this method as too simplistic and failing to consider the differences in the salience of certain frames (Entman, 1993). The main limitation of a deductive approach is that, though it enables the researcher with an easy replication of steps, an ability to work with large samples and comparative quantitative advantages, it is highly likely to overlook new frames that are not prevalent in previous research and therefore are not expected to appear beforehand (Matthes & Kohring, 2008; Saldaña Villa, 2017; Semetko & Valkenburg, 2000). Therefore, the researcher must have a clear vision of a set of frames they expect to find in a text, which is not applicable to the studies within specific fields or contexts.

Opposite to deductive approach is *an inductive approach*. The main feature of this strategy is an open technique of detecting frames "with very loosely defined preconceptions" of the frames and their elements (Semetko & Valkenburg, 2000, p. 94). Therefore, frames emerge from the material itself, though the approach is commonly criticized for being hard to replicate (de Vreese, 2005). The methodology within the inductive approach is quite diverse since both qualitative and quantitative, as well as mixed methods are used in order to detect frames in texts. Within the quantitative methodology, the inductive way of deriving frames is prevalent in half of the existing papers (Guenther et al., 2023; Matthes, 2009). One way of frame detection within this approach which is fully automated is frame mapping which essentially is automatic detection of frames within a large corpus of texts (see Miller, 1997); the prime example of this approach is topic modeling for studying the thematic structure of communication texts though it is important to note that topics gathered by such modeling are not

necessarily interpretable as frames. In general, the inductive approach offers the potential to work with big amounts of textual data, as well as to detect new framing patterns that may be overlooked within the deductive approach that requires a priori set of frames. In my thesis, I will implement the inductive strategy of frame detection with topic modeling as prime method, also taking into account the Saldana Villa's (2017) approach of detecting frames not directly but rather of interpreting certain thematic features of communication texts in terms of framing functions.

Media coverage of sexual violence: evidence from research

Media coverage of rape tends to be characterized by "rape myths", in particular images of the actor as a "monster" and the victim as either a "virgin" or a dishonorable woman who is responsible for the violence; this is harmful to victims when people who share such myths are involved in the criminal justice system (O'Hara, 2012, Benedict, 1993). The classical media image of the "virgin" also resonates with the denial of agency of victims of sexual violence, who are represented in news stories about violence as unable to protect themselves and making no attempt to do so, which contradicts survey data regarding protection in violent situations (Hollander & Rodgers, 2014).

Furthermore, media coverage of sexual violence is framed in the optics of mutual (i.e., on the part of both the actor of violence and the victim) responsibility for the violence committed (Easteal et al, 2015), and readers' comments on news articles about sexual violence are represented by distrust and doubt about the validity of the evidence of the violence committed (Harmer & Lewis, 2022).

The other characteristic of the coverage of sexual violence in mass media is sensationalism that has been present in the coverage of sexual violence at least since the 18th century (Benedict, 1993; Block, 2002). The sensationalist nature of covering the cases of violence is present when the violent cases follow the stereotypes about rape, such as "stranger danger" and other "rape myths", making the media, on the one hand, are "fascinated" with the stories of sex crime, and on the other hand, ignore the real, most common conditions under which such crime occurs (Hindes & Fileborn, 2020;

Serisier, 2017). The examples of the ways of sensationalizing the stories of sexual violence that are commonly used in media are "excessive detailing of the acts of sex and violence, exaggerated use of color headlines and emotional rhetoric, invasive pictures, and so on" (Lemish, 2004, p. 44). In overall, the sensationalistic approach is believed to be a strategy to attract more audience's attention (Boranijašević, 2018). The sensationalism has also been connected to the portrayals of the victims as traumatized and unable to perform as the active agents that are able to deal with the consequences of the violence and recover (Alcoff & Gray, 1993).

Research objectives and hypotheses

This research on framing of sexual violence against women in Russian news media is targeted at determining the ways of coverage of sexual violence that newspapers use, including finding specific themes within the discussion and their further interpretation in terms of the framing functions. The text analysis used in the exploration of the articles is implemented in an inductive approach, where the themes and framing patterns within the discussion are not defined a priori or coded – the themes are expected to be derived through the automatic methods of topic retrieval from the corpus. Therefore, the primary research objective of this thesis is the following:

RO1. To identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media.

The RO1 is essentially an explorative task and therefore doesn't require any hypothesis testing. However, the literature review allows me to make several assumptions on what the thematic structure of the discussion will be and what features of the media coverage of sexual violence against women may be expected to be observed in the texts of articles within the discussion.

As described in the literature review, the coverage of sexual violence in mass media is often characterized with "rape myths" and undermining the agency of the women who were raped, assaulted or experienced sexual violence in other forms (Alcoff & Gray, 1993). Therefore, the first assumption regarding the RO1 is the following:

A1. There will be themes within the thematic structure of the discussion on sexual violence against women that indicate the absence of agency of the victims of sexual violence.

Moreover, according to the literature, there are patterns of shared responsibility placed upon the victim as well as the perpetrator for the sexual violence (Easteal et al, 2015). Moreover, the framing research suggests that different mass media sources use different attribution of responsibility patterns, either attributing the problem to individuals or to larger entities such as the government or society (Iyengar, 1996; Semetko and Valkenburg, 2000). Connecting these two fields of research, I construct the second assumption:

A2. Within the thematic structure of the discussion on sexual violence against women, newspapers will tend to attribute responsibility for the violence to individuals rather than the government or society.

One the features of the coverage of sexual violence against women in mass media is that the cases of sexual violence are tend to be sensationalized, or presented in a form of unexpected, even shocking, individual situations, rather than portrayed in terms of a wider context and connected to sexual violence as a social problem (Alcoff & Gray, 1993**;** Benedict, 1993; Block, 2002; Hindes & Fileborn, 2020; Lemish, 2004; Serisier, 2017). Since sensationalism is understood as the "excessive detailing of the acts of sex and violence, exaggerated use of color headlines and emotional rhetoric, invasive pictures, and so on" (Lemish, 2004, p. 44), I expect to find sensationalism in a form of usage of detailed descriptions of violence, as well as in a form high emotional coloring in coverage of the violent cases. Therefore, the third assumption regarding the thematic structure of the coverage of sexual violence is the following:

A3. There will be a sensationalist rhetoric present in the Russian news media articles that cover sexual violence.

To address the connection of sensationalism in a form of the emotional coverage of sexual violence to the patterns of framing the sexual violence, I will examine the tone of the articles in their relation to the dominant themes within the discussion on sexual violence. The corresponding research objective is the following:

RO2. To explore the relationship between the themes present in the articles on sexual violence against women and the tone that is used by Russian news media to cover such violence.

Since sensationalism is mostly connected to the coverage of the cases of sexual crimes and also to reducing the problem of sexual violence to individual cases, I expect to find emotional rhetoric only in connection to specific themes within the corpus. Therefore, the hypotheses for the RO2 are the following:

H1. The themes that explore the cases of sexual violence will be correlated to either positive or negative sentiment in the coverage of sexual violence.

H2. The themes that explore sexual violence as a social problem will be correlated to the neutral sentiment in the coverage of sexual violence.

H3. There will be no correlation between the other themes and level of sentiment in the articles covering sexual violence against women.

Finally, the literature suggests that there are some temporal characteristics of media framing of the social problems present in the newspaper articles. Particularly, Entman and colleagues (2009) argue that the process of framing is diachronic in a way "that exposure during a given period is presumed to increase probabilities of particular responses during a future period, while diminishing the probability of thinking about other potentially relevant objects or traits" (p. 177). Therefore, the third research obkective of my thesis is the following:

RO3. To explore the temporal differences in the thematic coverage of sexual violence against women in Russian news media.

To achieve this task, I will examine the temporal differences in the coverage of sexual violence for the themes that are generally present in the discussion as a whole, excluding, for example, the possible issue-specific and situational themes that are, by essence, temporal. As Entman and colleagues (2009) argue, certain frames are finite, as once they appear in communication texts enough times to be internalized into personal cognitive schemas, the need of usage of these frames fades through time. Therefore, the hypothesis that will be tested in regard of the third research objective is the following:

H4. The general themes within the discussion on sexual violence in Russian news media will tend to be fading over time in terms of their presence in the articles.

CHAPTER II.
DATA COLLECTION AND METHODS FOR ANALYSIS

This thesis is made on how the Russian media frame sexually violent cases and the problem of sexual violence itself. The main purpose of the research, therefore, is to identify the ways through which the cases of sexual violence against women and of the discussion on sexual violence against women in general are covered by the Russian news media, as well as the interpretive frameworks, or frames, through which such coverage is carried out. To achieve this goal, I set several research objectives: (1) to identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media, (2) to explore the relationship between the themes present in the articles on sexual violence against women and the tone that is used by Russian news media to cover such violence, and (3) to explore the temporal differences in the thematic coverage of sexual violence against women in Russian news media. Therefore, the empirical research object of this thesis is the texts of Russian news media articles that describe cases of sexual violence against women or contribute to the discussion on sexual violence against women in other ways. The final study sample consisted of 15 342 texts of the articles on sexual violence against women published by 65 Russian news media outlets throughout the period of years 2016-2023, which then were analyzed through various methods to explore the characteristics of the coverage of sexual violence in Russian news media.

In the methodological chapter, I first define key concepts of the study, then describe methods used for the analysis and the process of data collection in detail. Further, I describe the process of model fitting, analysis and creating the variables that describe the thematic and emotional structure of the articles. I finish the chapter with the structured overview of the methods used for the hypotheses testing.

Main concepts of the study

Sexual violence is "any sexual act, attempt to obtain a sexual act, unwanted sexual comments or advances, or acts to traffic, or otherwise directed, against a person's sexuality using coercion, by any person regardless of their relationship to the victim, in

any setting, including but not limited to home and work" (Krug et al., 2002). Moreover, sexual violence, occurring in many forms, is always characterized to the absence of consent to the sexual act (Alcoff, 2009; Dartnall & Jewkes), and is also recognized as a social problem (United Nations, 1995).

The definition of a frame adopted in this thesis is based on the genre of defining frames in terms of framing functions, or what frames generally do (Entman, Matthes & Pellicano, 2009). Therefore, *to frame* is to "select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and / or treatment recommendation for the item described" (Entman, 1993; p. 52).

The definitions of the framing functions are also taken from Entman (1993). *Problem definition* is a function based on determining what a causal agent is doing with what costs and benefits, usually measured in terms of common cultural values. *Diagnosing causes* is identifying the forces creating the problem. Making *moral judgments* is evaluating causal agents and their effects. *Remedies suggestion* is a function based on offering and justifying treatments for the problems and predicting their likely effects.

Issue-specific frames are pertinent only to specific topics or events, and *generic frames* can be identified within different thematic scopes and are not limited by the time of occurrence or different cultural contexts (De Vreese, 2005).

Episodic frames stand for merely an illustration of some issue, depicting it though concrete events or people, while *thematic frames* illustrate issues in a more abstract way yet puts them into a wider context (Iyengar, 1996).

Sensationalism in the cover of sexual violence by the news media is a rhetoric based on presenting the issue in terms of unexpected, individual cases, based on common myths about the issue (Benedict, 1993; Block, 2002; Hindes & Fileborn, 2020; Serisier, 2017). The sensationalism usually happens in forms of "excessive detailing of the acts of sex and violence, exaggerated use of color headlines and emotional rhetoric, invasive pictures, and so on" (Lemish, 2004, p. 44).

Methods for data analysis

One of the research tasks of this thesis is finding the dominant themes in the discussion on sexual violence in Russian mass media through years 2016 – 2023. For this task, I use topic modeling as a method that allows to automatically determine the common themes in the texts based on the co-occurrences of certain words in them. For the task of exploring the connection between the themes and sensationalism, I use sentiment analysis with the further correlation analysis to connect the retrieved topics to the levels of article sentiment. For the task of determining whether the framing changed in time, I used the one-way ANOVA to explore differences in topic prevalence between the groups formed by the year of publication.

In this subchapter, I dive into topic modeling and sentiment analysis as the main methods of my research, which allowed me to discover a thematic structure of the discussion on sexual violence as well as to create variables used in further analysis.

Topic modeling: a strategy for discovering hidden thematic structure in texts

As already stated in the literature review section, journalism studies in their modern form have adopted many automated research strategies to work with big amounts of textual data which became common within both framing research and the media studies field as a whole. In general, texts analysis methods can be divided into two groups: deductive methods, which require pre-defined categories and researchers' clear understanding of what these categories contain, and inductive methods, which are aimed to explore the texts in order to find such categories of analysis. Günther and Quandt (2016) make this distinction between these two branches of texts analysis methods as well as differentiate between several methods within these branches. For the branch of deductive methods which require defining categories prior to the analysis, they outline three methods: text extraction, or rule-based, methods, dictionaries and supervised learning methods. As for the branch of inductive methods, they differentiate between document clustering and topic modeling (see Fig. 4 for a roadmap of text analysis approaches). The second branch of methods is key for this thesis which main

aim is to map the ways that Russian mass media frame the discussion on sexual violence on the large amount of textual data with a certain level of uncertainty of what constitutes these ways of framing. There, it is important to make a distinction between two common methods of inductive text analysis, document clustering and topic modeling, since both these methods deal with automated dividing textual data into different categories and detecting the inner topic, or thematic, structure of documents.

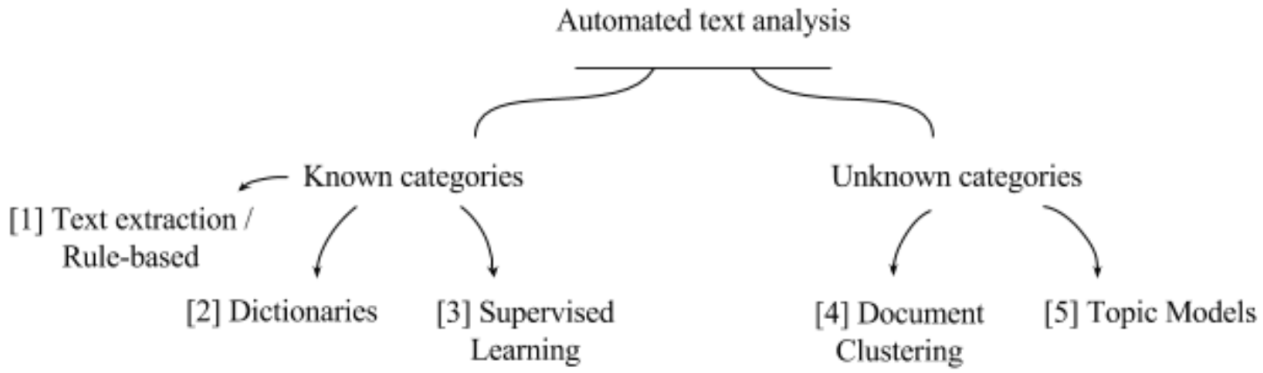

Figure 4. Overview of common text analysis methods (Günther & Quandt, 2016)

Both document clustering and topic models are built on word frequencies – how often do certain words co-occur with each other, though document clustering is aimed to divide documents into different categories where every document belongs to only one category, and topic modeling is aimed to divide parts of the texts into different categories, therefore allowing the multi-categorizing each of the documents. Günther and Quandt (2016) characterize topic modeling approach as the one working "with a statistically more sophisticated model" than document clustering, since "this approach describes each document as a mix of several latent topics and provides a thematic representation of the text collection" (p. 77). Taking into consideration the typical length and complexity of mass media articles (compared, for example, to short and thematically narrow texts, like tweets or online comments on social media platforms), topic modeling is an appropriate method for mapping the patterns of the discussion on

sexual violence against women with taking into account the potential thematic diversity both in the text corpus as a whole and also within individual documents.

Topic modeling methodology, according to DiMaggio et al. (2013) may be defined as "a suite of machine learning methods for discovering hidden thematic structure in large collections of documents" (p. 577) which produce interpretable topics (constituted of sets of words that are have a high probability of co-occurrence within the same documents) and hold such advantages for social science researches as the ability to code large amounts of data that cannot be coded manually as well as the potential to unravel hidden thematic structures that may be overseen even through manual coding. Various topic models are, in essence, probabilistic models since they treat "large collections of text as observations that arise from a generative probabilistic process that includes hidden variables which in turn reflect the thematic structure of a text" (Saldaña Villa, 2017, p. 57). This means that, again, the base of any topic modeling algorithm comes from a premise that there are groups of words that are more likely to co-occur within a same document and which form variables representing a consistent latent structure of themes that can be discovered in texts.

Topic modeling as a method was first introduced by Blei, Ng and Jordan (2003) and was built upon the algorithm of Latent Dirichlet Allocation (LDA), "a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics", and where each topic is "modeled as an infinite mixture over an underlying set of topic probabilities" (p. 993). Topic modeling of documents using the LDA method, thus, is based on a premise that each document given to a model input is constituted by a set of certain topics, where each of the topics is constituted by a certain distribution of words – in terms of the probability of being encountered within a given topic. Thus, the LDA topic model, given a set of documents on the input, generates on output 1) a distribution of words for each topic and 2) a distribution of topics for each document (Burns et al., 2011).

It should be noted, again, that this model, as well as all other models within the topic modeling family, is probabilistic and therefore requires major textual data

preparation since it deals with documents represented as “bag of words”, where the internal structure of the document in terms of words and / or sentences order and placement, as well as the grammar, is not taken into account (Qader et al., 2019). Nevertheless, the LDA itself is a strong method for working with unlabeled data and was proven to be a more efficient than many other common computer-assisted text analysis methods. For example, Guo et al. (2016) found that LDA outperforms another, “perhaps the most popular automated analysis approach in social science research” (p. 350) – the dictionary-based analysis which is based on an a priori construction of a set of keywords corresponding to the topics, also defined a priori, for the algorithm to detect in textual data. The researchers studied the two methods on a corpus of 77 million tweets about the 2012 U.S. presidential election and stated that, compared to the dictionary-based analysis, LDA had better performance scores: it was able to interpret more tweets and in a more nuanced and detailed way; there was also the advantages in terms of cost-efficiency and results validity. Moreover, Barde and Bainwad (2017), comparing the performance of different inductive text-analysis approaches (as opposed to the research cited above), found that LDA had “a better disambiguation of words and a more precise assignment of documents to topics” (p. 748) in comparison to other topic modeling techniques, such as, for example, probabilistic latent semantic analysis (pLSA).

LDA in its initial configuration though has many limitations. Such limitations include, for example: 1) the representation of the document as bag-of-words which doesn’t take into account the linguistic structure of the document, 2) difficulties to model non-independent, or correlated, topics, 3) disregard of any other topic structure of the documents. These limitations were pointed out by computer scientists and were considered as a ground to develop several extensions of the LDA model to make it more general and suitable for various types of complex textual data.

One of the extensions built for the classical LDA model is Hierarchical topic models based on the same allocation with a Bayesian perspective to creating a hierarchy of topics. The Hierarchical Latent Dirichlet Allocation model (or shortly, hLDA model) was developed by Blei, Griffiths, Jordan, and Tenenbaum (2003) and is intended to

account for complex, incommensurate structures of textual datasets used for topic modeling. This model, on the one hand, allows to build hierarchies of topic within text document with a given model depth, and, on the other hand, doesn't require an a priori setting of the k number of topics since it operates through hierarchical clustering. Having tested the hLDA model on the simulated data, authors claim that the model resulted is flexible and general for topic hierarchies, and also "naturally accommodates growing data collections" (p. 7).

To account for the other limitation of the standard LDA analysis – its inability to model topic correlations – Blei and Lafferty (2006) developed the Correlated Topic Model (CTM), where the variability of topic proportions was derived from the logistic normal distribution instead of a single Dirichlet distribution. The CTM model was tested on a dataset containing articles from the journal Science, where it showed a better performance than a simple LDA model and also allowed to make natural visualizations from the dataset. However, as Li and McCallum (2006) note, though the CTM model allows researchers to retrieve non-independent topics from textual data, only pairwise correlations are modeled because of the specific nature of a logistic normal distribution it deals with. In order to overcome such a limitation, Li and McCallum (2006) developed the Pachinko Allocation Model (PAM), which was able to capture arbitrary topic correlations and showed better results in terms of keywords coherence and topic interpretability in comparison to the hLDA (Blei et al., 2003) and CTM (Blei & Lafferty, 2006) models. The PAM model was also further enhanced by Mimno, Li and McCallum (2007) into the Hierarchical Pachinko Allocation Model (hPAM) which allowed explicit representation of a topic hierarchy.

Since all the extensions of the LDA model were built based on the premises of certain structure of textual data used by the researches, it is fair to note that literature-based (as in: based on the methodological overview above) model selection may be most suitable for research when the researcher clearly understands or can make valid assumptions about the structure of the data they are working with, for example, if there are topics that co-occur more frequently than the others, or if there may be a complex

hierarchical structure of the thematical variance in the dataset. Moreover, all of the models that demonstrate some advantages opposed to the others were tested on different types of data which differ from the textual data used in this thesis – for example, tweets or synthetic datasets. Therefore, there is no way of clearly knowing a priori, or based on given methodological studies, how the models will behave on a certain dataset, or, in the case of this thesis, the dataset containing texts of the articles of Russian mass media about sexual violence against women. In order to avoid the potential overlooking of the suitable model due to an only-literature-based model selection, I tested all of the aforementioned topic models using different hyperparameter settings on my data. The main aim there was to select the best model by evaluating its results through both quantitative and qualitative methods and therefore obtain the optimal quality of final topics structure and topics interpretability.

Therefore, the models fitted on the data in order to obtain the best results in terms of quality and interpretability were: the Latent Dirichlet Allocation (LDA) model, the Hierarchical Latent Dirichlet Allocation (hLDA) model, the Correlated Topic Model (CTM), the Pachinko Allocation Model (PAM) and the Hierarchical Pachinko Allocation Model (hPAM). The evaluation of the models was carried out both in quantitative (using the model Coherence score) and qualitative (manually assessing the interpretability and sensibility of the resulting topics) strategies.

Sentiment analysis of the article sentences

One of the research objectives of this thesis is to analyze sensationalism in the articles and its relation to the topics that are present in the discussion on sexual violence in the Russian news media. One of the ways of sensationalizing the stories on sexual crimes is the use of emotional rhetoric in the articles (Boranijašević, 2018).

The emotional rhetoric is one of the study subjects in the text mining and text analysis research field, where the measurements of the emotions are carried out by the sentiment analysis. Sentiment analysis is a commonly used method for the computational treatment of opinions, sentiments and subjectivity of the text which is

aimed to investigate attitudes and emotions towards some entity which represents individuals, events or topics (Medhat, Hassan & Korashy, 2014). The sentiment analysis field is constantly developing and includes many methods for emotion and opinion detection, such as dictionary-based, rule-based or modeled through machine learning algorithms and neural networks (Wankhade, Rao, Kulkarni, 2022), with latter proved to be the most successful, in terms of model quality, in the field (Taboada, 2016). Some of the common machine learning approaches to sentiment analysis, which is essentially a task for text classification into different emotions, are Maximum Entropy Classifier, Naïve Bayes Method and Support Vector Machine (Devika, Sunitha, & Ganesh, 2016), and with the usage of neural networks for the textual embeddings, especially the FastText model that is based on the n-grams, the sentiment analysis method shows the best results (Khomsah, Ramadhani, & Wijaya, 2022).

Therefore, in the analysis of the texts, for the sentiment measurements I will use the FastText model that is trained on the Russian-language data.

Data collection and preparation for analysis

In order to obtain the texts of Russian news media articles to use in the analysis, the automatic data collection was performed with several preparatory steps. First, the search of the relevant articles was conducted through the Medialogia resource in order to retrieve relevant article URLs for the further automatic collection from the websites. Then, after several methods for cleaning the list of obtained article URLs in order to ensure the article relevance, the final list of URLs was used to parse Russian mass media websites that contained relevant articles. In this subchapter, I dive into the methods of automatic data collection implemented in this thesis, with a step-by-step description of the database building process.

Timeline of the empirical study object

The empirical study object of this research is the texts of Russian news media articles mentioning cases of sexual violence against women or contributing to the discussion on sexual violence against women in other ways. Since the discussion on

sexual violence doesn't have the a priori starting or ending time, the time frame of the empirical object was chosen manually. The sample of the research, therefore, included texts of Russian news media articles published between January 1, 2016 and December 31, 2023. At that time a similar discussion was unfolding – about the legal regulation of domestic violence – starting from the year 2016 when the bill on the prevention of domestic violence was first introduced to the State Duma; approximately simultaneously with this discussion, a discussion was also unfolding about sexual violence against women as a form of gender-based violence closely related to domestic violence. Moreover, the timeline of the texts studied in this thesis is broad enough to include other features of the Russian context, such as wartime, which is usually characterized by elevated levels of sexual violence against women, where the sexual violence may be understood as a weapon of war (Skjelsbaek, 2001).

Retrieving relevant articles URLs using Medialogia

The initial set of article URLs was retrieved on 11th of April, 2024 from Medialogia, a media monitoring system created for the purposes of mass media and social media analysis. Medialogia stores a high scope of content from various media resources, especially in Russian language, and therefore was considered a primary source to collect data on the discussion of sexual violence against women present in Russian news media.

In order to reach a high variety of articles and to minimize the risk of overlooking a potentially relevant data, the query used to retrieve articles was quite broad and was targeted to all mentions of sexual violence within both article titles and article bodies. After the initial iterative process of looking for relevant keywords for the query in order to include as many articles as possible in the initial URLs dataset the following set of keywords was adopted: *((изнасил* /3 женщ*) | (секс* /3 насил*) | ("мужчина изнасиловал") | ("секусальное насилие") | ("сексуализированное насилие")).* Therefore, all the articles that mention "women" in the context of "rape" (there, the "*/3*" is used to limit said context to three words) as well those that mention "sexual violence"

in any word forms and wordings were included in the initial dataset. Also, two Medialogia built-in filters for media search were adopted in order to ensure high relevancy of articles along with their wide variety – "Mass media source level" and "Top-100 mass media sources". In terms of source level, I executed keyword search only among federal-level mass media sources. Also, I didn't include the ones that were not categorized the top 100 mass media sources in the initial URLs dataset. The purpose of such initial filtering of articles is, on the one hand, to initially clean the database from the most irrelevant publications, and, on the other hand, to use only the largest media sources that are presumed to be the most influential ones in terms of building the discourse and driving the discussion on sexual violence. Nevertheless, the initial set of articles retrieved from Medialogia included many irrelevant texts, where sexual violence wasnt one of the main topics of the article, and also many cases of sexual violence against children and minors, and therefore required major data cleaning. However, despite the fact that there were many irrelevant articles in the initial URLs dataset, the initially broad query used to collect URLs allowed me to look into the various article titles and ways that media phrase sexual violence cases to perform a further, more precise filtering by headers to clean data in a local format (see the section about the second stage of data cleaning below).

The whole study period includes articles from the 1st of January, 2016 to 31st of December, 2023, summing up into the 8 years of discourse on sexual violence in Russia mass media. The process of Medialogia data collection included searching and downloading the relevant article URLs for every month of a whole study period to take into account the download restrictions of the university account I used to work with Medialogia (the maximum size of a dataset containing article URLs available to download at once was 1000 units). Dividing the study period into separate months made it possible to download all the necessary URLs, since in almost all months the number of messages did not exceed 1000 units. However, within some month periods there were more than a thousand messages, and therefore the time period for one download of the URLs was reduced to 7-15 days. The total number of files downloaded from the

Medialogia for the whole period of study is 129 which is 14.33 files for every year in average; these files were then concatenated in order to create a full dataset of URLs for a future automatic retrieval of article texts.

The initial dataset included 87183 article URLs and then, after removing the duplicates, was reduced to a whole of 76222 of unique article URLs. Also, the initial dataset contained 7 variables – *id* (article id, unique only within one retrieved file from Medialogia), *header* (the title of the article), *date* (publication date), *newspaper* (mass media source name), *city* (mass media source location), *md_index* (Medialogia index for the message visibility[5]) and *url* (article URL). In order to prepare the initial URLs dataset for the automated data collection, I implemented two stages of data cleaning – first, cleaning the dataset from all the irrelevant units in terms of both articles and newspapers, and second, taking a slice of data using a filter of keywords dedicated to sexual violence against women.

URLs dataset preparation for automatic data collection

The first stage of URLs dataset cleaning was dedicated to deleting all the substantively irrelevant units of information and the units of information not available for automatic retrieval, first through article headers and then through newspapers.

The most important step there was to exclude all cases of sexual violence against children, minors, men and animals from the URLs dataset, since, on the one hand, the initial URL retrieval didn't account for different types of sexual violence victims, and, on the other hand, the further articles selection through the keywords would also skip the victim differentiation (see section below). In order to clean the URLs dataset from irrelevant cases I created a list of stopwords which contained words that would indicate that the article is about child abuse, not violence against women. Stopwords were collected through an iterative process of randomly selecting and reading publications

[5] Message visibility, according to Medialogia website, is a "cumulative parameter takes into account the "advertising equivalent" of a publication depending on the page number, the volume of the message, as well as circulation and attendance". See Medialogia website for the technologies' description: https://www.mlg.ru/about/technologies/

from the dataset until a repeatability of possible words that were indicators of child abuse was achieved. The final stopwords list for this step of data cleaning is following:

['10-лет', '10-лет', '11-лет', '11-лет', '12-лет', '12-лет', '13-лет', '13-лет', '14-лет', '14-лет', '15-лет', '15-лет', '16-лет', '16-лет', '17-лет', '17-лет', '2-лет', '2-лет', '3-лет', '3-лет', '4-лет', '4-лет', '5-лет', '5-лет', '6-лет', '6-лет', '7-лет', '7-лет', '8-лет', '8-лет', '9-лет', '9-лет', 'внуч', 'Внуч', 'воспитанни', 'Воспитанни', 'восьмилет', 'Восьмилет', 'геи', 'Геи', 'гей', 'Гей', 'гея', 'Гея', 'двухлет', 'Двухлет', 'девоч', 'Девоч', 'девятилет', 'Девятилет', 'десятилет', 'Десятилет', 'детдом', 'Детдом', 'детей', 'Детей', 'дети', 'Дети', 'детоубий', 'Детоубий', 'детск', 'Детск', 'детьми', 'Детьми', 'дочер', 'Дочер', 'дочк', 'Дочк', 'дочь', 'Дочь', 'животн', 'Животн', 'зоофил', 'Зоофил', 'извращ', 'Извращ', 'изнасиловала', 'Изнасиловала', 'изнасилованного', 'Изнасилованного', 'изнасилованном', 'Изнасилованном', 'изнасилованный', 'Изнасилованный', 'интернат', 'Интернат', 'классник', 'Классник', 'классниц', 'Классниц', 'маленькую', 'Маленькую', 'малолетн', 'Малолетн', 'малыш', 'Малыш', 'мальчи', 'Мальчи', 'младен', 'Младен', 'напал', 'Напал', 'несовершеннолетн', 'Несовершеннолетн', 'новорожд', 'Новорожд', 'одноклассниц', 'Одноклассниц', 'падчериц', 'Падчериц', 'педафил', 'Педафил', 'педофил', 'Педофил', 'племянни', 'Племянни', 'подвергшегося', 'Подвергшегося', 'подвергшемуся', 'Подвергшемуся', 'подвергшийся', 'Подвергшийся', 'подростк', 'Подростк', 'подростки', 'Подростки', 'подросток', 'Подросток', 'пятилет', 'Пятилет', 'развращ', 'Развращ', 'растлен', 'Растлен', 'растли', 'Растли', 'ребен', 'Ребен', 'ребён', 'Ребён', 'сверстни', 'Сверстни', 'семилет', 'Семилет', 'сирот', 'Сирот', 'соврати', 'Соврати', 'совращен', 'Совращен', 'сына', 'Сына', 'трехлет', 'Трехлет', 'трёхлет', 'Трёхлет', 'ученик', 'Ученик', 'учениц', 'Учениц', 'учитель', 'Учитель', 'четырехлет', 'Четырехлет', 'четырёхлет', 'Четырёхлет', 'шестилет', 'Шестилет', 'школ', 'Школ', 'школьни', 'Школьни', 'юнош', 'Юнош']

The number of articles dropped through this filter was 30750, leaving 45472 units in the URLs dataset.

The second step of data cleaning within this stage was checking the relevance and reachability of mass media sources. First, all units where the article URL was not available were deleted. Next, manually I checked each mass media source for relevance and accessibility. Some of the mass media sources included in the dataset did not fit the acceptable type of media outlet – these were websites of TV channels and radio stations, while for my thesis only textual news media, magazines and news agencies were relevant. I also removed media from the dataset whose sites had been deleted and were

no longer available for collecting information. In total, out of the original 131 media, there were 62 irrelevant ones. After removing all the irrelevant mass media sources, 31817 publications remained in the dataset.

The second stage of preparing data for automatic text collection was to identify relevant publications using keywords dedicated to cases of sexualized violence against women, as well as discussions about sexualized violence in general. The generation of a list of keywords was accomplished by selecting and reading random titles until repeatability of language describing cases or discussions about sexualized violence against women was achieved. The final keywords list for this step of data selection is following:

['harassment', 'Harassment', 'metoo', 'Metoo', 'MeToo', 'Metoo', 'домогавшегося', 'Домогавшегося', 'домогавшийся', 'Домогавшийся', 'домогался', 'Домогался', 'домогательств', 'Домогательств', 'домогаться', 'Домогаться', 'женщин', 'Женщин', 'жертв', 'Жертв', 'защит', 'Защит', 'защищ', 'Защищ', 'избиение', 'Избиение', 'избил', 'Избил', 'избитая', 'Избитая', 'изнасилов', 'Изнасилов', 'маньяк', 'Маньяк', 'на женщ', 'На женщ', 'На женщ', 'На женщ', 'надругал', 'Надругал', 'надругательст', 'Надругательст', 'нападен', 'Нападен', 'насилова', 'Насилова', 'насилуют', 'Насилуют', 'насильник', 'Насильник', 'насильственные сексуальные действия', 'Насильственные сексуальные действия', 'обвинение', 'Обвинение', 'обвиняемый', 'Обвиняемый', 'похититель', 'Похититель', 'похищение', 'Похищение', 'самооборон', 'Самооборон', 'секс-преступлен', 'Секс-преступлен', 'секс-скандал', 'Секс-скандал', 'скандал', 'Скандал', 'убийств', 'Убийств', 'убийц', 'Убийц', 'убил', 'Убил', 'харассмент', 'Харассмент', 'я тоже', 'Я тоже', 'Я тоже', 'Я тоже']
[['полиц', 'секс'], ['Полиц', 'секс'], ['полиц', 'Секс'], ['Полиц', 'Секс'], ['прав', 'женщин'], ['Прав', 'женщин'], ['прав', 'Женщин'], ['Прав', 'Женщин'], ['секс', 'нападен'], ['Секс', 'нападен'], ['секс', 'Нападен'], ['Секс', 'Нападен'], ['секс', 'насил'], ['Секс', 'насил'], ['секс', 'Насил'], ['Секс', 'Насил'], ['секс', 'преступлен'], ['Секс', 'преступлен'], ['секс', 'Преступлен'], ['Секс', 'Преступлен'], ['секс', 'преступник'], ['Секс', 'преступник'], ['секс', 'Преступник'], ['Секс', 'Преступник'], ['секс', 'рабство'], ['Секс', 'рабство'], ['секс', 'Рабство'], ['Секс', 'Рабство']]

After this step of articles selection, 17110 articles from the total of 68 mass media sources remained in the URLs dataset. The final URLs dataset contained the same 7 variables as the initial one– *id* (article id, unique only within one retrieved file from Medialogia), *header* (the title of the article), *date* (publication date), *newspaper* (mass

media source name), *city* (mass media source location), *md_index* (Medialogia index for the message visibility) and *url* (article URL). The last variable, *url*, was then used as a list to collect full article texts for the further analysis of the discussion on sexual violence present in Russian mass media articles.

Automatic collection of mass media texts from newspaper websites

Data collection was carried out through automatic collection, or parsing, of mass media article texts published on the corresponding mass media websites using a list of URLs obtained at the previous stage of data collection. Article texts, as most of the information on the Web, are stored in the HTML documents and therefore is considered unstructured data. In order to retrieve article texts stored in an HTML format on the media websites using URLs leading to the relevant webpages, I used the Requests python library which main aim is retrieving information from Internet search queries automatically from the local computer coding environment (Jupyter Notebook in my case). To retrieve the HTML code in a readable format from the source code of the webpage retrieved through the Requests library, I used the BeautifulSoup python library which allows to dive into the structure of the HTML document and search for the relevant tags and classes in order to retrieve relevant pieces of information. Requests and BeautifulSoup are one of the most common libraries used in web scraping though they have significant limitations since they work only on static pages (Chaulagain et al., 2017), however, for my case of data collection there is no need to use libraries that work with dynamic requests and imitate the browser (such as the Selenium python library) since all the textual data needed is stored on the single pages available through URLs that were previously collected through Medialogia.

In my URLs dataset in its final and ready-for-parsing version I had a total of 17 110 links leading to unique articles from 65 mass media sources. It is important to note that the way different media websites store data highly differ from each other – each mass media source has its own style of structuring the HTML code within the webpages, such as different class names and properties for storing pieces of displayable

data. Therefore, it was impossible to create a universal coding sequence to retrieve texts from the HTML documents among the various mass media outlets, and for each of the 65 mass media sources in my URLs dataset, the process of automatic textual data collection was carried out separately.

The entire process of automatic data collection consisted of a combination of two functions wrote in python coding language. The first function is a scraping function, created individually for each of the 65 mass media sources and their unique ways of storing textual data in the HTML format on their webpages. The function consisted of making a request to a webpage to obtain its source code in the HTML format through the Requests library, and also extracting the necessary text components from this source code using the BeautifulSoup library. The components of the article scraped through this function were article headline, article description and article body, stored then in three separate variables. The second function within the process of data collection was the parsing function. This function was universal for all mass media sources and was aimed to looping through all the URLs for a given mass media source, applying a scraping function to each one of them, and saving the collected three string variables into a dataframe for each mass media source, also saving the relevant URLs for each article into the same dataframe. The individual dataframes were also saved into the files in the .xlsx format locally and then were concatenated into a general dataframe containing parsed texts, also then saved to the .xlsx format for the local storage. The example of a scraping function, as well as the parsing function, may be found in the Appendix 2 (Fig. 1 and Fig. 2 respectively).

It is important to note that though most mass media websites allowed scraping the page source code freely, some mass media websites (for example, Известия (iz.ru) or Фонтанка (fontanka.ru)) had restrictions that prevented the webpage text from being retrieved through the Requests library which led to a code 403 error while trying to request a page. However, this problem was easily resolved by using API keys additionally to making a request. The API keys were taken from the third party resource – ZenRows, created to scrape large amounts of data without getting blocked. This part

of data collection was implemented in a separate Jupyter Notebook file; however, the whole process was, except for the API key usage, identical to the process of data collection described above.

The final dataframe, obtained as a result of automatic collection of texts from media publications, consisted of N rows and 4 columns*: title, description, article body,* and also the *url* column containing the link to the publication for subsequent comparison and concatenation of this dataframe with a dataframe containing links to publications and additional variables (such as time of publication and MD index) obtained at the previous stage of data collection. The total number of texts retrieved through automatic data collection was 15 342. The number of publications which texts were collected during the parsing process is less than the original number of relevant links because not all web pages allowed data to be collected from them in the usual way. The most common errors that made it impossible to collect the text were: "*HTTPSConnectionPool(host=www.perm.kp.ru, port=443): Read timed out. (read timeout=20)"*, meaning that the website page took too much time to load, and "*KeyError"*, meaning that the scraping function cannot access the dictionary key within the collected data. This means that this web page is not suitable for scraping and uses a different method for storing text data. Such publications were ignored as exceptions and were not additionally collected after the main data collection process.

Database building

The main database building was based on the concatenation of the dataset of article headlines, descriptions and texts that were automatically collected by article URLs, and the initial URLs dataset which, along with the article headlines and article URLs, contained variables-characteristics of the articles: date of publication, newspaper name, city where the newspaper is located and Medialogia article visibility index. The two datasets were concatenated by the unique article IDs which were the common variable for both of them.

The main dataset for the further analysis, thus, is built along the separate and unique articles published by the Russian media outlets on the discussion of sexual violence which are the main units of analysis in my thesis. As shown in Table 1 that displays a randomly selected row in the main dataset, there are several variables for each unit of analysis: *id* (the unique identification number of an article in the dataset), *headline* (article headline), *description* (an extended version of the article headline or qualifier that is used before the body of the article), *article_body* (the article text itself), *date* (the date that the article was published), *newspaper* (the name of the newspaper where the article was published), *city* (the city where the newspaper is located), *md_index* (Medialogia visibility index for the articles) and *url* (the URL that leads to the article). The dataset then was further supplemented by several variables indicating the topic probabilities extracted from the Pachinko allocation topic model in the process of analyzing the thematic structure of the articles, as well as by the variables indicating the level of several types of sentiment extracted from the FastText social network model used in the process of sentiment analysis to determine the emotional scope used in the coverage of sexual violence. Both of the analyses and correspondent variable building processes are described in the next subchapter of this thesis.

Table 1. Article example from the SV dataset, randomly selected

id	headline	description	article_body	date	newspaper	city	md_index	url
4326	Насиловал женщин: в Самаре полиция разыскивает опасного маньяка	Подозреваемый нападал на женщин в Самаре, Санкт-Петербурге и Волгограде	В Самарской области разыскивают опасного преступника. Мужчина нападал на женщин и насиловал их, угрожая убить. Маньяк действовал в трёх городах России — Самаре, Волгограде и Санкт-Петербурге.— В Санкт-Петербурге против неизвестного возбуждено уголовное дело по факту изнасилования и иных насильственных действий сексуального характера, сопряженных с угрозой убийства, — прокомментировали в ГУ МВД России по Самарской области. Преступления были совершены ещё в 2016 году, однако поймать преступника пока не получается. Полиция просит самарцев помочь с поисками подозреваемого.Приметы: мужчина, на вид 30-40 лет, славянской внешности, рост 165-175 см., среднего телосложения, волосы короткие черного цвета с лобной залысиной, стрижка короткая, голова овальной формы, на лице имеются следы оспиной сыпи, уши оттопырены. В разговоре с прохожим говорил, что он из Р. Башкортостан. Если вы знаете этого человека или видели, где он находится, сообщите информацию в ближайший отдел полиции или по телефонам: (846) 3737406, 89370627007 или 020 (102 с мобильного федеральных операторов сотовой связи).	2018-08-17 00:00:00	Комсомольская правда (kp.ru)	Москва	1,812	https://www.samara.kp.ru/online/news/3207490/

Analyses and procedures

One of the objectives of this thesis is to identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media. To achieve this task, I use the LDA-based topic modeling, for which I implement the model fitting and selection process that is described in this subchapter. Moreover, I aim to explore the relationship between the themes present in the articles on sexual violence against women and the tone that is used by Russian news media to cover such violence. For this task, I use sentiment analysis in order to classify texts into different emotional groups and to evaluate each article in terms of presence of different emotions in it, which is also described in this subchapter.

Fitting the LDA-based topic models for analysis

Due to their probabilistic nature, topic models deal with a set of separate words on the input and therefore the original texts that are to be modeled have to be preprocessed. Typically, the text preprocessing process includes the following steps: 1) splitting texts into lists, 2) punctuation and special symbols removal, 3) transforming words to a general, universal form by trimming some word parts (stemming) or bringing them to their original lexical form (lemmatization), 4) removing stopwords – words that are often found in any text, but do not carry any semantic load (Hickman et al., 2022).

For the topic modeling in my thesis, I implemented all of the common steps for text preprocessing, taking into account the features of Russian language. First, I removed all punctuation symbols using the list of punctuation symbols created by me as an extension of the ready-made list from the NLTK python library since this list didn't catch all the special symbols my data contained. Then I split the article texts into lists of words and transformed the words to their primary lexical forms using the MorphAnalyzer model from the Pymorphy python library created for working with natural language. In essence, MorphAnalyzer works as a lemmatizer that uses vocabulary and morphological analysis of words and transforms them into their primary, or dictionary, lexical forms. I used lemmatization instead of stemming, which simply

removes derivational suffixes and inflections from words, to take advantage of its ability to detect different word forms as well as synonyms, which is crucial when working with a language as rich morphologically as the Russian one. Moreover, lemmatization tends to show the best results compared to other forms of words preprocessing for document retrieval (Balakrishnan, 2014). The next step after words lemmatization was cleaning word lists from the stopwords – words that often occur in texts without adding any substantive meaning to them, for example, adpositions or conjunctions. The stopwords list for words removal was also constructed by me as an extension of the stopwords list from the NLTK python library since the ready-made list was short and didn't contain all the necessary words and word forms due to the multilingual nature of the library.[6] The preprocessed article texts then were added to my SV dataset as a separate variable called texts_lem – this variable was then used to form a collection of documents for topic modeling.

According to Blei (2012), given a collection of texts at input, a typical topic model "algorithmically finds a way of representing documents that is useful for navigating and understanding the collection" (p. 8) by returning a set of topics, or recurring themes, that are present in the collection. However, as Blei (2012) argues, the model itself doesn't provide the researcher with the results that themselves best fit the data and thematic field of study – it is the researcher's task to select the configurations of the model to obtain a set of valid and well-interpreted topics. In order to select a model which would be more suitable for my dataset and would produce more interpretable and sensible results, I tested five LDA-based probabilistic topic models – the Latent Dirichlet Allocation (LDA) model, the Hierarchical Latent Dirichlet Allocation (hLDA) model, the Correlated Topic Model (CTM), the Pachinko Allocation Model (PAM) and the Hierarchical Pachinko Allocation Model (hPAM) – on the texts of Russian mass media articles about sexual violence against women, with various model hyperparameters (see Table 2 below) set manually in order to determine which model

[6] Both custom punctuation and stopwords lists, as well as the Jupyter Notebook file implemeting text preprocessing of the article texts, may be found in my thesis GitHub repository: https://github.com/kmlapshina/bachelor-thesis

would show the best result in terms of thematic structure of the documents as well as topic interpretability.

Table 2. Hyperparameters used in model fitting for the LDA-based topic models

Model	Depth	N of topics	Alpha	Subalpha	Eta	Gamma
LDAModel		5, 10, 15, 20, 25, 30, 40, 50, 60, 70	0.001		0.00001, 0.001, 0.1, 0.5, 1	
HLDAModel	3		0.001		0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5, 0.7, 1	0.001
PAModel	3	5, 10, 15, 20, 25, 30, 40, 50, 60, 70	0.001	0.001	0.00001, 0.001, 0.1, 0.5, 1	
HPAModel	3	5, 10, 15, 20, 25, 30, 40, 50, 60, 70	0.001	0.001	0.00001, 0.001, 0.1, 0.5, 1	
CTModel		5, 10, 15, 20, 25, 30, 40, 50, 60, 70	0.001		0.00001, 0.001, 0.1, 0.5, 1	

Typically, a topic model requires a manual setting the k number of topics to model. The number of topics usually discovered by scholars in topic modeling implementations varies within very wide ranges – some research suggests small numbers, like 6-topic solutions, to best fit their data, while some suggests several hundred topics to be the best k (for examples, see DiMaggio et al., 2013; Günther, E., & Domahidi, 2017; Saldaña Villa, 2017; Shahin, 2016). As Saldaña Villa (2017) states, for large corpora the best topic solution is application specific and the optimal k number of topics should not rely only on the theoretical or previous research guidelines and is best to be determined by several iterations of model fitting. Therefore, the main varying hyperparameter I used in fitting the LDA-based topic models for finding the best model was the number of topics (except for the hLDA model which determines the number of topics automatically by a hierarchical cluster analysis). In my model fitting, the k number of topics model hyperparameter ranged from 5 to 70 with a step of 5, then 10 units. The whole range of the k number of topics hyperparameter was 5, 10, 15, 20, 25, 30, 40, 50, 60 and 70 topics except for the hLDA topic model. The other varying hyperparameter used in model fitting was the eta hyperparameter – a smoothing parameter for word weights within topics. For all models except hLDA, the eta

parameter range was 0.00001, 0.001, 0.1, 0.5 and 1 for model fitting. For the hLDA topic model, I used the same range of eta hyperparameter except for the shorter step since eta was the only varying hyperparameter for this model – 0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.5, 0.7 and 1 for model fitting. Other hyperparameters for model fitting were static. The alpha hyperparameter, which, as well as eta, is a smoothing parameter, but for the topic weights for documents instead of words weights for topics, was set to 0.001 for all models (and also the subalpha = alpha = 0.001 for the Pachinko allocation models). I chose the low value of the alpha parameter specifically to achieve the more visible presence of top topics within a document for a future analysis since low alpha gives the most weight of the topic distribution within a document to a single topic.[7] For hierarchical models, I used the depth = 3 hyperparameter for the number of levels in topic hierarchy, and a gamma = 0.001 hyperparameter for the hLDA model, following Kolstov and colleagues' parameter setting in their implementation of the model (Koltsov et al., 2021). All the hyperparameters used in fitting the LDA-based topic models are systematized in Table 2 presented above.

Table 3. Number of models fitted for model selection, by model type

Model type	**Number of models fitted**
LDAModel	50
HLDAModel	10
PAModel	50
HPAModel	50
CTModel	50
Total count	**210**

To select the best model, I fitted five types of LDA-based models based on 10 various k number of topics and 5 eta values, or just 10 eta values for the hLDA model, with other hyperparameters as static numbers. The total number of models fitted on the text documents containing articles on sexual violence against women was 210 (see

[7] For a better understanding of the hyperparameters alpha and eta, see the python implementation of topic modeling with different model configurations here: https://ethen8181.github.io/machine-learning/clustering/topic_model/LDA.html

Table 3 above): for the Latent Dirichlet Allocation (LDA) model, the Correlated Topic Model (CTM), the Pachinko Allocation Model (PAM) and the Hierarchical Pachinko Allocation Model (hPAM) it was 50 fits for each of the model types (according to the combinations of the k number of topics and eta hyperparameter values) and the Hierarchical Latent Dirichlet Allocation (hLDA) model it was 10 fits (according to the eta hyperparameter values).

For the implementations of the models, I used the Tomotopy python library created specifically for topic modeling. It contains the implementations for most of the common topic models, including the LDA-based topic models that I chose to fit and evaluate on my data, and is similar to one of the most common Python libraries for topic modeling, Gensim, though Tomotopy has several performance advantages such as in running time or stability.[8] Moreover, the Tomotopy library contains implementations for all the methods which are necessary for the evaluation of the models. For example, it allows to access dictionaries used in model fitting, to get word distributions for modeled topic structures, to get topic distributions for the documents passed to the models, and also some common performance evaluation metrics, such as log-likelihood, perplexity score and coherence score.

The process of model fitting consisted of looping through all models and hyperparameters, fitting each of the 210 models on a set of parameters, evaluating each of the models and writing the results into a csv file, with 100 iterations for each model fitting. The whole fitting process was executed in 28 hours 52 minutes.

Topic model selection and building the topic structure

In order to retrieve topic from the article texts, I fitted five LDA-based topic models – the Latent Dirichlet Allocation (LDA) model, the Hierarchical Latent Dirichlet Allocation (hLDA) model, the Correlated Topic Model (CTM), the Pachinko Allocation Model (PAM) and the Hierarchical Pachinko Allocation Model (hPAM) – with differing

[8] For performance comparisons between Tomotopy and Gensim see Tomotopy library documentation: https://bab2min.github.io/tomotopy/v0.12.6/en/#performance-of-tomotopy

k number of topics and eta hyperparameters, resulting in 210 different model configurations. To get interpretable and valuable insights from the article texts in a form of topics, I implemented a model selection process based on the Coherence score metric and further assessing the interpretability of topics modeled by the models with best coherence.

Topic coherence score is essentially a metric which measures the degree of semantic similarity between high scoring words in the topic which allows to judge the results of a topic models in terms of them being humanly interpretable (Stevens et al., 2012). Therefore, a good coherence score (which is the closer to zero the better) essentially means that the model produced topics which are, on the one hand, substantively differentiated from each other and, on the other hand, are humanly interpretable in opposition to being only statistically produced by the algorithm. This evaluation metric is commonly used along with the model perplexity metric which evaluates the predictive ability of the model, as in "how efficiently a model can handle new data it has never seen before" (Hasan et al., 2021, p.3). However, the predictive ability of the models in my analysis is not as important as the ability of the model to model the meaningful thematic diversity of texts, so the models were evaluated only on the basis of coherence. In the process of model selection, I used the coherence score based on the UMass method which, as for the present day, is that approximates human ratings of the model results better than other methods (Röder, Both, & Hinneburg, 2015). The implementation of the coherence score was taken, as the model implementations themselves, from the Tomotopy library. The coherence score was calculated based on top 50 words in each topic to get the evaluation based on the words that characterize each topic more than other words assigned to it.

Table 4. Top 5 models fitted on SV data docs by coherence score

№	Model	Depth	N of topics	Alpha	Subalpha	Eta	Coherence	Log-likelihood
86	PAModel	3	30	0.001	0.001	0.1	-1.2457179301339154	-10.646885743101738
93	PAModel	3	15	0.001	0.001	0.5	-1.2534100402663058	-10.287369551568238
102	PAModel	3	10	0.001	0.001	1	-1.2718038364140059	-9.83193860466609
138	HPAModel	3	50	0.001	0.001	0.001	-1.2953333188568883	-9.440748825795408
135	HPAModel	3	25	0.001	0.001	0.1	-1.3072488732739878	-9.504060927078278

The best results in terms of coherence score were showed by the Pachinko allocation models, both flat and hierarchical types of the model. As showed in Table 4 (and for the full display of model fitting results see Table A2 in Appendix 3), the best model in terms of coherence score (coherence = -1.246) is the PA model for k=30 number of topics and eta hyperparameter set to 0.1. The two following best models (coherence scores -1.253 and -1.272 respectively) are flat PAM models for 15 and 10 topics with eta hyperparameter set to 0.5 and 1 respectively. The next two best models are hierarchical PA models for 50 and 25 topics with eta hyperparameter set to 0.001 and 0.1 which showed coherence scores of -1.295 and -1.307, respectively.

Table 5. Top 5 models fitted on SV data docs by coherence score, best model within each group

№	Model	Depth	N of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
42	LDAModel		10	0.001		1		-1.490926863053502	-8.573678756177065
86	PAModel	3	30	0.001	0.001	0.1		-1.2457179301339154	-10.646885743101738
60	HLDAModel	3	70	0.001		1	0.001	-1.3551410899687797	-8.416453023365115
138	HPAModel	3	50	0.001	0.001	0.001		-1.2953333188568883	-9.440748825795408
191	CTModel		5	0.001		0.5		-1.4562519565059382	-8.04699631497335

Due to the disproportionality of the distribution of models by coherence scores, I decided to select a model from the best models in each group of LDA-based topic models according to the human interpretability of modeled topics. Table 5 shows the top 5 models adjusted for model type – the Latent Dirichlet Allocation (LDA) model,

the Hierarchical Latent Dirichlet Allocation (hLDA) model, the Correlated Topic Model (CTM), the Pachinko Allocation Model (PAM) and the Hierarchical Pachinko Allocation Model (hPAM), – where in each group, the model with the best coherence score was taken for the evaluation of topic human interpretability. As seen in Table 5, the “classical” LDA model showed a coherence score of -.491 with a number of topics set to 10 and the eta hyperparameter set to 1. The best hierarchical version of the LDA model showed a slightly better coherence score of -1.355 with eta parameter set 1 and with the 4 optimal number of topics. Finally, the best result of a CTM model is coherence score of -1.456 with 5 topics and eta parameter set to 0.5.

Table 6. Topic words for the top-topic and super-topics from the HPA model topic word distributions

ID	Top 50 topic words	Unique words within the top 50 topic words
0	мужчина женщина изнасиловать задержать девушка изнасилование полиция дом подозревать дело преступление произойти возбудить район уголовный житель насильник квартира жертва сообщать сообщить ранее год москва местный улица совершить место напасть злоумышленник знакомый время находиться орган область потерпевший стать преступник молодой правоохранительный избить пострадать город действие рассказать сексуальный жительница сотрудник обратиться надругаться	напасть злоумышленник знакомый избить надругаться
1	мочь человек говорить ребёнок очень год жизнь знать сказать история хотеть просто большой мужчина делать женщина почему никто рассказать семья время хороший самый жить рассказывать наш мать думать случай жертва видеть сделать считать слово родитель нужно стать понимать идти насилие поэтому должный хотя ситуация работать сторона происходить девочка понять бояться	очень просто делать почему никто жить рассказывать думать видеть родитель нужно понимать идти поэтому хотя понять бояться
2	стать время день год несколько тело первый выйти однако жена оказаться пока найти решить друг последний жертва дом месяц убийца рука убить поздний спустя смерть видео пытаться место новый имя вернуться слово написать появиться фото начать затем момент сын брат остаться лишь конец сразу взять человек работать вместе сделать александр	тело пока рука поздний спустя смерть пытаться вернуться затем лишь сразу взять
3	человек мочь число право должный случай проблема иметь работа вопрос новый ситуация являться преступление принять считать большой общественный представитель социальный закон результат место часть согласно наш среди группа решение мера ответственность мнение страна количество лицо защита безопасность центр связь отметить сеть глава однако информация внимание цель система последний любой член	иметь социальный согласно количество внимание цель система

For all the models from the list of the best models displayed in Table 5, I built the distribution of the top words in the topics, and for ease of evaluation and interpretation, I identified unique words within each topic (based on the top 50 words in the topic) which make the topic stand out from the rest. Table 6 shows the top 50 words, as well as the unique words for the first few topics of the HPAM model, to illustrate what the data looked like when human assessment of model topic coherence was performed.

Human evaluation of the top five models showed that the best result was achieved by the HPAM model with k=50 number of topics and eta hyperparameter set to 0.001. The remaining models were not suitable for being used in further analysis, both in terms of the number and content of topics, and in the degree of their interpretability, since the models showed too general, identical or too many purely "machine" results that were difficult to interpret and occurred only due to what appeared to be a random co-presence of words in text documents. However, the HPAM model still did not model the topic structure perfectly and also required a detailed human evaluation of each topic to remove topics from the data that were not suitable for further analysis.

The human evaluation process was implemented by assessing the top documents in terms of topic probability. Typically, for each of the 50 topics modeled by the HPAM model, 20 articles with the highest probability of containing the corresponding topics were taken for the human assessment and topic interpretation. There, after carefully assessing the topic structure modeled by the HPAM model (k = 50, eta = 0.001), I removed all the «nonsensical» topics from the topic structure. The final topic structure consisted of 37 topics in total, including one top topic (the one that connects all the articles together and is modeled by the HPAM model as the first-level "cluster"), three super topics (the second-level "cluster" in the HPAM model) and 33 general topics, and was used in further analysis. The final topic structure is displayed in Table A5 (Appendix 5), and the interpretation of the topics is provided in Chapter III of this thesis.

Building sentiment variables for the articles

One of the research objectives of this thesis was to explore the thematic structure of the discussion in its relation to the sensationalist rhetoric. In order to evaluate sensationalism, understood as the emotional tone used in the coverage of sexual violence, in the articles, I used sentiment analysis – the method in the text mining field that is aimed to investigate attitudes and emotions towards some entity which represents individuals, events or topics (Medhat, Hassan & Korashy, 2014). Sentiment analysis, therefore, allows a researcher to extract a scope of different emotions that are expressed in a text, as well as quantitative indicators reflecting the degree of presence of each emotion in the text.

For building the variables indicating a sentiment in a text, I used the Dostoevsky python library created for sentiment analysis of the Russian-language texts. The library contains the FastText-based social network model trained on RuSentiment dataset,[9] which I used in my analysis to create sentiment variables. The implementation of sentiment analysis in the Dostoevsky library allows to classify the texts into five categories of sentiment – "positive", "negative", "neutral", "speech" and "skip", as well as assign all these labels to a text along with the degree of confidence of the model in the presence of this type of sentiment in the text, expressed quantitatively. It is important to note here that sentiment analysis methods work best with short and thematically and / or emotionally narrow texts, and therefore when passing large texts as articles to the model, it tends to classify them as just emotionally neutral, ignoring the emotional coloring of individual sentences in the text. Therefore, I conducted sentence-level sentiment analysis with a further building of article-based sentiment variables from the distributions of sentiment in the sentences of the article.

There were several steps in the analysis. First, I split the article texts into sentences using PunktSentenceTokenizer from the NLTK library,[10] which is a training-based model that, at the application stage, automatically recognizes sentences in the text

[9] For implementation details, see the Dostoevsky library documentation: https://github.com/bureaucratic-labs/dostoevsky
[10] For details of the method, see the documentation: https://www.nltk.org/api/nltk.tokenize.PunktSentenceTokenizer.html

and carries out appropriate tokenization of the text. Then, for the implementation of the model, I tokenized the sentences with the RegexTokenizer and then passed them to the FastText-based social network model from the Dostoevsky library. The model output was essentially a sentence-level dictionary with five types of sentiment and the level of the presence of the corresponding types of sentiment in the sentences. These sentiment variables, indicators of their presence in sentences, as well as original sentences and article IDs were collected into a separate dataset during the analysis process for further calculations of article-level sentiment variables. The result of the sentence-level sentiment analysis for one of the articles sampled randomly (with a 425 article id) is presented in Table 7.

As shown in Table 7, the sentence-level sentiment analysis resulted in obtaining five interval sentiment variables – “positive”, “negative”, “neutral”, “speech” and “skip” – for each sentence in each article. In order to compute the article-level sentiment variables, several approximation methods were implemented. The first one was to simply get the mean value of each of the sentence-level variables, which for the example in Table 7 would be 0.029 for positive sentiment, 0.248 for negative sentiment and 0.681 for neutral sentiment. The second approximation method was based on the 75 percentile value of sentence-level sentiment values since the mean value would reduce too much the significance of sentences where sentiment is expressed strongly, if this sentiment is not expressed throughout the rest of the text. For the variables in the example displayed in Table 7, the values of this article-level sentiment variables would be would be 0.039 for positive sentiment, 0.245 for negative sentiment and 0.895 for neutral sentiment. Finally, the maximum values of sentence-level sentiment variables were taken to take into account the level of emotions to which the text of the publication as a whole reaches – for example, if the text contains an important evaluative exclamation, which is erased if we take only the average values of the emotionality of sentences. For the variables in the example displayed in Table 7, the values of such article-level sentiment variables would be would be 0.08 for positive sentiment, 0.725 for negative sentiment and 0.953 for neutral sentiment.

Table 7. Sentence sentiment values for one of the articles

	article_id	sentence	positive	negative	neutral	speech	skip
1	425	Фото с сайта Kaktus.media Прокуратура Жайлыкского района Чуйской области Киргизии возбудила уголовное дело в отношении сотрудников районного отдела внутренних дел после того, как в здании ОВД была убита девушка.	0,013	0,197	0,803	0,041	0,095
2	425	Об этом сообщает Turmush.Убийство произошло 27 мая.	0,006	0,554	0,943	0,000	0,057
3	425	Перед этим в милицию обратился Турдаалы Кожоналиев, который сообщил, что его дочь Бурулай похитили и собираются выдать замуж против ее воли.	0,068	0,321	0,415	0,003	0,245
4	425	Во все райотделы Бишкека и Чуйской области были отправлены ориентировки, и через несколько часов в селе Сосновка Чуйской области сотрудники ДПС задержали машину, в которой находились Бурулай и двое мужчин.	0,020	0,156	0,570	0,023	0,169
5	425	Их доставили в ОВД Жайлыкского района.Пока сотрудники ОВД оформляли документы, один из задержанных, 29-летний Б.М., прошел в комнату приема граждан, где находилась Бурулай, и запер дверь изнутри.	0,027	0,192	0,743	0,003	0,156
6	425	Услышав звуки борьбы, сотрудники милиции взломали дверь и обнаружили, что похититель нанес ножевые ранения девушке и себе.	0,053	0,363	0,301	0,004	0,207
7	425	Их доставили в больницу, где Бурулай, не приходя в сознание, скончалась.	0,007	0,725	0,766	0,001	0,101
8	425	Напавший был прооперирован, сейчас он находится в больнице.	0,009	0,725	0,257	0,001	0,245
9	425	Второй пассажир был передан в следственную службу ОВД Жайлыкского района.Дело против сотрудников милиции возбуждено по статье «Халатность».	0,042	0,148	0,699	0,002	0,148
10	425	Им грозит до пяти лет тюрьмы.	0,080	0,239	0,024	0,000	0,644
11	425	В прокуратуре сообщили, что подозреваемые правоохранители допустили целый ряд нарушений.	0,052	0,129	0,896	0,001	0,053
12	425	«Во-первых, сотрудники, которые остановили автомобиль, должны были составить рапорт.	0,035	0,192	0,831	0,001	0,041
13	425	Во-вторых, осмотреть задержанных на наличие холодного или огнестрельного оружия.	0,022	0,123	0,926	0,001	0,017
14	425	В-третьих, – на наличие телесных повреждений, наркотического или алкогольного опьянения», – рассказали изданию в прокуратуре.Представитель ведомства также отметил, что, когда задержанных доставили в Жайлыкский ОВД, милиция и там никого не осмотрела, поэтому мужчина и смог пронести нож.	0,024	0,207	0,644	0,025	0,095
15	425	«Кроме того, девушку и задержанных двух мужчин оставили одних в кабинете, а этого не должно быть.	0,016	0,197	0,813	0,001	0,228
16	425	В результате убийство произошло под носом у сотрудников милиции», – сообщил он.	0,009	0,165	0,910	0,000	0,072
17	425	Также возбуждено уголовное дело по статье «убийство».По оценке правозащитников, в Кыргызстане с целью принуждения к браку ежегодно похищают от 8 до 12 тысяч женщин.	0,021	0,066	0,766	0,014	0,129
18	425	Из этих похищений две трети несогласованные.	0,007	0,207	0,500	0,001	0,370
19	425	Около двух тысяч похищенных женщин подвергаются изнасилованию.	0,019	0,129	0,672	0,002	0,321
20	425	В 2012 году был принят закон, устанавливающий уголовную ответственность за похищение невест, однако он пока не смог остановить сложившуюся практику.	0,080	0,187	0,446	0,017	0,104
21	425	Кроме того, правозащитники указывают, что в милиции к подобным инцидентам иногда относятся без должного внимания и не возбуждают уголовные дела.	0,020	0,251	0,902	0,009	0,028

Therefore, for each type of sentiment (I do not take “speech” and “skip” types into the analysis due to their procedural nature which doesn’t deal with the emotion in texts) three article-based variables were computed, with a total of nine variables – *positive_mean, negative_mean, neutral_mean, positive_75, negative_75, neutral_75, positive_max, negative_max* and *neutral_max* – used in further analysis of sensationalism in articles.

Methods, concepts and variables: summary

The main units of analysis in this the unique articles published by various Russian media outlets on the cases of sexual violence and on the discussion on sexual violence in general through the years 2016-2023. For article texts as the unit of analysis in a database the variables-characteristics of the articles are added into the dataset: date of publication, newspaper name, city where the newspaper is located, Medialogia article visibility index, 37 variables for topic probabilities, prevalence of positive, negative and neutral sentiment and year or publication computed from the date variable.

To summarize the methods used to achieve the research objectives of this thesis as well as the variables used in the analysis, see Table 8.

Table 8. Summary of the research methods

Research objective	Hypotheses & assumptions	Methods for analysis	Variables
RO1. To identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media.	A1. There will be themes within the thematic structure of the discussion on sexual violence against women that indicate the absence of agency of the victims of sexual violence. A2. Within the thematic structure of the discussion on sexual violence against women, newspapers will tend to attribute responsibility for the violence to individuals rather than the government or society. A3. There will be a sensationalist rhetoric present in the Russian news media articles that cover sexual violence	Topic modeling with the Pachinko allocation topic model	*headline, description, article_body, date*, 37 topic probability variables
RO2. To explore the relationship between the themes present in the articles on sexual violence against women and the tone that is used by Russian news media to cover such violence	H1. The themes that explore the cases of sexual violence will be correlated to either positive or negative sentiment in the coverage of sexual violence. H2. The themes that explore sexual violence as a social problem will be correlated to the neutral sentiment in the coverage of sexual violence. H3. There will be no correlation between the other themes and level of sentiment in the articles covering sexual violence against women.	Pearson's correlation coefficient (with preliminary sentiment analysis for emotion detection)	*positive_mean, negative_mean, neutral_mean, positive_75, negative_75, neutral_75, positive_max, negative_max neutral_max,* 37 topic probability variables
RO3. To explore the temporal differences in the thematic coverage of sexual violence against women in Russian news media	H4. The general themes within the discussion on sexual violence in Russian news media will tend to be fading over time in terms of their presence in the articles.	One-way ANOVA	*year,* 10 topic probability variables

CHAPTER III.
HOW RUSSIAN MASS MEDIA FRAME SEXUAL VIOLENCE AGAINST WOMEN

This study is aimed to identify the ways through which the cases of sexual violence against women and of the discussion on sexual violence against women in general are covered by the Russian news media, as well as the interpretive frameworks, or frames, through which such coverage is carried out. The empirical research object, therefore, is the texts of Russian news media articles that describe cases of sexual violence against women or contribute to the discussion on sexual violence against women in other ways. In order to analyze the ways of coverage of sexual violence in Russian mass media, I aimed to detect and analyze a thematic structure of said discussion where the discovered topics or topic groups, along with the emotions following these topics in coverage, could perform the framing functions which would indicate different framing strategies in the coverage of sexual violence. To discover a thematic structure of the discussion, I implemented the hierarchical Pachinko allocation topic model with a total of 37 topics and then described and interpreted the topics, and further connected them to sensationalism in the articles as well as tested the anticipation of the temporal change inherent to the common frames used in media coverage of social problems.

This chapter starts with the sample overview. Then follows the part on the thematic structure of the discussion on sexual violence against women in Russian news media, then followed by the analysis of the article sentiments related to the topics and the analysis on the temporal changes in the topic usage by the media.

The coverage of sexual violence: articles overview

The final sample of the study consists of 15 342 articles from the Russian news media covering the discussion on sexual violence through the years 2016 – 2023. Thus, the whole study period is constituted by the eight years of the coverage of sexual violence by various media outlets in a Russian language, located in Russia. Within the distribution of the number of articles by the year of publication, generally speaking,

there is a good fullness of each class (where each year of coverage is taken as a class), though the number of publications is not the same for each year. As shown in Table 9, the coverage was the highest in the year 2018 with a total of 2312 articles, followed by the year 2019 with a total of 2229 articles. The lowest coverage in terms of the article count is seen in the year 2020 where the total number of articles was only 1562, supposedly to the shift of the general public and media discourse to health themes during the Covid-19 pandemic.

Table 9. Number of articles by the year of publication

Year	Article count
2016	1610
2017	1475
2018	2312
2019	2229
2020	1562
2021	1966
2022	2118
2023	2070
Sum	**15342**

The daily coverage of sexual violence is seemingly evenly distributed through the whole study period as shown in Figure 5, which demonstrates the total number of articles on sexual violence published by various Russian news outlets everyday throughout the years 2016 – 2023.

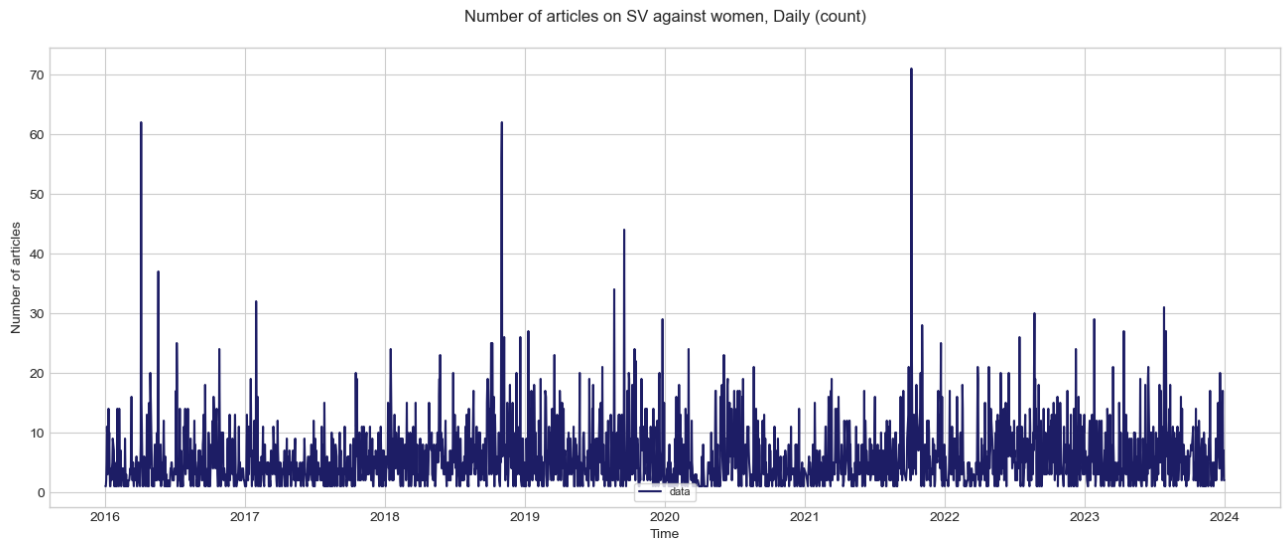

Figure 5. The dynamic of the coverage of SV in Russian mass media, daily count

There are several noticeable outliers in the daily distribution of publications at the beginning of the year 2016, at the end of the year 2018 and at the end of the year 2021. In this regard, I assumed that outliers may be associated with high-profile cases of sexual violence during these periods of peak publication. However, a careful examination of publications on each day where an outlier was observed did not show any substantive consistency in the themes covered by Russian news media on those days. Therefore, I consider outliers in the number of publications on some days of the period to be random.

Figure 6 shows a similar distribution of the number of publications by Russian news media within the framework of the discussion on sexual violence, where the number of publications is summarized by month. There is a noticeable dynamic in the coverage of sexual violence which extends the observations that are made on the basis of the year-by-year coverage shown in Table 9. It can be seen that, within the 2018-2019 time period, which itself is characterized by a great degree of number of articles on sexual violence, the most coverage of the problem happened in the last months of the year 2018. The coverage of sexual violence in Russian media news then declines in 2020 and for about the whole next year, but sin the year 2021 the new wave of attention to the issue may be observed, with the overall coverage peaking at the end of 2021.

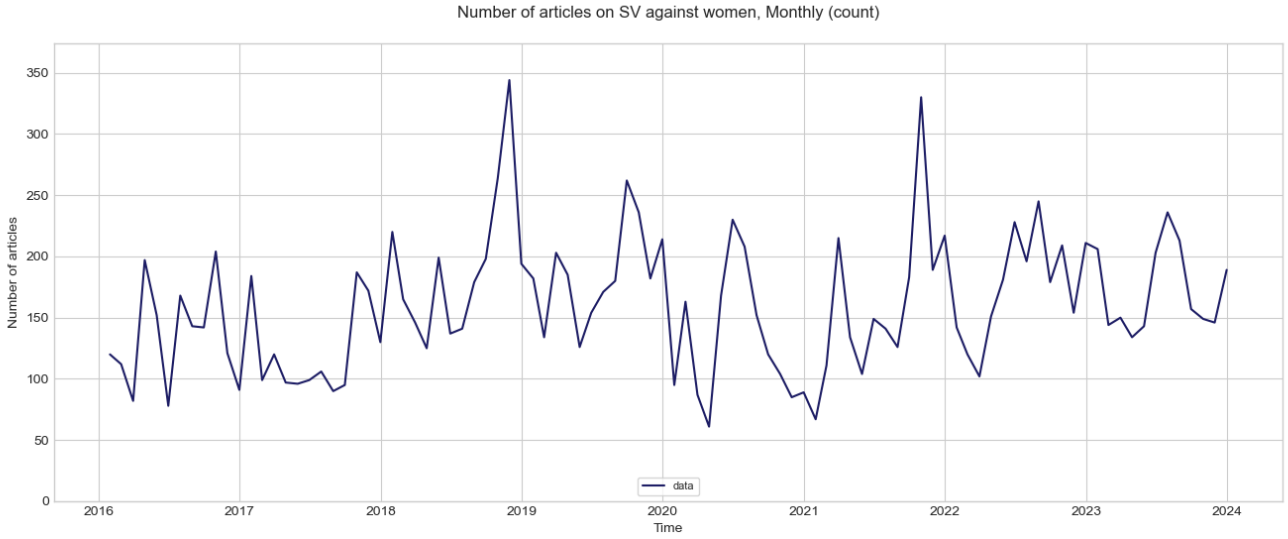

Figure 6. The dynamic of the coverage of SV in Russian mass media, monthly count

There is very high variability of the media outlets that published the articles on in the study sample, both in terms of number of the outlets and the number of articles published by each outlet. The total number of the newspapers in the dataset is 65 (for the full list of papers see Table A1 in Appendix 1), with the top 50% of the media outlets accounting for about 90% of the total number of articles covering sexual violence. As shown in Table 10, the most prevalent newspaper in the corpus is "Комсомольская правда" with a total 1465 number or articles on sexual violence collected through the data collection process. The other top mass media outlets by the number of articles analyzed in this thesis are large newspapers such as "Газета.Ru" and "Известия" (with a total of 1085 and 420 articles respectively) and information agencies such as "РИА Новости" (with a total of 619 articles).

Table 10. Top mass media sources by article count

Newspaper name	**Article count**
Комсомольская правда (kp.ru)	1465
Газета.Ru	1085
Lenta.Ru	1074
ИА Regnum	864
Московский Комсомолец (mk.ru)	708
Life.ru	691
ТАСС	645
РИА Новости	619
ИА Росбалт	511
VSE42 (vse42.ru)	461
Известия (iz.ru)	420
ИА Красная весна	407
NEWS.ru	384
Российская газета (rg.ru)	338
Sports.ru	333
Фонтанка (fontanka.ru)	332

The news media outlets also show some spatial variability. As shown in Table 11, most newspapers that covered sexual violence through years 2016-2023 are located in Moscow with a total number of 12 920 (84.2%) articles in a corpus. The other large locations in terms of the total coverage are Saint-Petersburg (total of 1139 articles),

Kemerovo (total of 461 articles) and Ekaterinburg (total of 461 articles). The total number of locations for the newspapers in the corpus is 11.

Table 11. The number of articles by the city of the newspaper, count

City	Article count
Москва	12920
Санкт-Петербург	1139
Кемерово	461
Екатеринбург	361
Казань	128
Волгоград	90
Пермь	78
Тюмень	61
Химки	42
Красногорск	36
Нижний Новгород	26

The dataset analyzed in this thesis also contains a variable displaying the level of visibility for each article taken from Medialogia, which essentially is an advertising equivalent of the article, combining such parameters as the number of page where the article is published, circulation and attendance for the published article. This index is an interval variable meaning that divisions in a dimension can be interpreted and compared quantitatively; though this is an index variable meaning that variable values can only be interpreted relatively to each other and do not appellate to the count of real entities. Moving to the values of the visibility index, as shown in Table 12, the mean value of this variable is 1.234 with a 95% confidence level that the real mean value lies within the [1.223; 1.245] interval. The median value is 1.148 which is slightly lower than the mean value, indicating the uneven distribution in the way of the higher values of the variable. All the values of the visibility index are distributed from the value 0, which is the lowest value, to the value 3.996, which is the highest value. The variance and standard deviation for this variable are 0.48 and 0.692, respectively.

Table 12. Descriptive statistics for the Medialogia visibility index variable

Statistic	Value
Mean	1,23380
95% Confidence Interval for Mean	[1,22284; 1,24476]
5% Trimmed Mean	1,19003
Median	1,14800
Variance	0,480
Std. Deviation	0,692805
Minimum	0,000
Maximum	3,996
Range	3,996
Interquartile Range	0,964
Skewness	0,866
Kurtosis	0,934

The statement on the uneven nature of the variable distribution mentioned above was checked with a normality test. The Kolmogorov-Smirnov test, displayed in Table 13, shows that the empirical distribution of the visibility index is not normal: the null statistical hypothesis of no difference between the empirical and theoretical distribution is rejected in favor of the alternative at the 99% confidence level (p-value < 0.01).

Table 13. Results of the normality test for the topic probability variables

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
md_index	0,058	15342	0,000

a. Lilliefors Significance Correction

The articles in the corpus can also be described by the level of different types of sentiment prevalent in the coverage of sexual violence. The construction of the variables indicating positive, negative and neutral sentiment in articles was described in the previous chapter. Essentially, these variables show the level of positive, negative and neutral emotions used in articles that describe cases of sexual violence or the discussion on sexual violence itself and are considered quantitative, or, more precisely, interval. The values within the sentiment variables indicate the cumulative (computed by mean, 75 percentile or maximum values) probabilities of the corresponding sentiment types in the articles covering the discussion on sexual violence and are derived from the sentence

sentiment variables, since the sentiment analysis itself was conducted on the articles split into sentences instead of being conducted on the whole article texts themselves. For example, the variable *positive_mean* indicates the mean value for the probabilities of the positive sentiment counted on the sentence basis.

Table 14. Descriptive statistics for the sentiment variables

	N	Minimum	Maximum	Mean	Std. Deviation
positive_mean	15342	0,002	0,263	0,059	0,026
positive_75	15342	0,002	0,345	0,074	0,034
positive_max	15342	0,002	1,000	0,183	0,150
negative_mean	15342	0,028	0,453	0,173	0,051
negative_75	15342	0,011	0,618	0,221	0,070
negative_max	15342	0,041	0,997	0,398	0,178
neutral_mean	15342	0,241	0,966	0,684	0,102
neutral_75	15342	0,255	1,000	0,820	0,106
neutral_max	15342	0,269	1,000	0,928	0,093
Valid N (listwise)	15342				

Table 14 displays the key descriptive statistics for all the sentiment variables used in the further analysis. As shown in Table 15, all the variables are not normally distributed: according to the Kolmogorov-Smirnov test, the empirical distributions of the sentiment variables are not normal: the null statistical hypothesis of no difference between the empirical and theoretical distributions is rejected in favor of the alternative at the 99% confidence level (p-value < 0.01) for all the sentiment variables.

Table 15. Results of the normality test for sentiment variables

	Kolmogorov-Smirnov[a]		
	Statistic	**df**	**Sig.**
positive_mean	0,08	15342	0
positive_75	0,085	15342	0
positive_max	0,175	15342	0
negative_mean	0,032	15342	0
negative_75	0,048	15342	0
negative_max	0,084	15342	0
neutral_mean	0,019	15342	0
neutral_75	0,068	15342	0
neutral_max	0,22	15342	0

a. Lilliefors Significance Correction

Finally, the dataset contains 37 interval variables indicating the probabilities of encountering different topics in the articles, with the topics being constructed in the process of topic modeling fitting, choosing the best model and then evaluating and interpretating its results (for the description of all the corresponding procedures see the previous chapter). Therefore, the articles are characterized by the level of certain topics prevalence in them in terms of probabilities. The descriptive statistics for the topic probability variables for all the topics used in further analysis are displayed in Table A3, Appendix 4. The normality test was also conducted for these variables and resulted in rejecting the null statistical hypothesis of no difference between the empirical and theoretical distributions in favor of the alternative at the 99% confidence level (p-value < 0.01) for all the variables (for the results of the normality test, see Table A4 in Appendix 4).

Dominant themes in the discussion on sexual violence in Russian mass media

One of the research objectives of this thesis was to identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media. In this subchapter, I address this task by describing the thematic structure obtained through the textual analysis of the corpus, assessing the prevalence and visibility rates of different topics within the discussion, then moving to the description and interpretation of the topics obtained through text analysis.

Thematic structure of the discussion on sexual violence

The thematic structure of the discussion on sexual violence in Russian news media was obtained and analyzed through topic modeling as the method most suitable for retrieving topics from the texts, based on the probabilistic method of clustering the word co-occurrences. In order to retrieve topics from the article texts, I fitted five LDA-based topic models with differing k number of topics and eta hyperparameters, resulting in 210 different model configurations. To get interpretable and valuable insights from the article texts in a form of topics, I implemented a model selection process based on

the Coherence score metric and further assessing the interpretability of topics modeled by the models with best coherence. Human evaluation of the top five models showed that the best result was achieved by the hierarchical Pachinko allocation topic model with k=50 number of topics and eta hyperparameter set to 0.001. The further model evaluation, targeted at cleaning the obtained thematic structure of the corpus from all the irrelevant topics, was implemented by assessing the top 20 documents in terms of topic probability for each of the topics from the initial model. There, after carefully assessing the topic structure modeled by the HPAM model, I obtained the final thematic structure of the corpus consisting of 37 topics in total. For each of the topics, the model output contained topic word distributions (simply, from what words the topic is constructed), along to which I computed a set of unique words based on the top 50 words for each of the topics to ease the process of topic interpretations in terms of assessing those words that make a certain topic stand out from the rest.

The topics obtained during topic modeling cannot be called equivalent both in terms of prevalence levels in the texts and in terms of content characteristics. Within the thematic structure of the corpus, at the stage of interpretation of topics, I carried out a classification, combining content-similar topics into separate groups. In total, during the classification process, I identified six groups of topics: “General” with a total of 6 topics, “Social problem” with a total of 4 topics, “Criminal cases & process” with a total of 5 topics, “Perpetrator types” with a total of 3 topics, “Spheres of violence” with a total of 6 topics and “Cases” with a total of 13 topics. The whole thematic structure of the corpus, classified into these groups, can be seen in Figure 7.

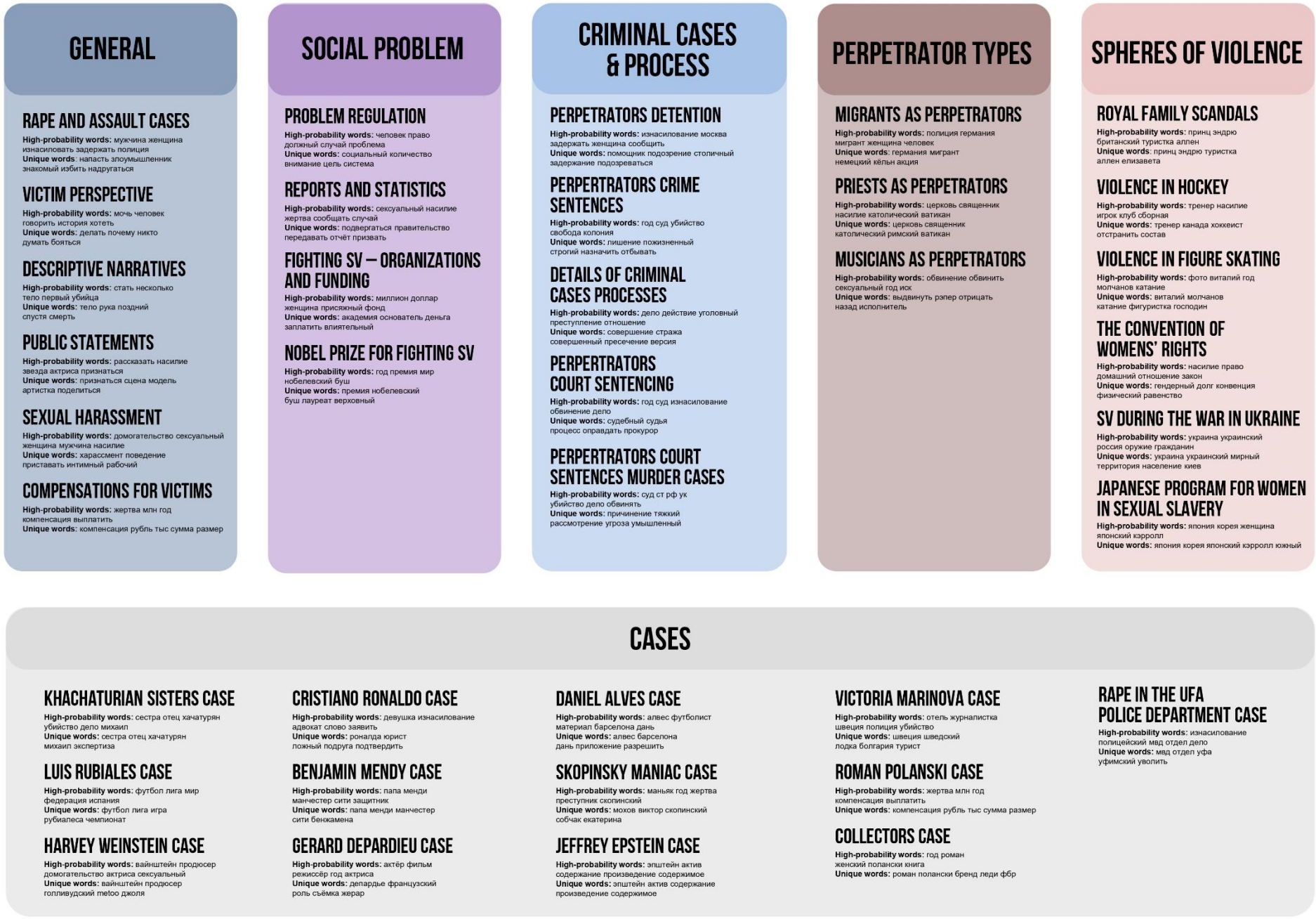

Figure 7. The thematic structure of the articles on SV in Russian mass

The “General” topic group contains topics which explore different general themes within the articles covering sexual violence in Russian news media, as well as ways of describing violence. The “Social problem” topic group includes the topics that essentially portray sexual violence in terms of a problem that holds societal causes as well as societal consequences. The “Criminal cases & process” topic group is based on the topics that deal with descriptions of criminal cases of sexual violence perpetrators, bringing the legal consequences discourse into the discussion. Topics in the “Perpetrator type” group are concentrated on a certain social or professional group of people being the perpetrators of sexual violence. The “Spheres of violence” topic group is constituted by the topics which were too narrow to be interpreted as general topics and yet too wide to be interpreted as cases since they contain many cases united by specific circumstances such certain place or professional sphere. Finally, the “Cases” topic group contains many topics that explore the coverage of specific cases that were popular in the media and discourse throughout the whole study period and mostly are associated with the names of a victim or a perpetrator of sexual violence. A detailed description of the topics included in these content groups is contained in further parts of this subchapter.

The structure of the discussion on sexual violence, consisting of 37 topics, is complex not only in terms of the substantive differences in the topics, but also in the degree of presence and visibility of each topic in this discussion. To assess this complexity and differences in the presence and visibility of topics, one can look at the prevalence of topics in the corpus, calculated in several ways, as well as the average visibility index values for topics in the articles where these topics were the most probable (for illustrations see Figures 8-11).

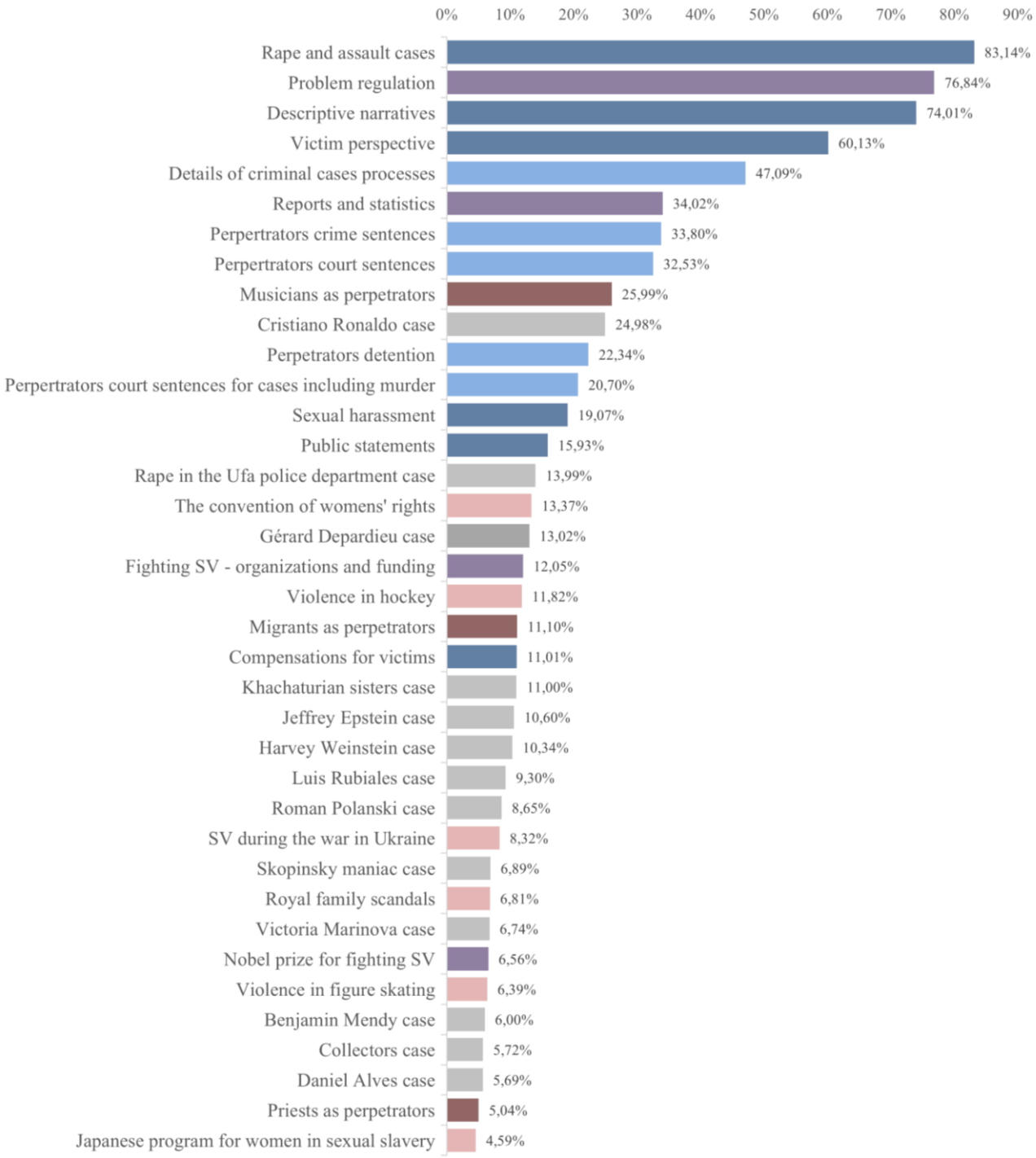

Figure 8. Total topic prevalence within the thematic structure of the discussion on sexual violence in Russian media

Figure 8 shows topic prevalence rates calculated as the share of articles where this topic can be seen with a probability greater than 0.1. Therefore, this is a general indicator of the occurrence of a topic in articles without taking into account how dominant this topic is within each article compared to the other topics that are encountered in the same article.

As shown in Figure 8, the most prevalent topics in the corpus according to this indicator are the topics from the "General" topic group, as well as two topics from the "Social problem" group and one topic indicating the criminal cases process. The most prevalent topics, in terms of being encountered in the article with the probability of 0.1 or higher, are "Rapes and assault cases" (occurring in 83.14% of the articles in the corpus), "Problem regulation" (occurring in 76.84% of the articles in the corpus), "Descriptive narratives" (occurring in 74.01% of the articles in the corpus), "Victim perspective" (occurring in 60.13% of the articles in the corpus), "Details of criminal cases process" (occurring in 47.09% of the articles in the corpus) and "Reports and statistics" (occurring in 34.02% of the articles in the corpus) topics. The most prevalent topics according to this indicator are topics of a fairly general nature that do not differ in specificity, and some can be interpreted as generic frames and can be encountered regardless of the general topic of the publication.

Topics indicating cases and spheres of violence are expectedly found in fewer publications due to their situational nature and specificity. The least prevalent topics in the corpus according to the corresponding indicator of prevalence indicator are "Japanese program for the women in sexual slavery" (occurring in 4.59% of the articles in the corpus), "Priests as perpetrators" (occurring in 5.04% of the articles in the corpus), "Daniel Alves case" (occurring in 5.69% of the articles in the corpus), "Collectors case" (occurring in 5.72% of the articles in the corpus) and "Benjamin Mendy case" (occurring in 6% of the articles in the corpus) topics.

However, it is also important to assess topic prevalence that takes into account the differences in topic probabilities in the articles, for which I explore maximum and mean values of topic probabilities within the corpus.

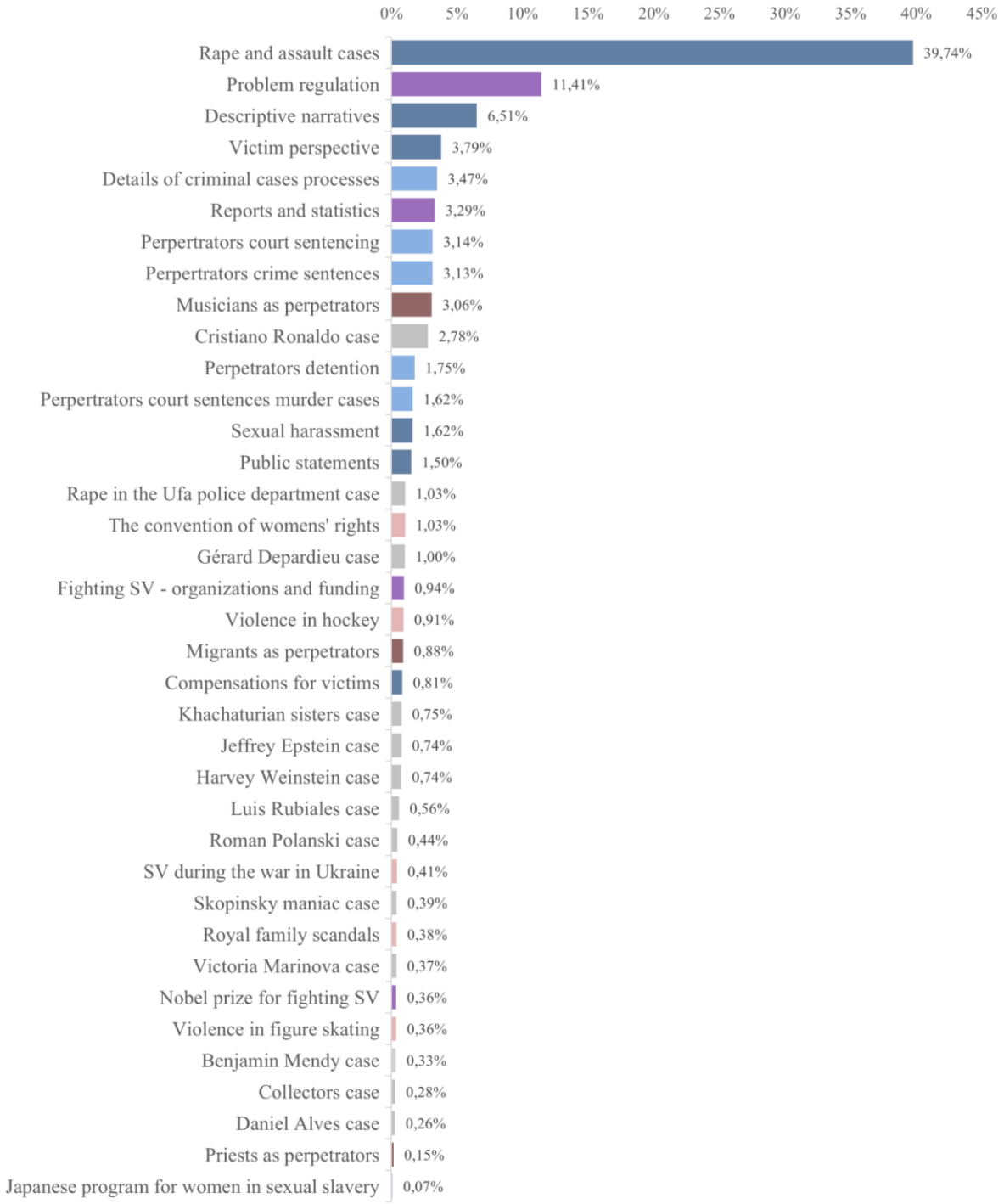

Figure 9. The most probable topics per article prevalence within the thematic structure of the discussion on sexual violence in Russian media

Figure 9 shows topic prevalence rates calculated as the share of the articles where the corresponding topic is dominant, as in having the highest probability to be encountered within the article in comparison to the other topics in the topic distribution for the articles.

As shown in Figure 9, the “Rape and assault cases” is the most prevalent topic according to this indicator, having a prevalence indicator that is much higher than these indicators for other topics. Thus, in 39.74% of articles, this topic is the most predominant in terms of the probability of being encountered in these articles. This can be interpreted in such a way that these publications are essentially the publications about cases of violence, embedded in the discussion about sexualized violence, that is, they directly describe cases of sexual violence towards women. The other dominant themes in the discussion, according to the shares of the articles where these themes are the most probable, are “Problem regulation” (is the most probable in 11.41% of the articles), “Descriptive narratives” (is the most probable in 6.51% of the articles), “Victim perspective” (is the most probable in 3.79% of the articles) and “Details of criminal cases process” (is the most probable in 3.47% of the articles) topics.

The least shares of the articles, as for the previous indicator, are had by the cases topics. These topics are the same: “Japanese program for the women in sexual slavery” (is the most probable in 0.07% of the articles), “Priests as perpetrators” (is the most probable in 0.15% of the articles), “Daniel Alves case” (is the most probable in 0.26% of the articles), “Collectors case” (is the most probable in 0.28% of the articles) and “Benjamin Mendy case” (is the most probable in 0.33% of the articles). Again, the low scores for case topics are due to their situational nature, in contrast to the general nature of the “big” topics, which are included in the discussion through general, applicable to many situations, ways of describing violence or particular approaches to the problem of violence.

In general, the results of this indicator turned out to be similar to the results of the previous indicator, since they consider the occurrence of topics in publications from the point of view of the share of publications.

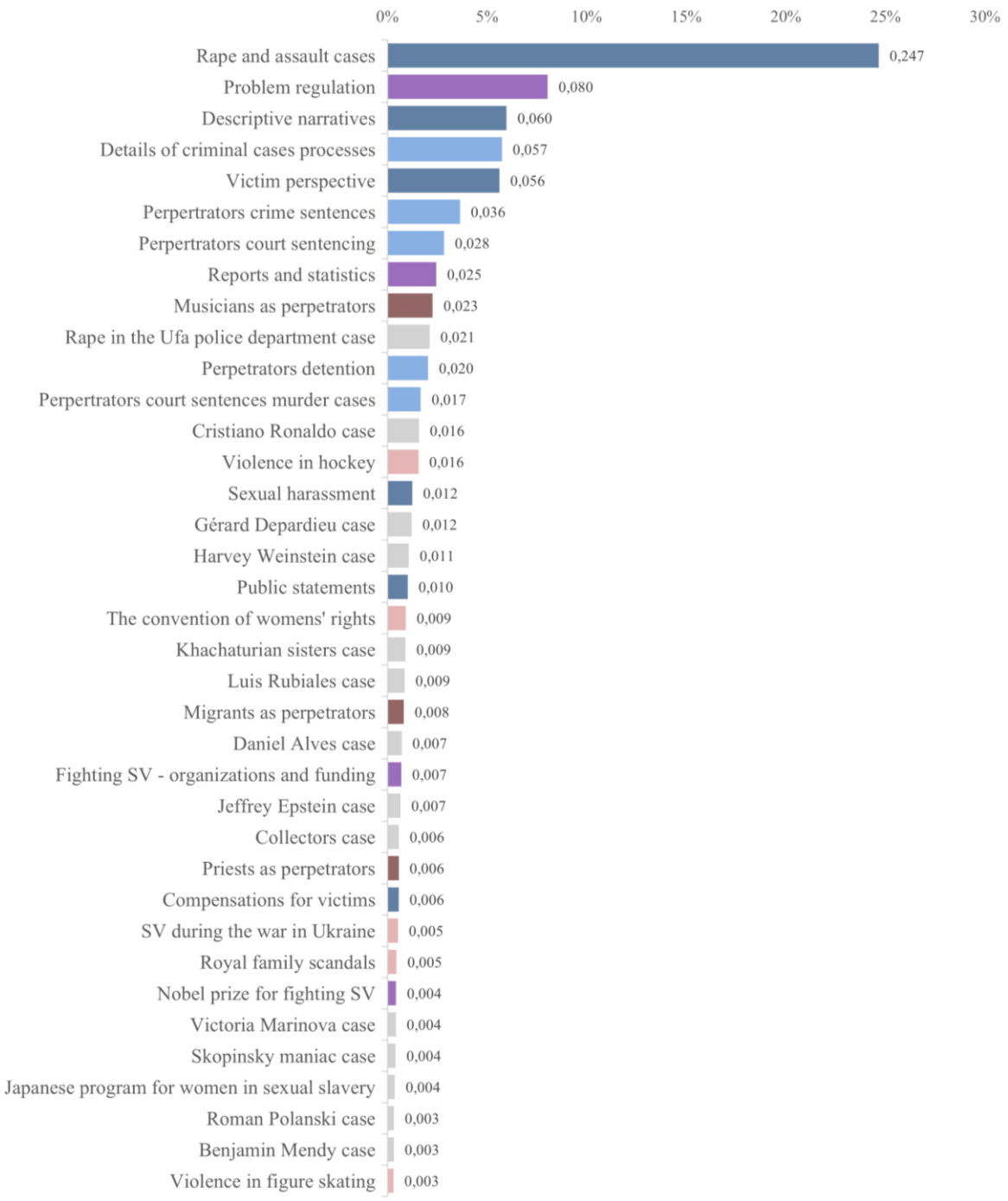

Figure 10. Mean probability per topic prevalence within the thematic structure of the discussion on sexual violence in Russian media

Figure 10, in contrast, shows the mean values of the topic probabilities across all articles in the corpus. This indicator of prevalence may be understood as the probability of encountering a certain topic by randomly choosing an article from a corpus. It is different from the previous prevalence rate indicators in terms of accounting for topic probability values across all the article texts.

Nevertheless, the distribution of the dominant topics is quite similar to the ones assessed above. The topics that have the highest probabilities within the corpus are "Rape and assault cases" (with 0.247 mean probability of topic), "Problem regulation" (with 0.08 mean probability of topic), "Descriptive narratives" (with 0.06 mean probability of topic), "Details of criminal cases process" (with 0.057 mean probability of topic) and "Victim perspective" (with 0.056 mean probability of topic). Interestingly though, the case topic regarding the rape in the Ufa police station has relatively high mean probability (0.021) which is possible to be interpreted as high topic consistency, meaning that if the topic is to be encountered in the article, the article is more likely on the corresponding case.

The least probable topics in terms of the mean probability through the corpus are mostly case topics. Among these are "Violence in figure skating" (with 0.003 mean probability of topic), "Benjamin Mendy case" (with 0.003 mean probability of topic), "Roman Polanski case" (with 0.003 mean probability of topic), "Japanese program for women in sexual slavery" (with 0.004 mean probability of topic) and "Skopinsky maniac case" (with 0.004 mean probability of topic). Again, the low scores for case topics are due to their situational nature, in contrast to the general nature of the "big" topics, which are included in the discussion through general, applicable to many situations, ways of describing violence or particular approaches to the problem of violence.

The indicators discussed above are based on the distributions of the topics in the corpus and therefore are directly dealing with the topic occurrence rates. The alternative method for determining which topics are the most prevalent in the discussion, I used another variable for measuring prevalence – Medialogia visibility index.

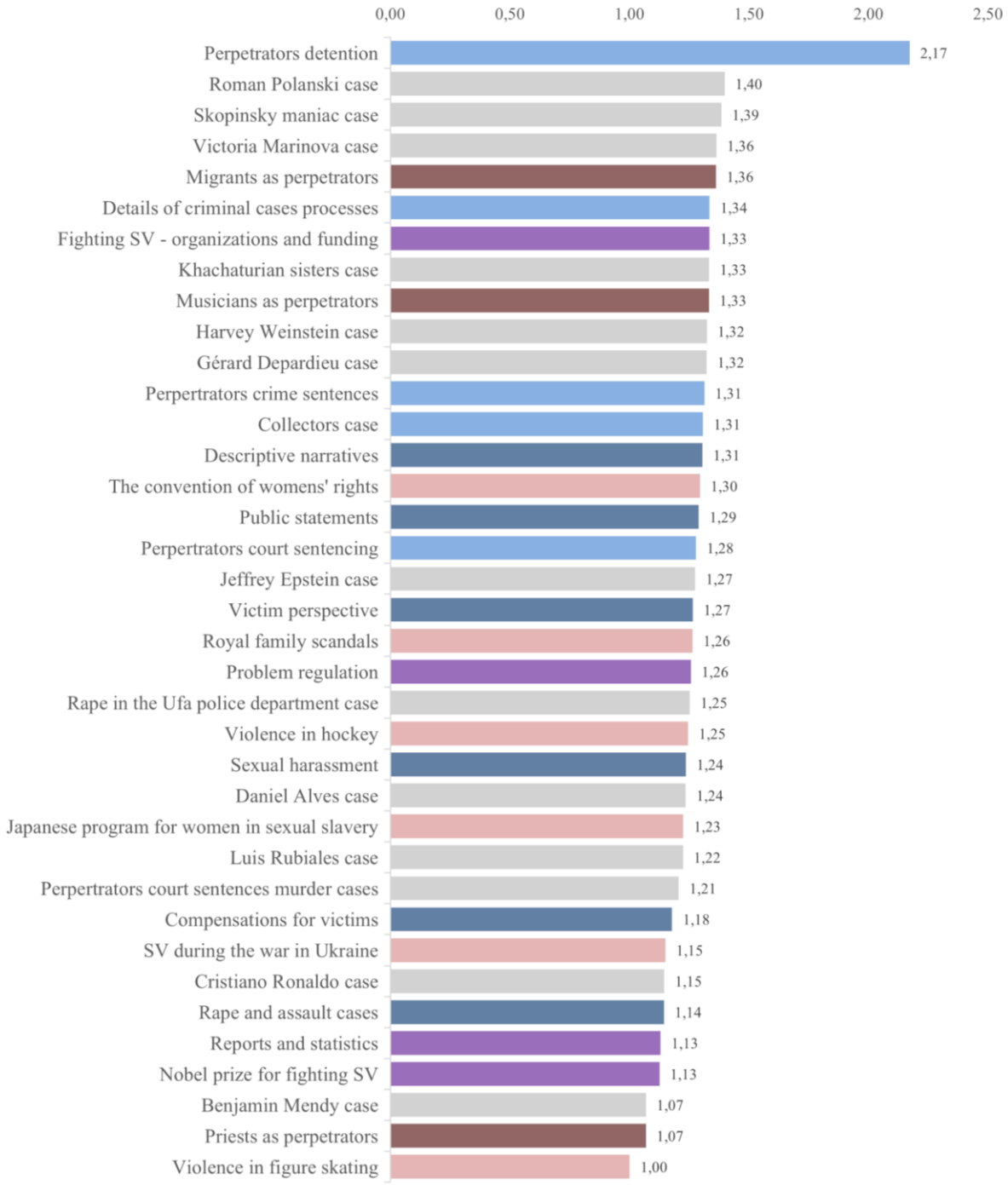

Figure 11. Mean values of the visibility index for the most probable topics in the articles

Figure 11 essentially shows how visible the topics in the corpus are in terms of turnover, views and other media visibility parameters associated with the topics. For assessing topic visibility, I determined the topics that are more probable for each of the articles and then computed mean article visibility scores for these topics. Therefore, this indicator of topic visibility shows, if an article is based on a certain topic, how much it would be interesting for the audience (since the visibility index is computed from the quantitative parameters that indicate audience's attention paid for the articles).

There, the distribution of the topics within this topic dominance scale differs from the previous methods for assessing it. The highest mean visibility index values tend to be in the articles where the most probable topics are "Perpetrators detention" (2.17 mean index value), "Roman Polanski case" (1.4 mean index value), "Skopinsky maniac case" (1.39 mean index value), "Victoria Marinova case" (1.36 mean index value). "Rape and assault cases", which had the highest prevalence scores according to the previous measurements, has, on the other hand, a 1.14 mean visibility index within the articles where it is the most probable, which makes it one of the less interesting to the audience despite the high rates of occurrence. Also interestingly the articles on migrants being the perpetrators of sexual violence have one of the highest (1.36) visibility scores which indicates the media successful, in terms of audience engagement, interaction with themes of nationalism, as well as placing responsibility for violence on outsiders of society.

Thus, the most visible and interesting for the audience are the articles associated with public cases of violence, which were either committed by a celebrity, or the case of violence itself had a wide publicity. Topics from the general group, as well as topics related to sexual violence as a social problem, are less dominant in terms of visibility and audience interest, and for them the index values are distributed differently than the occurrence indicators.

In the further sections of this subchapter, I explore each of the topics in the six topic groups in more detail, along with their prevalence and visibility within the thematic structure of the discussion on sexual violence.

General topic group

The “General” topic group contains topics which explore different general themes within the articles covering sexual violence in Russian news media, as well as ways of describing violence. The topics in this group are: “Rape and assault cases”, “Victim perspective”, “Descriptive narratives”, “Public statements”, “Sexual harassment”, “Compensations for victims”.

Rape and assault cases

According to the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, “Rape and assault cases” is the most prevalent topic in the corpus. This topic is present, having more than 0.1 probability of being discussed, in 83.14% articles (as shown in Figure 8), being the most probable topic in 39.14% articles (as shown in Figure 9). The mean value of probability for this topic is 0.247 as shown in Figure 10. However, for the articles where this topic is the most probable, the mean value of the visibility index is only 1.14 which is, as shown in Figure 11, is significantly lower compared to most other topics in the corpus. Therefore, the “Rape and assault cases” topic is though the most prevalent, yet not the most prominent topic within the discussion on sexual violence.

The “Rape and assault cases” topic includes such words as “напасть”, “злоумышленник” and “надругаться” and essentially describes the very subject of the discussion on sexual violence – rape, assaults and other forms of sexual violence against women. Articles highly associated with this topic are headlined as, for example, “В российском регионе мужчина взломал дверь дома и изнасиловал женщину” (Lenta.Ru, 06.05.2023) or “Восемнадцатилетнюю саратовчанку в лифте изнасиловал сосед-наркоман” (Комсомольская правда, 19.12.2016). As shown in Figure 12, the topic usage in the coverage of sexual violence against women in Russian news media is more or less evenly distributed along the study timeline, with peaks of

monthly topic probability (and therefore, the topic prominence within the monthly period) at the end of the year 2019 and also the year 2022.

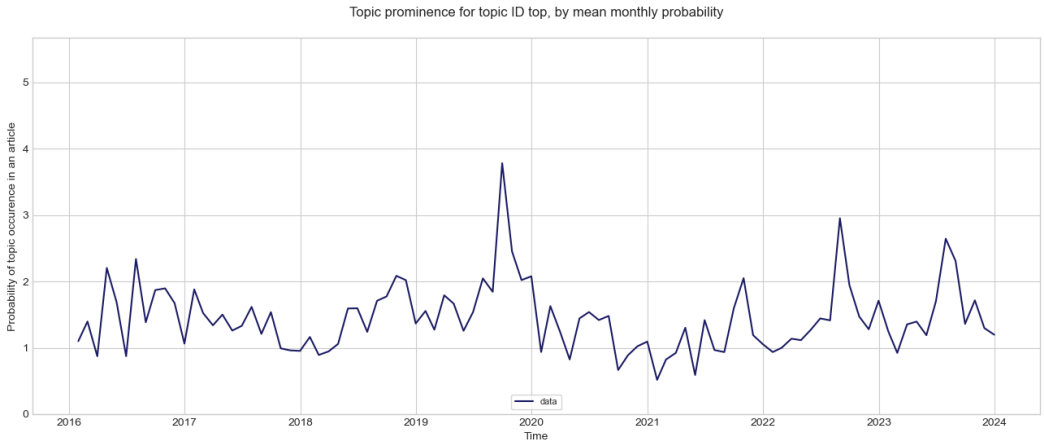

Figure 12. The dynamic of the Rape and assault cases topic, by monthly mean probability

Victim perspective

According to different measurements of topic prevalence, the "Victim perspective" topic is one of the most prevalent topics in the corpus. This topic is present, having more than 0.1 probability of being discussed, in 60.13% articles (as shown in Figure 8), being the most probable topic in 3.79% articles (as shown in Figure 9). The mean value of probability for this topic is 0.056 as shown in Figure 10. However, for the articles where this topic is the most probable, the mean value of the visibility index is only 1.27 which is, as shown in Figure 11, is lower compared to about half of the other topics in the corpus. Therefore, the "Victim perspective" topic, being one of the most prevalent ones in the corpus, is less prominent than half the themes discussed in the articles.

The "Victim perspective" topic includes such words as "человек", "история", "говорить" and "бояться" and essentially explores victims' stories and narratives about the violence that happened to them. The examples of the articles highly associated with this topic are: "#яНеБоюсьСказать: сотни женщин рассказали в Фейсбуке о сексуальном насилии над собой" (Комсомольская правда, 12.08.2018) and "«Ты привыкаешь, если тебя учат этому с детства»: RT поговорил с жертвой

сексуального насилия в семье" (RT, 15.07.2019). Moreover, along with the victims' narratives, this topic explores the ways of dealing with the consequences of sexual violence, mostly psychological assistance: "Как помочь жертве сексуального насилия, психологическая помощь в Тюмени" (72.ru, 13.12.2023). As shown in Figure 13, the topic usage in the coverage of sexual violence against women in Russian news media is more or less evenly distributed along the study timeline, however there is a growth visible in the topic usage by the end of the year 2017, which corresponds with the MeToo campaign and overall growth in interest in the victims' narratives within this period of time.

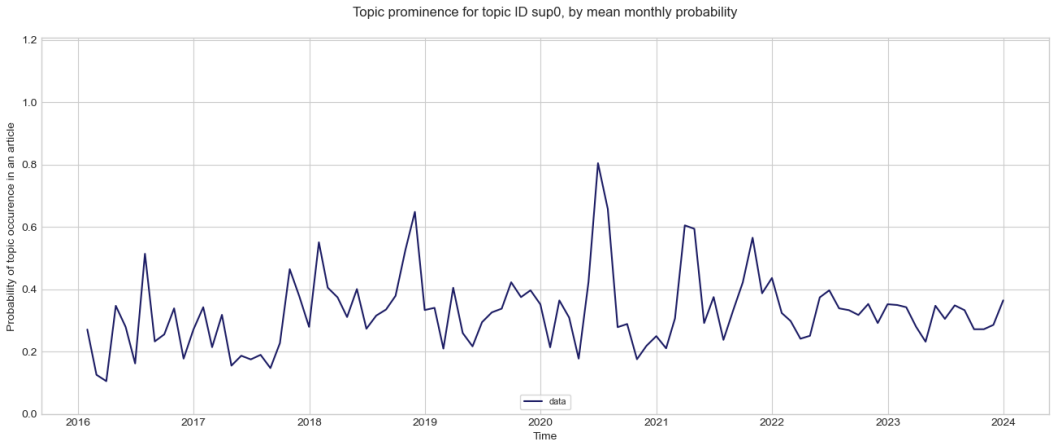

Figure 13. The dynamic of the Victim perspective topic, by monthly mean probability

Descriptive narratives

According to different measurements of topic prevalence, the "Descriptive narratives" topic is also one of the most prevalent topics in the corpus. This topic is present, having more than 0.1 probability of being discussed, in 74.01% articles (as shown in Figure 8), being the most probable topic in 6.51% articles (as shown in Figure 9). The mean value of probability for this topic is 0.06 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.31 which is, as shown in Figure 11, is lower compared to about one third of the other topics in the corpus. Therefore, the "Descriptive narratives" topic, being one of the most

prevalent ones in the corpus, is less prominent than one third of the themes discussed in the articles.

The “Descriptive narratives” topic includes such words as “тело” and “убийца” and is present in the articles where sexual violence cases covered with the descriptions, sometimes graphic, of what happened to the victim. The examples of the articles highly associated with this topic are: “Насильник задушил жертву после выхода из тюрьмы” (Дни.Ру, 03.10.2017) and “Новые подробности: у жертвы Алвеса нашли остатки спермы при медосмотре, а в туалете клуба были отпечатки футболиста” (Sports.ru, 25.01.2023). As can be seen from the examples of articles, the topic “Descriptive narratives” dives into the details of cases from a bodily point of view. As shown in Figure 14, the topic usage in the coverage of sexual violence against women in Russian news media is more or less evenly distributed along the study timeline, with a little higher mean monthly probability of being used in an article within the years 2018 – 2021.

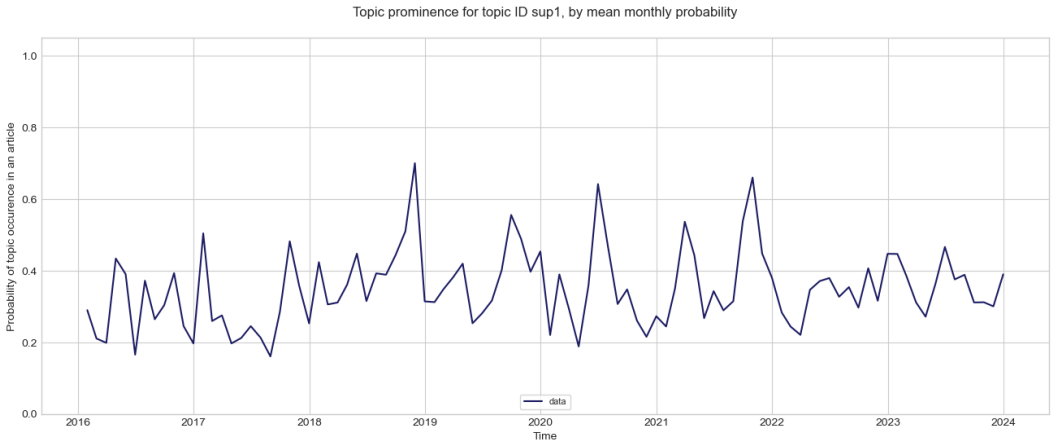

Figure 14. The dynamic of the Descriptive narratives topic, by monthly mean probability

Public statements

The “Public statements” topic is present, having more than 0.1 probability of being discussed, in 15.93% articles (as shown in Figure 8), being the most probable topic in 1.5% articles (as shown in Figure 9). The mean value of probability for this topic is 0.01 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.29 as shown in Figure 11.

The “Public statements” topic includes such words as “рассказать” and “признаться”, as well as “актриса” and “звезда”, since the public statements are mostly made by famous women regarding their personal experience with sexual violence. The topic, therefore, is present in the articles about the victims of violence telling the large public about sexual violence that happened to them, with the emphasis on the perpetrator accusation. The examples of the articles highly associated with this topic are: “Певица Бьорк рассказала о домогательствах датского режиссера” (РИА Новости, 16.10.2017) and “Известный фокусник Копперфильд обвинен в домогательствах” (ИА Росбалт, 25.01.2018). As shown in Figure 15, the topic usage in the coverage of sexual violence against women in Russian news media is growing throughout the study timeline, with peaks at the beginning of the year 2018 as well as the middle of the year 2020, and the highest peak and most overall prevalence throughout the year 2021.

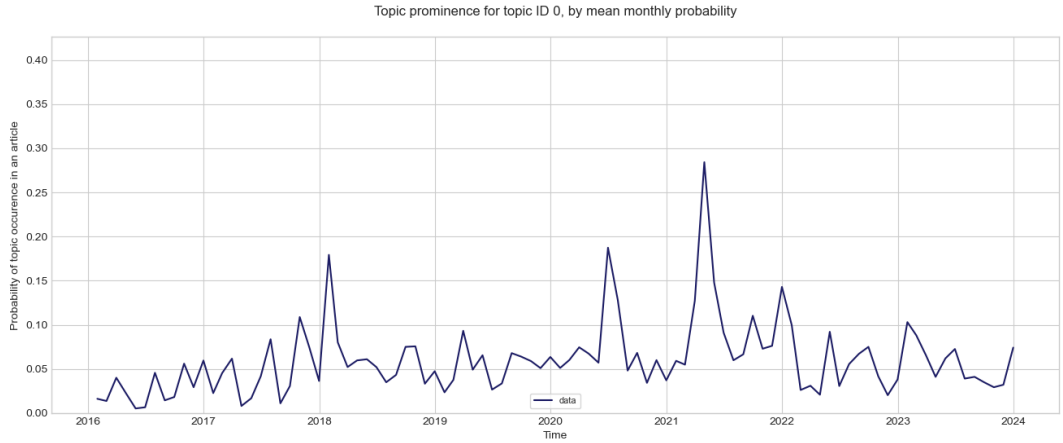

Figure 15. The dynamic of the Public statements topic, by monthly mean probability

Sexual harassment

The “Sexual harassment” topic is present, having more than 0.1 probability of being discussed, in 19.07% articles (as shown in Figure 8), being the most probable topic in 1.62% articles (as shown in Figure 9). The mean value of probability for this topic is 0.012 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.24 as shown in Figure 11.

The “Sexual harassment” topic includes such words as “домогательство”, “харассмент”, “приставать” and “рабочий”, indicating the prevalence of sexual harassment theme as one of the forms of sexual violence, including the discussion of harassment in the workplace, in the corpus. The examples of the articles highly associated with this topic are: “Большинство граждан РФ даже не знают о домогательствах на работе — опрос” (ИА Красная весна, 31.07.2020) and “Оксана Пушкина готовит закон о защите от домогательств” (Бизнес Online, 27.02.2018). As shown in Figure 16, the topic usage in the coverage of sexual violence against women in Russian news media is the highest around the end of the year 2017 and the year 2018, also peaking in the middle of year 2020.

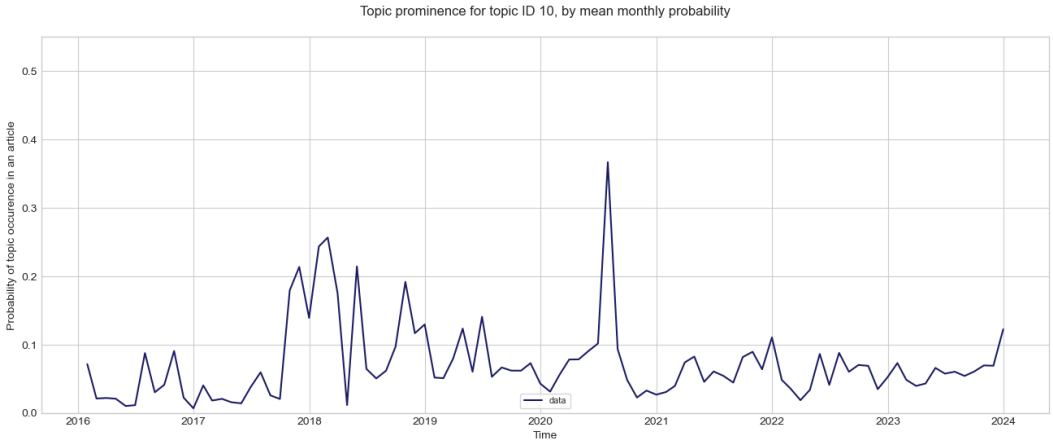

Figure 16. The dynamic of the Sexual harassment topic, by monthly mean probability

Compensations for victims

The “Compensations for victims” topic is present, having more than 0.1 probability of being discussed, in 11.01% articles (as shown in Figure 8), being the most probable topic in 0.81% articles (as shown in Figure 9). The mean value of probability for this topic is 0.006 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.18 as shown in Figure 11.

The “Compensations for victims” topic includes such words as “выплатить”, “компенсация” and “сумма”, and describes the economic retribution for the victims after the experienced sexual violence. The examples of the articles highly associated with this topic are: “В США жертвы гинеколога-насильника из университета UCLA получат $700 млн” (ИА Regnum, 25.05.2022) and “Федерация спортивной гимнастики США выплатит $ 215 млн жертвам домогательств” (Чемпионат.com, 31.01.2020). As shown in Figure 17, the topic usage is quite low throughout the study timeline, though there is a significant growth in topic prevalence in the coverage of sexual violence in the year 2021.

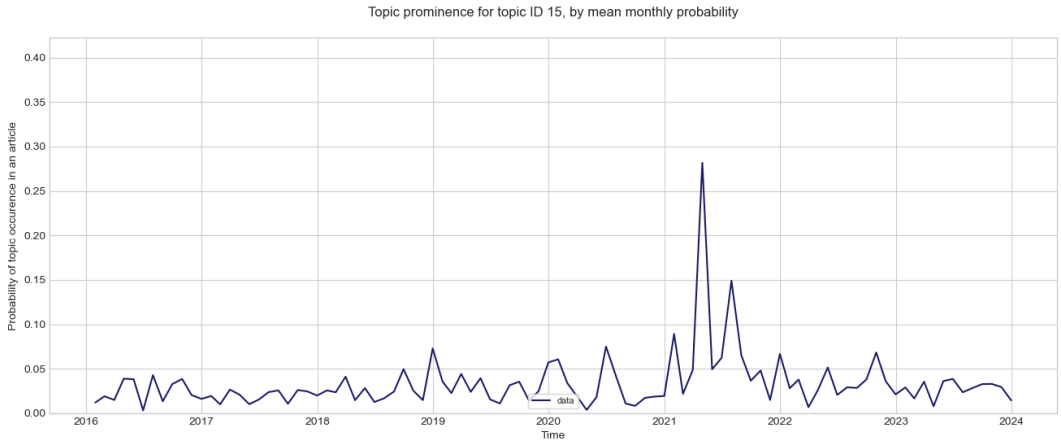

Figure 17. The dynamic of the Compensations for victims topic, by monthly mean probability

Social problem topic group

The second class of topics within the thematic structure of the corpus modeled by the HLDA model is the “Social problem” topics group which includes the topics that essentially portray sexual violence in terms of a problem that holds societal causes as well as societal and consequences. The topics included in this group are: “Problem regulation”, “Reports and statistics”, “Fighting SV - organizations and funding”, “Nobel prize for fighting SV”.

Problem regulation

According to the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, “Problem regulation” is the most prevalent topic in the corpus. This topic is present, having more than 0.1 probability of being discussed, in 76.84% articles (as shown in Figure 8), being the most probable topic in 11.41% articles (as shown in Figure 9). The mean value of probability for this topic is 0.08 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.26 as shown in Figure 11, therefore making this topic not very attention-catching despite the high rate of prevalence in the corpus.

The “Problem regulation” topic includes such words as “человек”, “право”, “социальный” and “система”, and explores the ways of dealing with the problem of sexual violence on a legislative level, such as regulations in the sphere of women’s rights and social protection programs. The examples of the articles highly associated with this topic are: “Власти заказали исследование в целях равных прав для женщин. Госдума изучит российский и зарубежный опыт в целях совершенствования законодательства в сфере защиты прав женщин” (Российская газета, 02.07.2018) and “Минтруд запланирует убрать «стеклянный потолок» и «липкий пол» для женщин” (Комсомольская правда, 07.08.2022). As shown in Figure 18, the topic became more prevalent by the end of the year 2018, with many

peaks throughout the study period and overall high prevalence in the corpus throughout the study period.

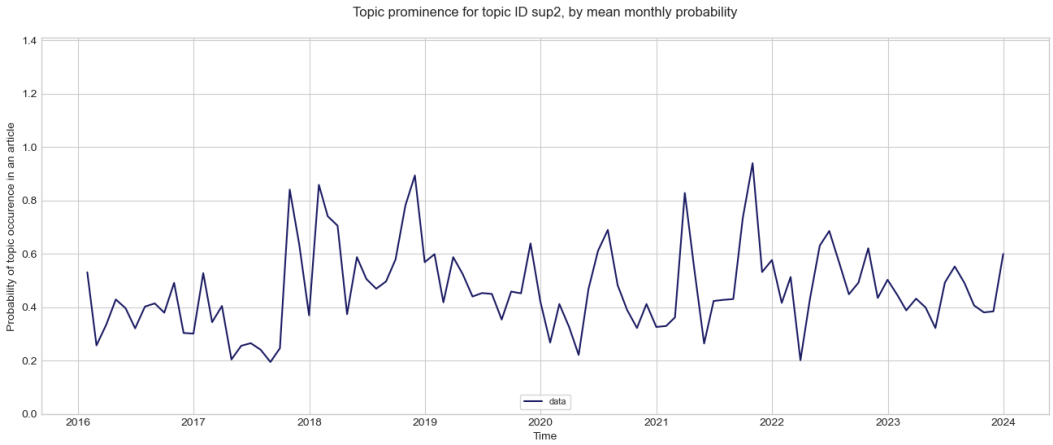

Figure 18. The dynamic of the Problem regulation topic, by monthly mean probability

Reports and statistics

The “Reports and statistics” topic is present, having more than 0.1 probability of being discussed, in 34.02% articles (as shown in Figure 8), being the most probable topic in 3.29% articles (as shown in Figure 9). The mean value of probability for this topic is 0.025 as shown in Figure 10. Thus, according to the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, the “Reports and statistics” topic is one of the most prevalent topics in the corpus. However, for the articles where this topic is the most probable, the mean value of the visibility index is 1.13 as shown in Figure 11, indicating that though the topic is highly prevalent in the coverage of sexual violence in Russian news media, the articles containing it aren’t most likely to catch high levels of the audience’s attention.

The “Reports and statistics” topic includes such words as “подвергаться”, “правительство” and “отчёт”, and combines all the ways in which the problem of sexual violence is estimated and evaluated, usually by either the national governments or international non-profit organizations in a form of statistical reports on cases or the

problem as a whole. Moreover, the topic includes mentions of methods for solving a problem as coming from reports. The examples of the articles highly associated with this topic are: “В Дублине на треть выросло число обращений по поводу изнасилований” (ИА Красная весна, 19.10.2023) and “ЕП подготовил проект резолюции по борьбе с сексуальными домогательствами” (РИА Новости, 25.10.2017). As shown in Figure 19, the topic became more prevalent by the end of the year 2017, with many peaks throughout the study period with a slight fall in the beginning of the year 2020 and then with the return to the topic usage in the coverage of sexual violence by the end of the year 2020.

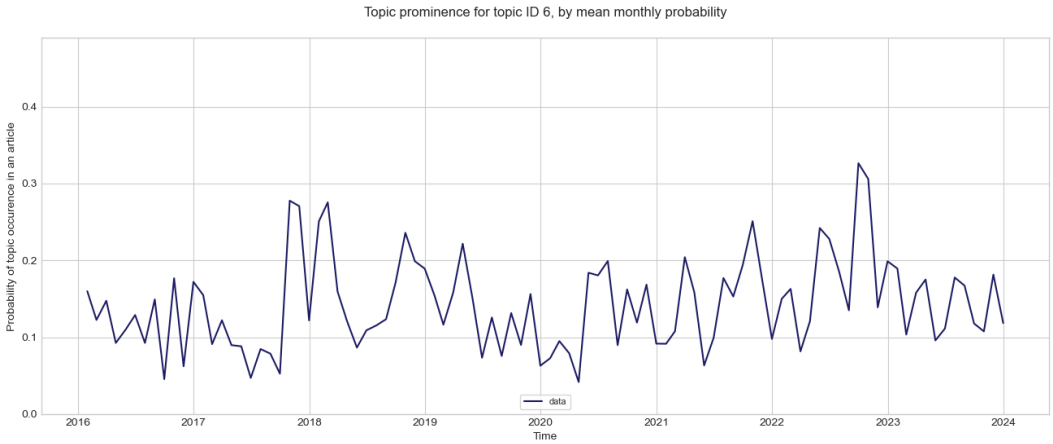

Figure 19. The dynamic of the Reports and statistics topic, by monthly mean probability

Fighting SV - organizations and funding

The “Fighting SV - organizations and funding” topic is present, having more than 0.1 probability of being discussed, in 12.05% articles (as shown in Figure 8), being the most probable topic in 0.94% articles (as shown in Figure 9). The mean value of probability for this topic is 0.007 as shown in Figure 10. Thus, according to the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, the “Reports and statistics” topic prevalence in the corpus is quite low. Nevertheless, for the articles where this topic is the most probable, the mean value of the visibility index is 1.33 as

shown in Figure 11, indicating that though the topic has low prevalence in the coverage of sexual violence in Russian news media, the articles containing are likely to catch high levels of the audience's attention.

The "Fighting SV - organizations and funding" topic includes such words as "фонд", "влиятельный" and "заплатить" and indicates the ways that various work with the consequences of the violence is organized and funded, including personal donations from famous people. The examples of the articles highly associated with this topic are: "Более 300 женщин из киноиндустрии США основали движение по борьбе с домогательствами" (Коммерсантъ. Новости информ. центра, 02.01.2018) and "Эмма Уотсон пожертвовала $1,4 млн на борьбу с домогательствами и дискриминацией женщин" (ТАСС, 18.02.2018). As shown in Figure 20, the topic had almost zero prevalence, in terms of monthly mean probability, in the coverage of sexual violence until the year 2018, however, the peak usage of this topic by Russian media ended within a year, with a further low usage throughout the remaining years within the study timeline.

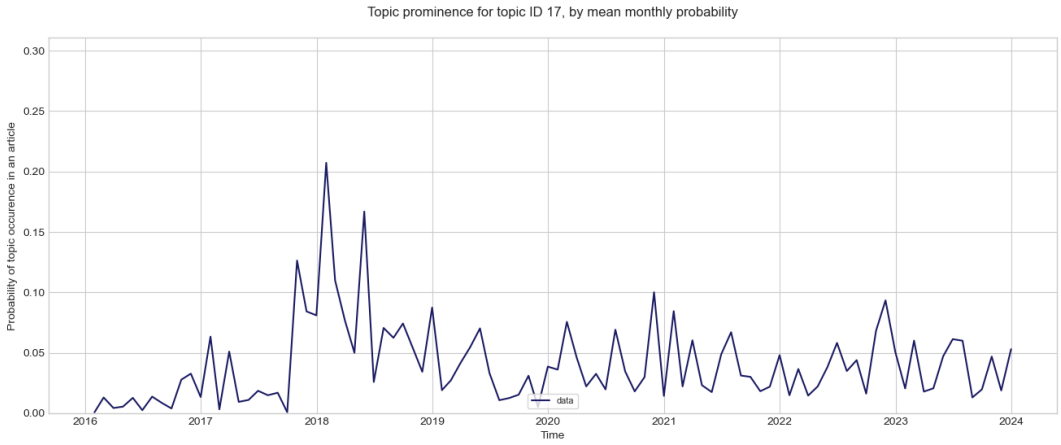

Figure 20. The dynamic of the Fighting SV - organizations and funding topic, by monthly mean probability

Nobel prize for fighting SV

This topic is the last within the “Social problem” topic group and indicates the case of the Nobel peace prize ceremony held in 2018, where, following the MeToo campaign, the laureates announced were people who were involved in the fight against sexual violence. Despite the fact that the case was prevalent in the news for a very short time (see Figure 21), I classified it as a part of the “Social problem” topic group because despite the narrow time frame, this topic also considers sexual violence as a social problem. The topic is present, having more than 0.1 probability of being discussed, in 6.56% articles (as shown in Figure 8), being the most probable topic in 0.36% articles (as shown in Figure 9). The mean value of probability for this topic is 0.004 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is also quite low for this topic – 1.13 as shown in Figure 11.

The “Nobel prize for fighting SV” topic includes such words as “премия”, “мир” and “нобелевский”. The examples of the articles highly associated with this topic are: “Нобелевская премия мира досталась борцам с сексуальным насилием в войнах” (Аргументы и факты, 05.10.2018) and “Лауреат Нобелевской премии рассказал о помощи женщинам — жертвам насилия” (Российская газета, 09.12.2018).

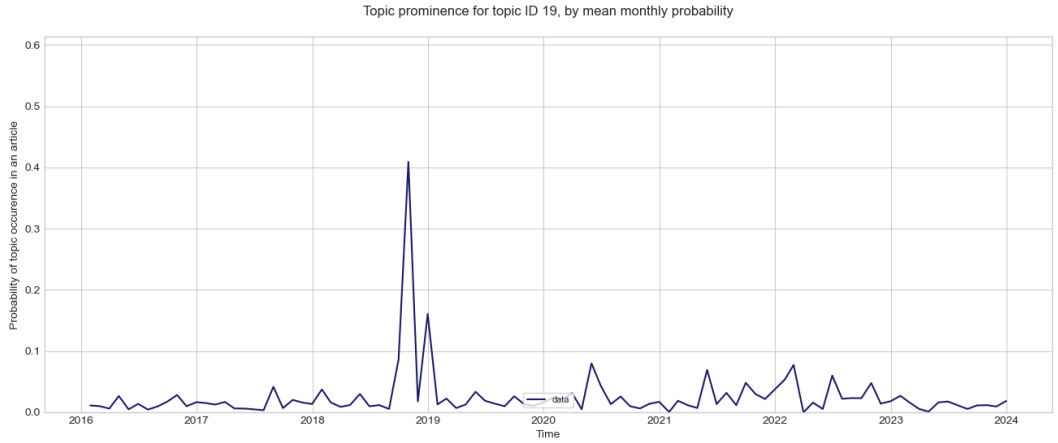

Figure 21. The dynamic of the Nobel prize for fighting SV topic, by monthly mean probability

Criminal cases & process topic group

This group consists of the following topics: “Perpetrators detention”, “Perpertrators crime sentences”, “Details of criminal cases processes”, “Perpertrators court sentencing” and “Perpertrators court sentences murder cases”.

Perpetrators detention

The “Perpetrators detention” topic is present, having more than 0.1 probability of being discussed, in 22.34% articles (as shown in Figure 8), being the most probable topic in 1.75% articles (as shown in Figure 9). The mean value of probability for this topic is 0.02 as shown in Figure 10. In terms of the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, the “Perpetrators detention” topic is quite prevalent in the corpus, though still is not one of the most prevalent ones such as the aforementioned topics from the “General” topic group as well as the other topics from the “Criminal cases & process” group as will be shown below. However, for the articles where this topic is the most probable, the mean value of the visibility index is 2.17 as shown in Figure 11, which is the highest value among all the topics that indicates that, within all the thematic structure of the discussion on sexual violence in Russian news media, the usage of this topic catches the audience’s attention the most.

The “Perpetrators detention” topic includes such words as “изнасилование”, “задержать” and “подозреваться”, and essentially contains procedural descriptions and reports on perpetrators’ detention as the initial part of the rape criminal cases. The examples of the articles highly associated with this topic are: “В Томской области задержали подозреваемого в серии изнасилований” (РИА Новости, 13.10.2019) and “В Москве арестовали насильника, ограбившего женщину в Битцевском парке” (РИА Новости, 09.07.2021). As shown in Figure 22, the topic usage in the coverage of sexual violence against women in Russian news media is more or less

evenly distributed along the study timeline, with several peaks of monthly topic probability in the years 2019, 2020 and 2022.

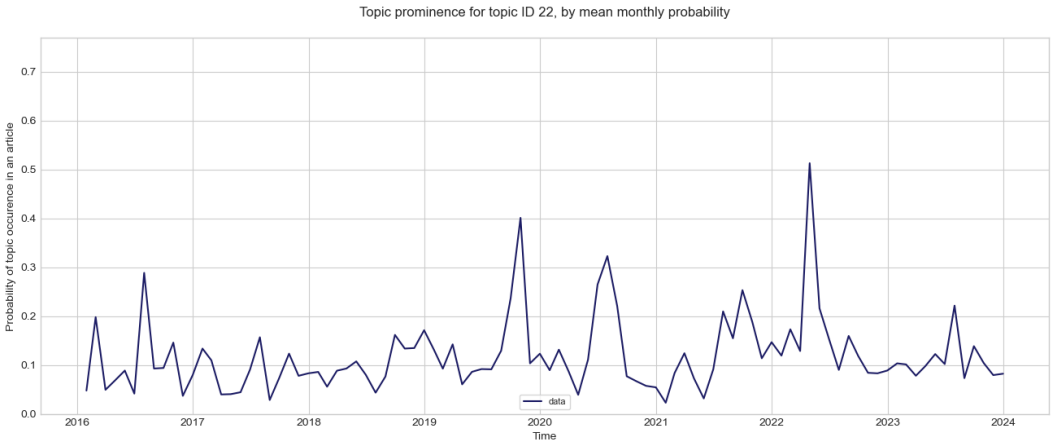

Figure 22. The dynamic of the Perpetrators detention topic, by monthly mean probability

Perpertrators crime sentences

The “Perpetrators crime sentences” topic is present, having more than 0.1 probability of being discussed, in 33.8% articles (as shown in Figure 8), being the most probable topic in 3.13% articles (as shown in Figure 9). The mean value of probability for this topic is 0.036 as shown in Figure 10. This topic, according to the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, is one of the most prevalent in the corpus. For the articles where this topic is the most probable, the mean value of the visibility index is 1.31 as shown in Figure 11, which makes it less prominent despite the high usage in the articles on sexual violence in Russian news media.

The “Perpetrators crime sentences” topic includes such words as “колония”, “пожизненный” and “назначать”, and essentially is about criminal penalties that rapists and other perpetrators of sexual violence bear for the committed crimes, therefore describing the final stages of the criminal processes against sexual violence perpetrators. The examples of the articles highly associated with this topic are: “Житель

Череповца получил пожизненный срок за убийство и изнасилование" (РИА Новости, 11.05.2017) and "Суд добавил к двум пожизненным срокам «ангарского маньяка» почти 10 лет" (РБК, 04.06.2021). As shown in Figure 23, the topic usage in the coverage of sexual violence against women in Russian news media is more or less evenly distributed along the study timeline, with several growths in usage around the years 2018-2019 and also the years 2021-2022.

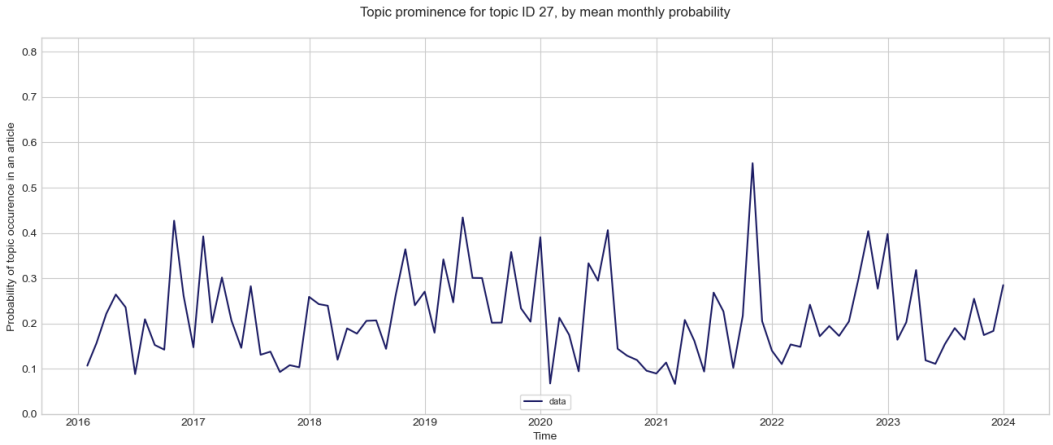

Figure 23. The dynamic of the Perpertrators crime sentences topic, by monthly mean probability

Details of criminal cases processes

The "Details of criminal cases processes" topic is present, having more than 0.1 probability of being discussed, in 47.09% articles (as shown in Figure 8), being the most probable topic in 3.47% articles (as shown in Figure 9). The mean value of probability for this topic is 0.057 as shown in Figure 10. This topic, according to the topic prevalence measured by the absolute number of appearances of the topic in the articles on sexual violence in Russian news media, as well as by the proportions of the articles by the most probable topics and mean topic probabilities in the corpus, is one of the most prevalent in the corpus. For the articles where this topic is the most probable, the mean value of the visibility index is 1.34 as shown in Figure 11, which indicates a more or less high levels of the audience attention to the articles using this topic.

The "Details of criminal cases processes" topic includes such words as "дело", "уголовный" and "версия" and may be interpreted as the topic that contains the

descriptions of the criminal cases for the sexual violence perpetrators, with the focus on the case process after the initial detention (“Perpetrators detention” topic) and the final sentencing (“Perpetrators crime sentences” topic). The examples of the articles highly associated with this topic are: “Дело об убийстве трех девушек в Оренбургской области. Главное” (ТАСС, 05.10.2021) and “Завершено расследование уголовного дела об изнасиловании женщины на западе Москвы” (РИАМО, 23.05.2023). As shown in Figure 24, the topic usage in the coverage of sexual violence against women in Russian news media is more or less evenly distributed along the study timeline, with a significant peak at the end of the year 2021.

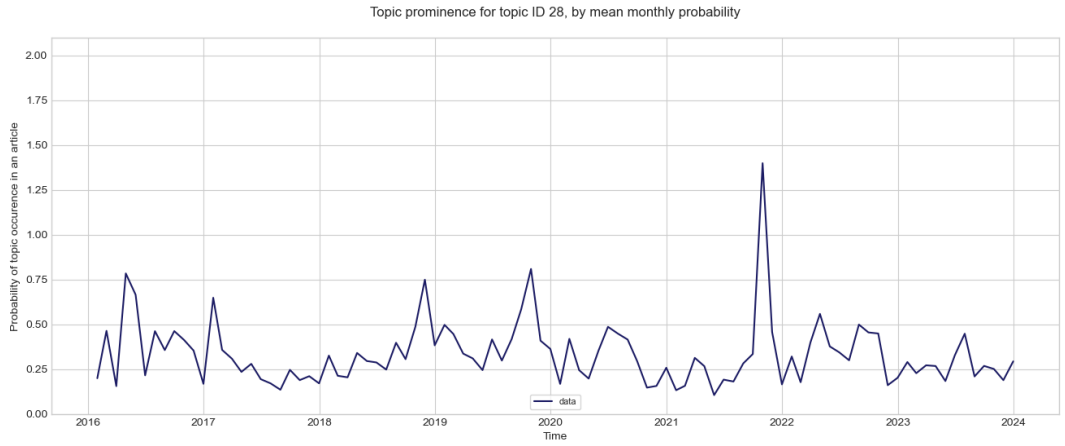

Figure 24. The dynamic of the Details of criminal cases processes topic, by monthly mean probability

Perpertrators court sentencing

The “Perpertrators court sentencing” topic is present, having more than 0.1 probability of being discussed, in 32.53% articles (as shown in Figure 8), being the most probable topic in 3.14% articles (as shown in Figure 9). The mean value of probability for this topic is 0.028 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.28 as shown in Figure 11, which indicates a more or less high levels of the audience attention to the articles using this topic.

The “Perpertrators court sentencing” topic includes such words as “судебный”, “процесс”, “обвинение” and “прокурор” and essentially is about the process of court trial process against sexual violence perpetrators. It is important to note that there is a difference of this topic from the “Perpetrators crime sentences” which is about the court sentences themselves while the “Perpertrators court sentencing” explores the process of a court trial that leads to such sentences. In addition, this topic shows significantly higher attention to foreign public rape cases than the “Perpetrators crime sentences” topic. The examples of the articles highly associated with this topic are: “Присяжные признали Вайнштейна виновным еще в одном изнасиловании” (Коммерсантъ. Новости Online, 20.12.2022) and “ СМИ: суд отклонил апелляцию Билла Косби, признанного виновным в домогательствах” (ТАСС, 10.12.2019). As shown in Figure 25, the topic usage in the coverage of sexual violence against women in Russian news media is gradually growing throughout the whole study period.

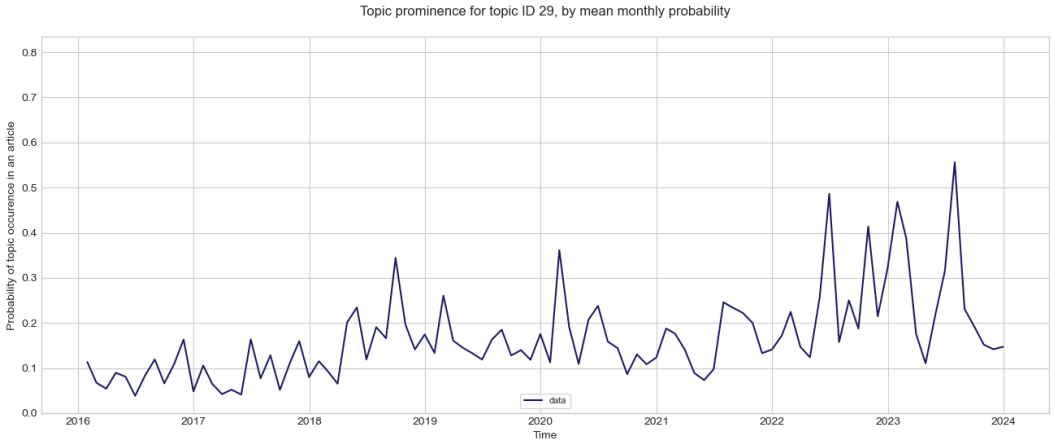

Figure 25. The dynamic of the Perpertrators court sentencing topic, by monthly mean probability

Perpertrators court sentences murder cases

The “Perpertrators court sentences murder cases” topic is present, having more than 0.1 probability of being discussed, in 20.7% articles (as shown in Figure 8), being the most probable topic in 1.62% articles (as shown in Figure 9). The mean value of probability for this topic is 0.017 as shown in Figure 10. For the articles where this topic

is the most probable, the mean value of the visibility index is 1.21 as shown in Figure 11.

The “Perpertrators court sentences murder cases” topic includes such words as “убийство”, “тяжкий” and “умышленный” and may be interpreted as the topic that contains the descriptions of the criminal sentences for the cases of the sexual violence perpetrators which also include the murder of the victim (compared to the “Perpetrators crime sentences” topic). Moreover, the murder is the central theme in the topic since the sentences are murder-related and sexual violence is included in the cases as the additional crime. The examples of the articles highly associated with this topic are: “В Оренбургской области убийцу трех студенток приговорили к пожизненному сроку” (Новые известия, 15.03.2023) and “Насильник и убийца в одном лице предстанет перед судом Твери” (Комсомольская правда, 11.07.2017). As shown in Figure 26, the topic usage in the coverage of sexual violence against women in Russian news media in terms of monthly mean topic probability is quite high since the year 2018.

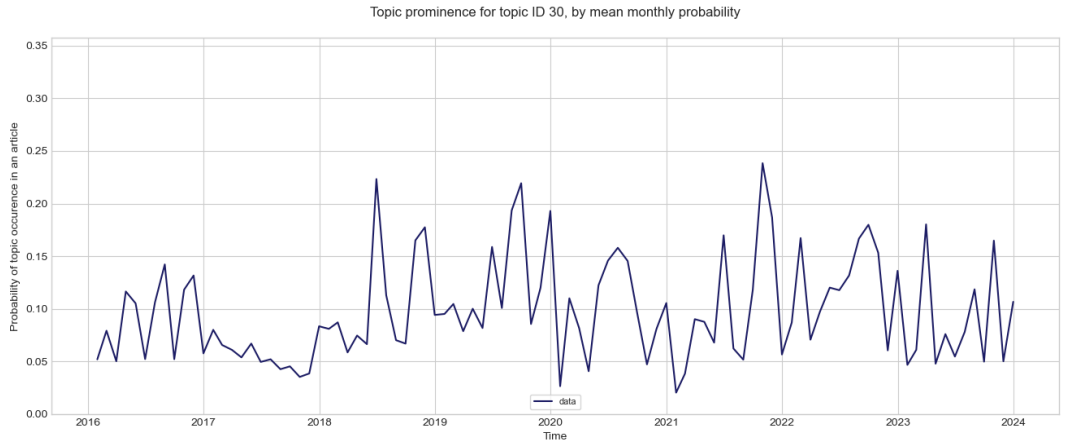

Figure 26. The dynamic of the Perpertrators court sentences murder cases topic, by monthly mean probability

Perpetrator types topic group

This topic group is thematical in a way that it includes topics which are concentrated on a certain social or professional group of people being the perpetrators of sexual violence. The “Perpetrator types” topic group, thus, includes the following

topics: “Migrants as perpetrators”, “Priests as perpetrators” and “Musicians as perpetrators”.

Migrants as perpetrators

The “Migrants as perpetrators” topic is present, having more than 0.1 probability of being discussed, in 11.1% articles (as shown in Figure 8), being the most probable topic in 0.88% articles (as shown in Figure 9). The mean value of probability for this topic is 0.008 as shown in Figure 10. Yet, being far from prevalent in terms of appearances in the articles, this topic is quite visible for the audience, being in the top-5 topics in terms of the mean value of the visibility index for the articles where this topic is the most probable, which is 1.36 as shown in Figure 11.

The “Migrants as perpetrators” topic includes such words as “мигрант”, “женщина” and “полиция” and essentially explores sexual violence, both cases and the problem, through the lens of a migrant being the perpetrator. The examples of the articles highly associated with this topic are: “В Якутии вспыхнули протесты против мигрантов после изнасилования женщины” (ИА Росбалт, 17.03.2019) and “Полиция Хельсинки готовится отбивать нападения мигрантов” (Дни.Ру, 06.02.2016). As shown in Figure 27, the topic usage is very low throughout the whole study period, with peaking just at the beginning of the year 2016.

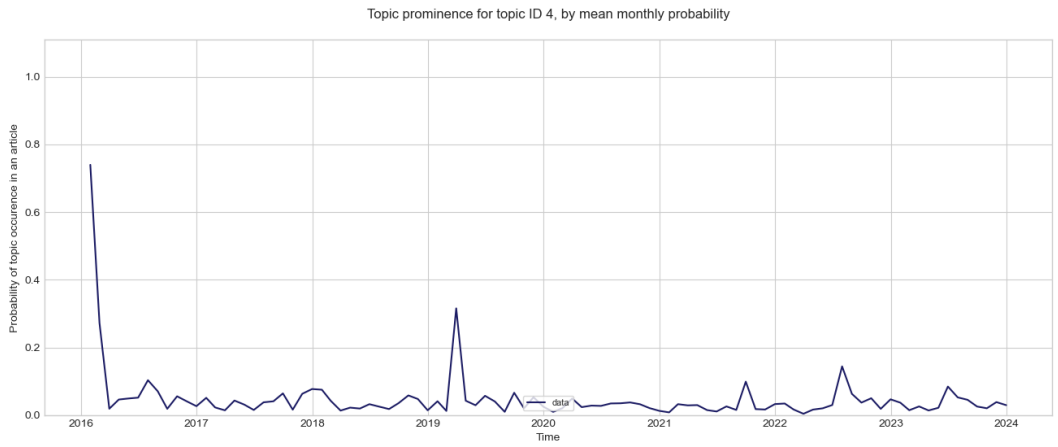

Figure 27. The dynamic of the Migrants as perpetrators topic, by monthly mean probability

Priests as perpetrators

The “Priests as perpetrators” topic is present, having more than 0.1 probability of being discussed, in 5.04% articles (as shown in Figure 8), being the most probable topic in 0.15% articles (as shown in Figure 9). The mean value of probability for this topic is 0.006 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.07 as shown in Figure 11. According to the parameters of prevalence and visibility discussed above, this topic is one of the less prevalent and less prominent in the corpus.

The “Priests as perpetrators” topic includes such words as “церковь” and “священник” and essentially explores sexual violence, both cases and the problem, through the lens of a priest being the perpetrator. The examples of the articles highly associated with this topic are: “Во Франции опубликовали доклад о сексуальном насилии священников” (РИА Новости, 05.10.2021) and “Папа Римский встретился с жертвами сексуального насилия священников” (Комсомольская правда, 17.01.2018). As shown in Figure 28, the topic usage is quite low throughout the whole study period, with peaking at the beginning of the year 2019.

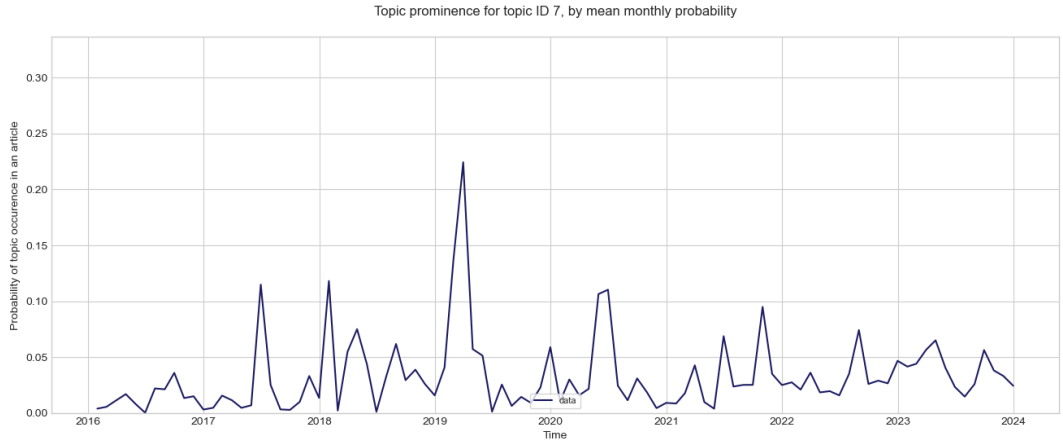

Figure 28. The dynamic of the Priests as perpetrators topic, by monthly mean probability

Musicians as perpetrators

The “Musicians as perpetrators” topic is present, having more than 0.1 probability of being discussed, in 25.99% articles (as shown in Figure 8), being the most probable

topic in 3.06% articles (as shown in Figure 9). The mean value of probability for this topic is 0.023 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.33 as shown in Figure 11.

The “Musicians as perpetrators” topic includes such words as “обвинить”, “рэпер” and “исполнитель” and essentially explores sexual violence, both cases and the problem, through the lens of a musician being the perpetrator. The examples of the articles highly associated with this topic are: “Против солиста группы Aerosmith Тайлера подали новый иск о сексуальном насилии” (ТАСС, 02.11.2023) and “Третья женщина обвинила рэпера Дидди в изнасиловании” (Газета.Ru, 24.11.2023). As shown in Figure 29, the topic usage almost gradually grows throughout the study period: the first growth in usage happened around the years 2018 and 2019, then, after a fall in usage, the topic became more probable to being used in the articles on sexual violence in Russian mass media since the year 2021.

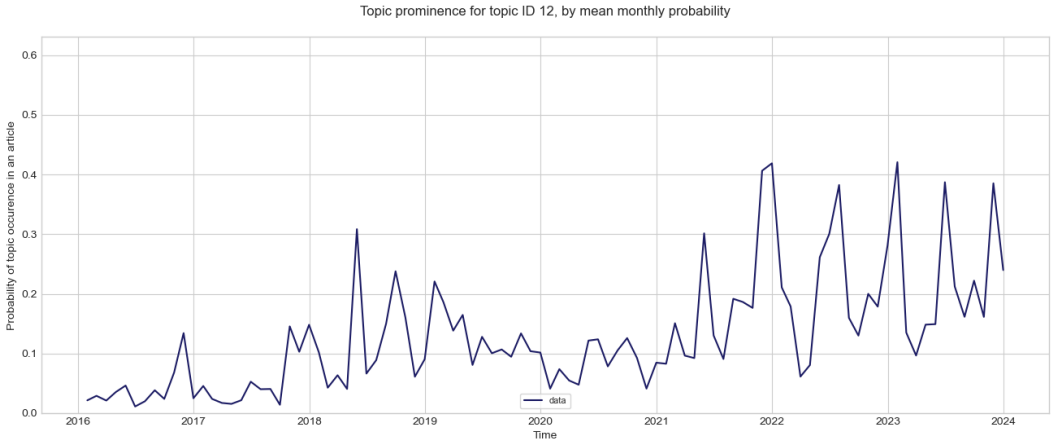

Figure 29. The dynamic of the Musicians as perpetrators topic, by monthly mean probability

Spheres of violence topic group

The “Spheres of violence” topic group is constituted by the topics which were too narrow to be interpreted as general topics and yet too wide to be interpreted as cases since they contain many cases united by specific circumstances such certain place or professional sphere. There topics in this group are the following: “Royal family scandals”, “Violence in hockey”, “Violence in figure skating”, “The convention of

womens' rights", "SV during the war in Ukraine", "Japanese program for women in sexual slavery".

Royal family scandals

The "Royal family scandals" topic is present, having more than 0.1 probability of being discussed, in 6.81% articles (as shown in Figure 8), being the most probable topic in 0.38% articles (as shown in Figure 9). The mean value of probability for this topic is 0.005 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.26 as shown in Figure 11, indicating that this topic is quite popular with the audience despite its low prevalence in the articles throughout the study period.

The "Royal family scandals" topic includes such words as "принц", "британский" and "елизавета", combining the indicators for the British royal family circumstances as well as the personalities within it. The examples of the articles highly associated with this topic are: "Британского принца Эндрю после изнасилования лишили воинских званий" (ИА Regnum, 13.01.2022) and "Объяснены скандальные действия британской королевы" (Lenta.Ru, 31.03.2022). As shown in Figure 30, the topic has non-zero yet low prevalence throughout the whole study period, with the peak in the year 2022 regarding the main prince Andrew sexual violence case.

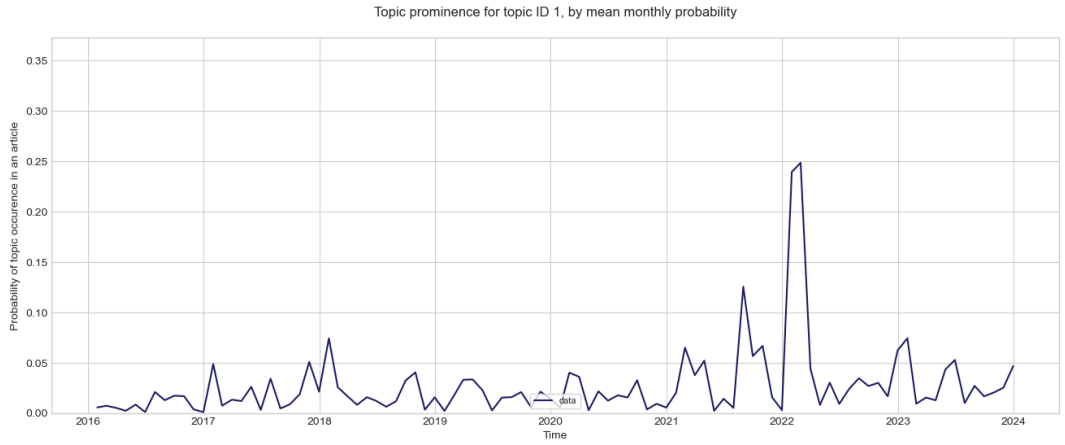

Figure 30. The dynamic of the Royal family scandals topic, by monthly mean probability

Violence in hockey

The “Violence in hockey” topic is present, having more than 0.1 probability of being discussed, in 11.82% articles (as shown in Figure 8), being the most probable topic in 0.91% articles (as shown in Figure 9). The mean value of probability for this topic is 0.016 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.25 as shown in Figure 11.

The “Violence in hockey” topic includes such words as “тренер”, “хоккеист” and “отстранить”, combining the cases of sexual violence within the hockey sport. The examples of the articles highly associated with this topic are: “Совет Федерации хоккея Канады поддержал главу организации в деле по сексуальному насилию” (Чемпионат.com, 05.09.2022) and “НХЛ начала расследование случая сексуального насилия на МЧМ-2003” (Р-Спорт, 23.07.2022). As shown in Figure 31, the topic has non-zero yet low prevalence throughout the whole study period, with the peak in the years 2021 and 2022 regarding the popular cases within this period.

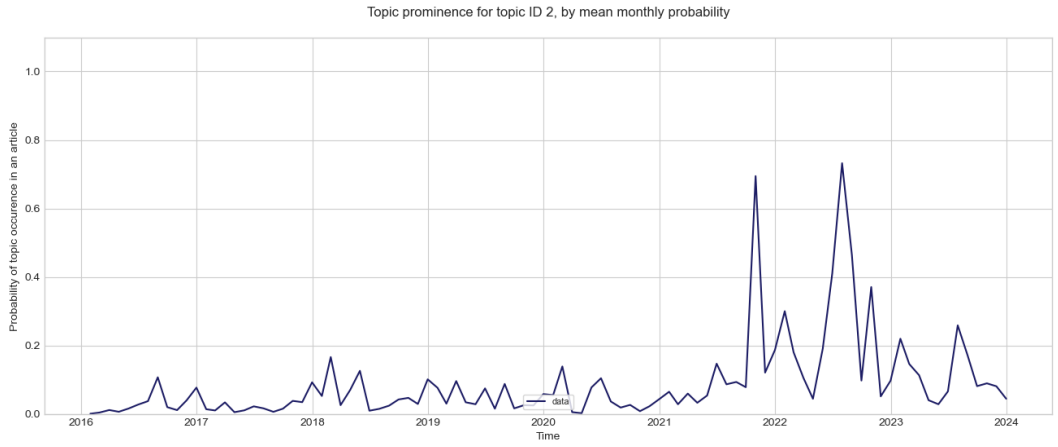

Figure 31. The dynamic of the Violence in hockey topic, by monthly mean probability

Violence in figure skating

The “Violence in figure skating” topic is present, having more than 0.1 probability of being discussed, in 6.39% articles (as shown in Figure 8), being the most probable topic in 0.36% articles (as shown in Figure 9). The mean value of probability for this topic is 0.003 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1 as shown in Figure 11, which is the lowest value among all topics.

The “Violence in figure skating” topic includes such words as “катание” and “фигуристка”, combining the cases of sexual violence within the figure skating sport. The examples of the articles highly associated with this topic are: “Глава Федерации фигурного катания Франции подал в отставку из-за скандала с изнасилованием” (RT, 08.02.2020) and “Французские фигуристки признались в домогательствах со стороны тренеров” (NEWS.ru, 30.01.2020). As shown in Figure 32, the topic has non-zero yet low prevalence throughout the whole study period, with the peak at the beginning of the year 2020.

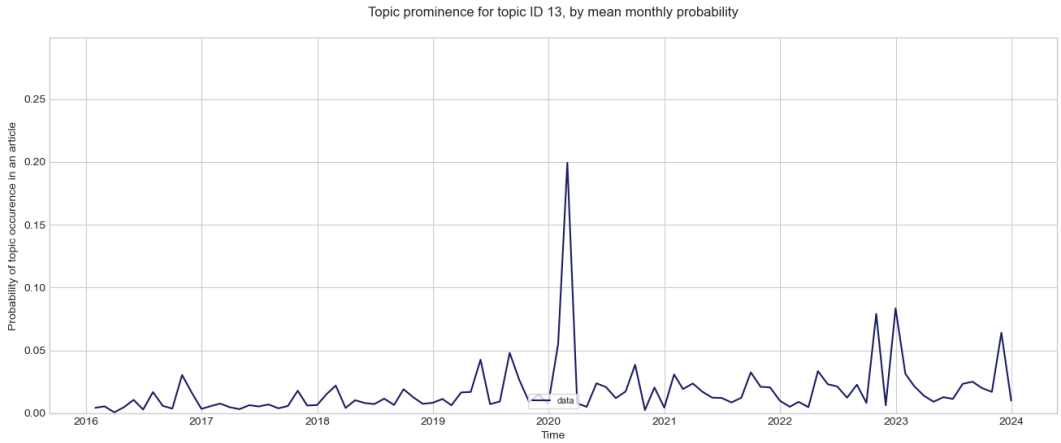

Figure 32. The dynamic of the Violence in figure skating topic, by monthly mean probability

The convention of womens' rights

The “The convention of womens' rights” topic is present, having more than 0.1 probability of being discussed, in 13.37% articles (as shown in Figure 8), being the most probable topic in 1.03% articles (as shown in Figure 9). The mean value of probability

for this topic is 0.009 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.3 as shown in Figure 11, which makes the topic quite visible regarding its low prevalence.

The “The convention of womens' rights” topic includes such words as “гендерный”, “конвенция” and “равенство”, indicating the coverage of the corresponding document in the sphere of regulation the violence against women. The examples of the articles highly associated with this topic are: “Турция вышла из Стамбульской конвенции о борьбе с насилием в отношении женщин” (Коммерсантъ. Новости Online, 20.03.2021) and “Пушкина оценила выход Турции из Стамбульской конвенции по защите женщин” (РБК, 21.03.2021). As shown in Figure 33, the topic has non-zero yet low prevalence throughout the whole study period, with a major peak at the beginning of the year 2021.

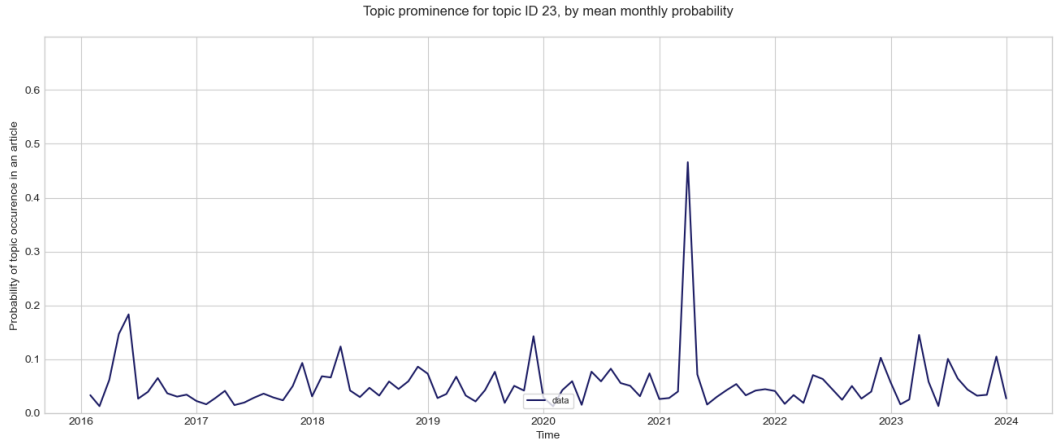

Figure 33. The dynamic of the The convention of womens' rights topic, by monthly mean probability

SV during the war in Ukraine

The “SV during the war in Ukraine” topic is present, having more than 0.1 probability of being discussed, in 8.32% articles (as shown in Figure 8), being the most probable topic in 0.41% articles (as shown in Figure 9). The mean value of probability for this topic is 0.005 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.15 as shown in Figure 11.

The “SV during the war in Ukraine” topic includes such words as “украина”, “россия” and “оружие”, indicating the media coverage of the cases of sexual violence that are occur during the military aggression of Russian Federation in Ukraine. The examples of the articles highly associated with this topic are: “Россия отметила всплеск сексуального насилия в подконтрольных Киеву районах” (NEWS.ru, 13.04.2022) and “Жена Зеленского заявила, что российским солдатам «приказали насиловать украинок»” (ИА SM-News, 29.11.2022). As shown in Figure 34, the topic usage has several peaks throughout the study period: first in year 2016, then in year 2020 and finally in the year 2022 when the war turned to a full-scale invasion stage.

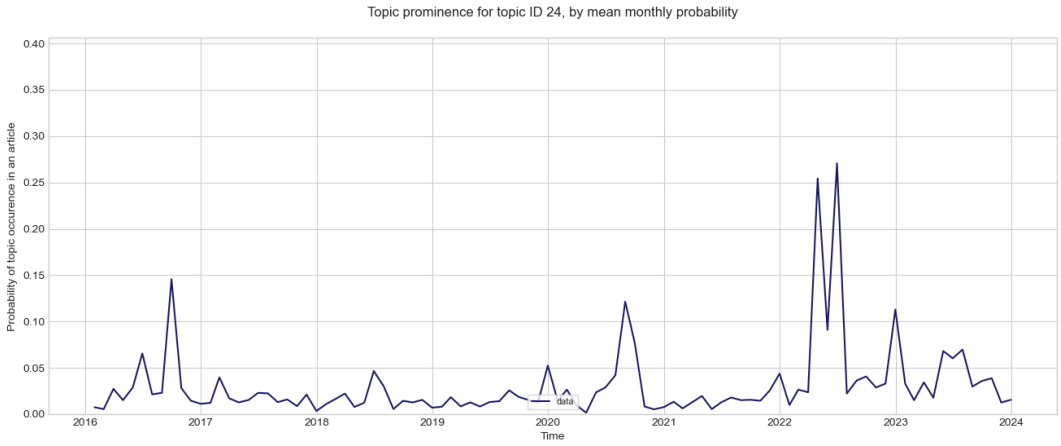

Figure 34. The dynamic of the SV during the war in Ukraine topic, by monthly mean probability

Japanese program for women in sexual slavery

The “Japanese program for women in sexual slavery” topic is present, having more than 0.1 probability of being discussed, in 4.59% articles (as shown in Figure 8), being the most probable topic in 0.07% articles (as shown in Figure 9). The mean value of probability for this topic is 0.004 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.23 as shown in Figure 11.

The “Japanese retributions for sexual slavery” topic includes such words as “япония”, “корея” and “женщина”, describing the conflict between Japan and Korea over Japan's payment of compensation for taking women into sexual slavery in the first

half of the 20th century, where all compensation goes to funds to help women who suffered violence. The examples of the articles highly associated with this topic are: “Токио заявил протест Сеулу в связи с проблемой «женщин для утешения»” (EaDaily.com, 01.03.2018) and “Президент Южной Кореи раскритиковал Японию за позицию по проблеме "женщин для утешения"” (ТАСС, 01.03.2018). As shown in Figure 35, the topic usage is more or less stable throughout the whole study period, with a peak in its usage at the beginning of the year 2017.

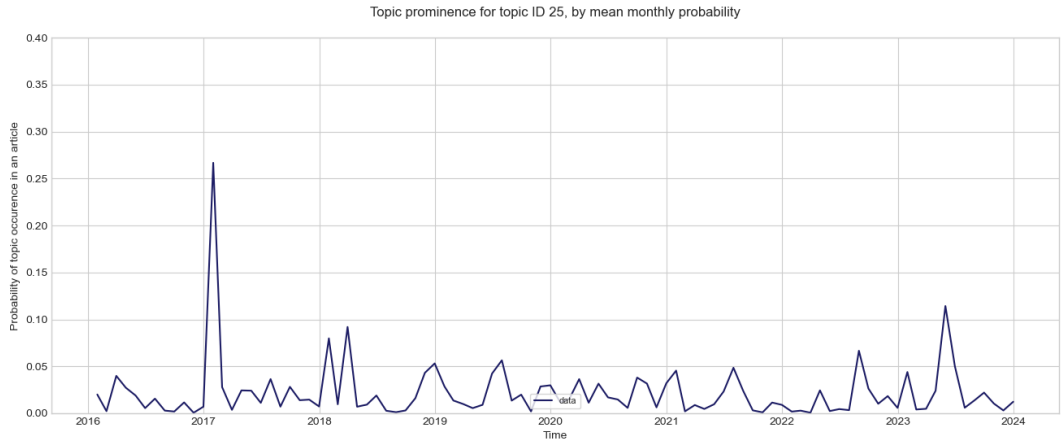

Figure 35. The dynamic of the Japanese program for women in sexual slavery topic, by monthly mean probability

Cases topic group

The “Cases” topic group contains many topics that explore the coverage of specific cases that were popular in the media and discourse throughout the whole study period and mostly are associated with the names of a victim or a perpetrator of sexual violence. The topics in this group are the following: “Khachaturian sisters case”, “Luis Rubiales case”, “Harvey Weinstein case”, “Cristiano Ronaldo case”, “Benjamin Mendy case”, “Gérard Depardieu case”, “Daniel Alves case”, “Skopinsky maniac case”, “Jeffrey Epstein case”, “Victoria Marinova case”, “Rape in the Ufa police department case”, “Roman Polanski case”, “Collectors case”.

Khachaturian sisters case

The “Khachaturian sisters case” topic is present, having more than 0.1 probability of being discussed, in 11% articles (as shown in Figure 8), being the most probable topic in 0.75% articles (as shown in Figure 9). The mean value of probability for this topic is 0.009 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.33 as shown in Figure 11 – one of the highest values among all the topics. Therefore, it can be said that the Khachaturian sisters case, despite low prevalence (due to the episodic coverage most likely), was very prominent in terms of the audience attention.

The “Khachaturian sisters case” topic includes such words as “сестра”, “отец” and “хачатурян”, describing the case of murder of Mikhail Khachaturian perpetrated by his daughters after the many years of domestic violence and sexual abuse. The examples of the articles highly associated with this topic are: “Защита сестер Хачатурян ожидает прекращения дела, заявил адвокат” (РИА Новости, 31.01.2020) and “Генпрокуратура признала убийство отца сестрами Хачатурян необходимой обороной” (Новые Известия, 31.01.2020). As shown in Figure 36, the topic usage corresponds with the case criminal process.

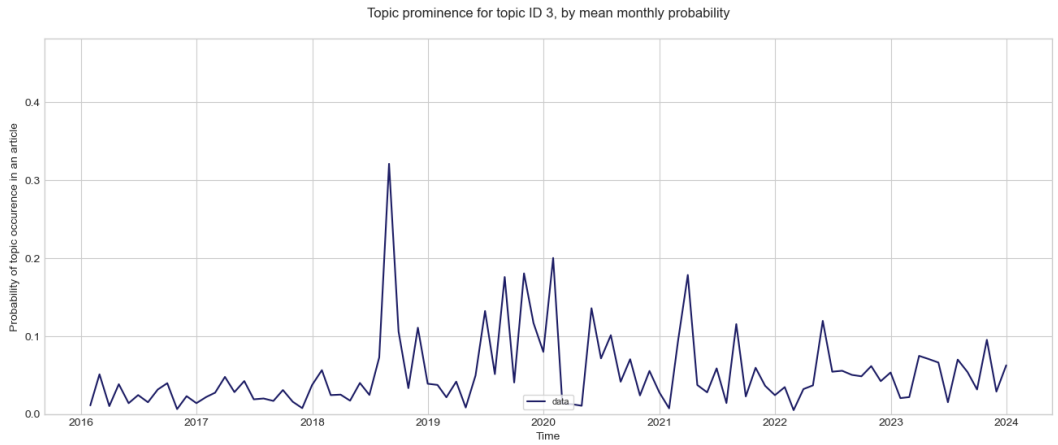

Figure 36. The dynamic of the Khachaturian sisters case topic, by monthly mean probability

Luis Rubiales case

The “Luis Rubiales case” topic is present, having more than 0.1 probability of being discussed, in 9.3% articles (as shown in Figure 8), being the most probable topic in 0.56% articles (as shown in Figure 9). The mean value of probability for this topic is 0.009 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.22 as shown in Figure 11.

The “Luis Rubiales case” topic includes such words as “футбол” and “рубиалес”, describing the Luis Rubiales sexual harassment case. The examples of the articles highly associated with this topic are: “Испанская прокуратура начала расследование в отношении домогавшегося футболистки главы RFEF” (Газета.Ru, 25.08.2023) and “Хави назвал скандал с поцелуем Рубиалеса печальной ситуацией” (Р-Спорт, 26.08.2023). As shown in Figure 37, the topic usage corresponds with the case occurrence and process.

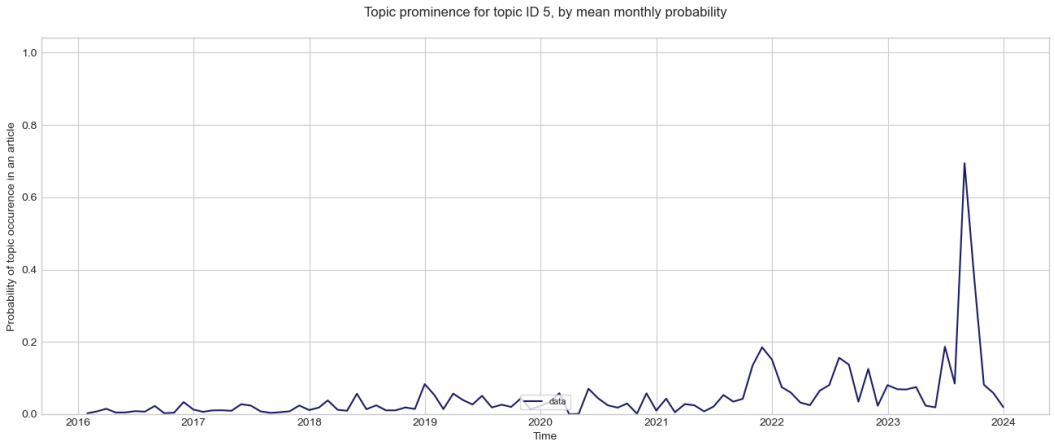

Figure 37. The dynamic of the Luis Rubiales case topic, by monthly mean probability

Harvey Weinstein case

The “Harvey Weinstein case” topic is present, having more than 0.1 probability of being discussed, in 10.34% articles (as shown in Figure 8), being the most probable topic in 0.74% articles (as shown in Figure 9). The mean value of probability for this topic is 0.011 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.32 as shown in Figure 11 – one of the highest values among all the topics. Therefore, it can be said that the Harvey Weinstein case, despite low prevalence (due to the episodic coverage most likely), was very prominent in terms of the audience attention.

The “Harvey Weinstein case” topic includes such words as “продюсер”, “домогательство” and “metoo”, describing the Harvey Weinstein harassment case. The examples of the articles highly associated with this topic are: “СМИ: почти 40 женщин обвинили американского режиссера в домогательствах” (РИА Новости, 22.10.2017) and “Харви Вайнштейна приговорили к 23 годам тюрьмы за сексуальное насилие” (Комсомольская правда, 11.03.2020). As shown in Figure 38, the topic usage corresponds with the case criminal process and is recurring even in the years after that.

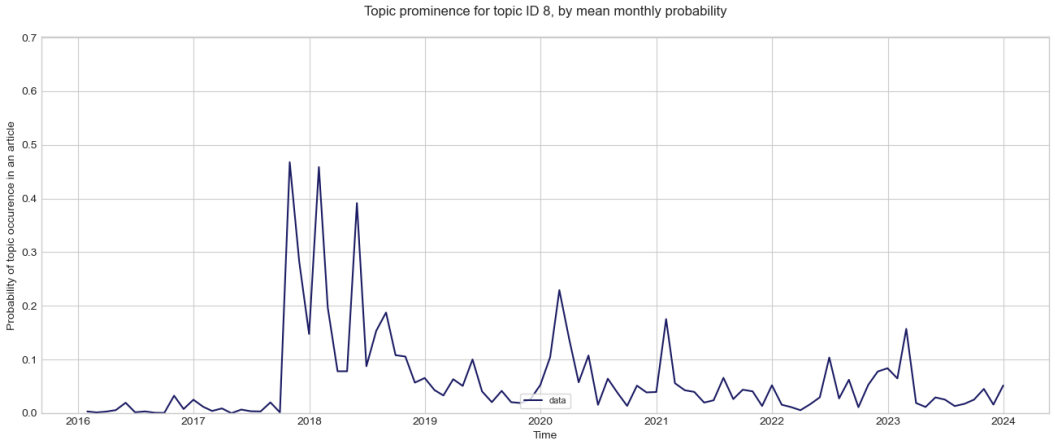

Figure 38. The dynamic of the Harvey Weinstein case topic, by monthly mean probability

Cristiano Ronaldo case

The “Cristiano Ronaldo case” topic is present, having more than 0.1 probability of being discussed, in 24.98% articles (as shown in Figure 8), being the most probable topic in 2.78% articles (as shown in Figure 9). The mean value of probability for this topic is 0.016 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.15 as shown in Figure 11. Therefore, being quite prevalent in terms of occurrences in articles on sexual violence in Russian news media, this case wasn’t popular with the audience in terms of visibility index.

The “Cristiano Ronaldo case” topic includes such words as “изнасилование”, “роналдо” and “заявить”, describing the Cristiano Ronaldo rape case. The examples of the articles highly associated with this topic are: “Американка обвинила Криштиану Роналду в изнасиловании – футболист считает, что женщина хочет прославиться” (Бинес Online, 29.09.2018) and “Суд отклонил иск модели, обвинившей Роналду в изнасиловании. Она требовала 56 млн фунтов компенсации” (Ведомости, 11.06.2022). As shown in Figure 39, the topic usage corresponds with the case criminal and media coverage process.

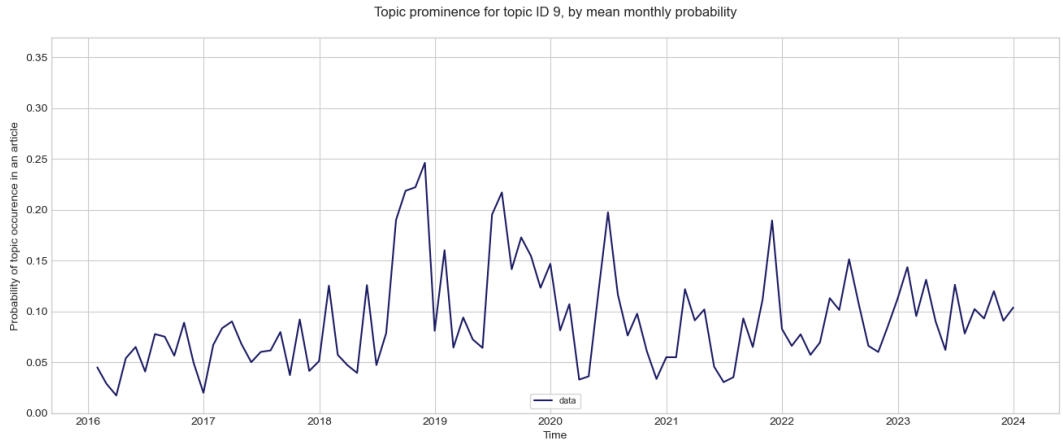

Figure 39. The dynamic of the Cristiano Ronaldo case topic, by monthly mean probability

Benjamin Mendy case

The “Benjamin Mendy case” topic is present, having more than 0.1 probability of being discussed, in 6% articles (as shown in Figure 8), being the most probable topic in 0.33% articles (as shown in Figure 9). The mean value of probability for this topic is 0.003 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.07 as shown in Figure 11.

The “Benjamin Mendy case” topic includes such words as “менди”, “манчестер” “защитник”, describing the Benjamin Mendy sexual violence case. The examples of the articles highly associated with this topic are: “Бенжамена Менди обвинили в 8-м изнасиловании. Всего ему вменяют 10 преступлений на почве секса” (Sports.ru, 01.06.2022) and “С игрока «Манчестер Сити» Менди сняли ряд обвинений в сексуальном насилии” (RT, 13.01.2023). As shown in Figure 40, the topic usage corresponds with the case criminal process.

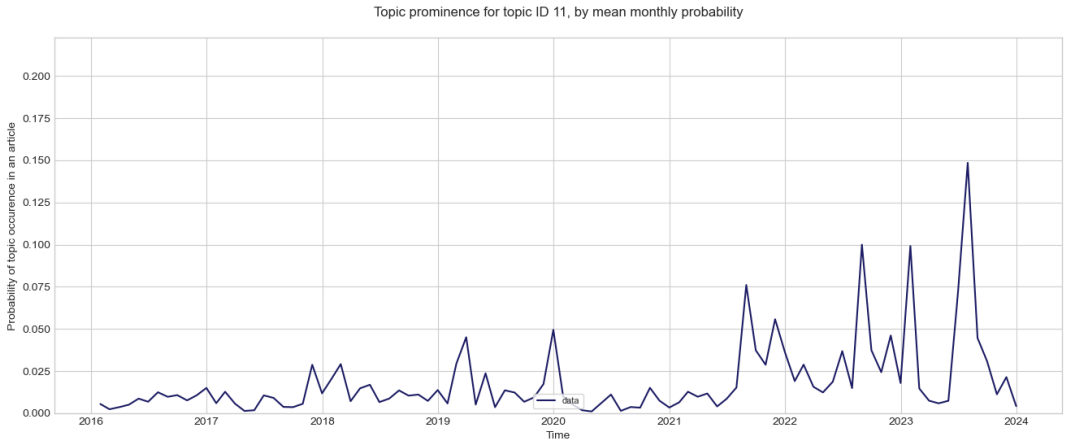

Figure 40. The dynamic of the Benjamin Mendy case topic, by monthly mean probability

Gérard Depardieu case

The “Gérard Depardieu case” topic is present, having more than 0.1 probability of being discussed, in 13.02% articles (as shown in Figure 8), being the most probable topic in 1% articles (as shown in Figure 9). The mean value of probability for this topic is 0.012 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.32 as shown in Figure 11 – one of the highest values among all the topics. Therefore, it can be said that the Gérard Depardieu case, despite low prevalence (due to the episodic coverage most likely), was very prominent in terms of the audience attention.

The “Gérard Depardieu case” topic includes such words as “актер”, “депардье” and “французский”, describing the Gérard Depardieu sexual harassment cases. The examples of the articles highly associated with this topic are: “Депардье обвинили в сексуальных домогательствах во Франции” (VSE42, 31.08.2018) and “Жерар Депардье домогался переводчицы в Северной Корее” (Газета.Ru, 08.12.2023). As shown in Figure 41, the topic usage corresponds with the cases of sexual violence regarding the actor.

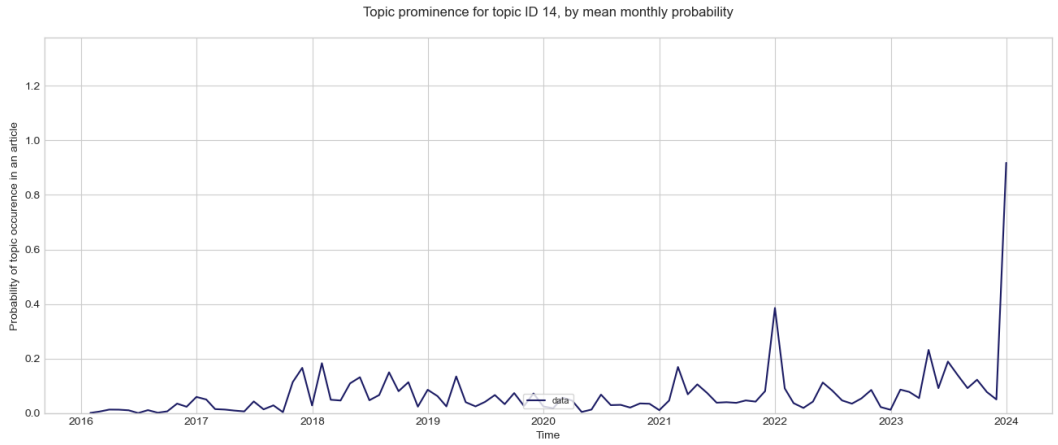

Figure 41. The dynamic of the Gérard Depardieu case topic, by monthly mean probability

Daniel Alves case

The “Daniel Alves case” topic is present, having more than 0.1 probability of being discussed, in 5.69% articles (as shown in Figure 8), being the most probable topic in 0.26% articles (as shown in Figure 9). The mean value of probability for this topic is 0.007 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.24 as shown in Figure 11.

The “Daniel Alves case” topic includes such words as “алвес” and “футболист”, describing the Daniel Alves sexual violence case. The examples of the articles highly associated with this topic are: “Marca: задержанному по обвинению в сексуальном насилии Алвесу может грозить до 12 лет” (Коммерсантъ. Новости информ. центра, 21.01.2023) and “СМИ: самого титулованного футболиста в истории Дани Алвеса обвиняют в сексуальном насилии” (ТАСС, 01.01.2023). As shown in Figure 42, the topic usage corresponds with the case criminal process.

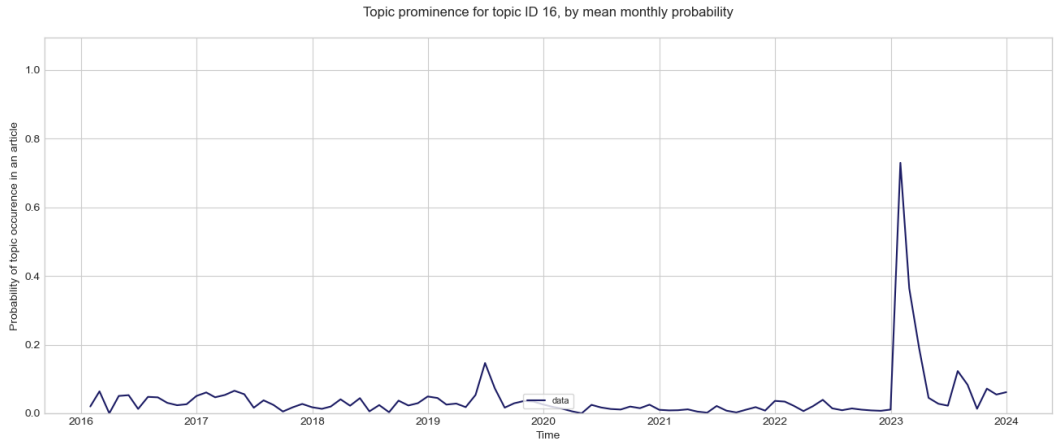

Figure 42. The dynamic of the Daniel Alves case topic, by monthly mean probability

Skopinsky maniac case

The “Skopinsky maniac case” topic is present, having more than 0.1 probability of being discussed, in 6.89% articles (as shown in Figure 8), being the most probable topic in 0.39% articles (as shown in Figure 9). The mean value of probability for this topic is 0.004 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.39 as shown in Figure 11 – one of the highest values among all the topics. Therefore, it can be said that the Skopinsky maniac case catched audience’s attention very highly and was relevant to the readers despite low prevalence (due to the episodic coverage most likely).

The “Skopinsky maniac case” topic includes such words as “маньяк”, “жертва” and “скопинский”, describing the Skopinsky maniac case. The examples of the articles highly associated with this topic are: “Жертва скопинского маньяка потребовала возбудить против него уголовное дело” (NEWS.ru, 30.03.2021) and “Жертве «скопинского маньяка» предоставили госзащиту” (РБК, 06.04.2021). As shown in Figure 43, the topic usage corresponds with the case criminal process.

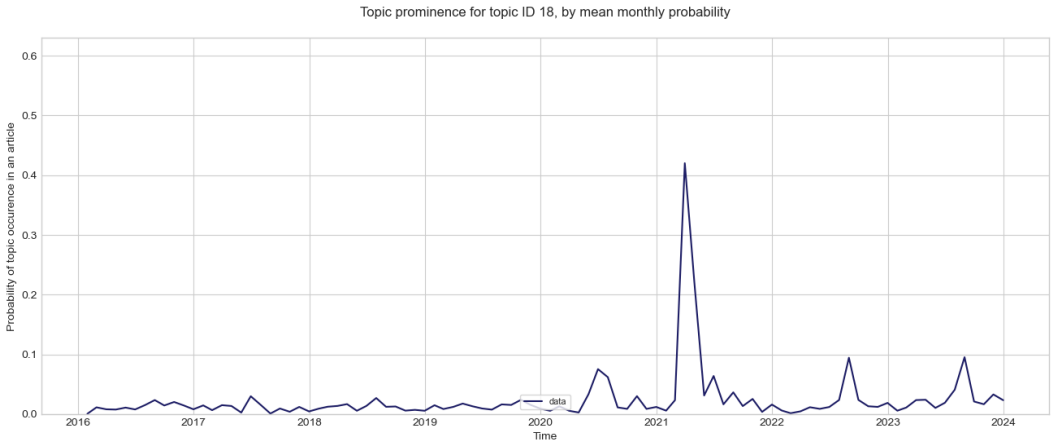

Figure 43. The dynamic of the Skopinsky maniac case topic, by monthly mean probability

Jeffrey Epstein case

The “Jeffrey Epstein case” topic is present, having more than 0.1 probability of being discussed, in 10.6% articles (as shown in Figure 8), being the most probable topic in 0.74% articles (as shown in Figure 9). The mean value of probability for this topic is 0.007 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.27 as shown in Figure 11.

The “Jeffrey Epstein case” topic includes such words as “эпштейн”, “актив” and “содержание”, describing the Jeffrey Epstein sexual violence case. The examples of the articles highly associated with this topic are: “Обнародовано секретное мировое соглашение между Эпштейном и его жертвой” (РИА Новости, 16.10.2021) and “Обвиняемый в секс-торговле миллиардер покончил с собой в тюрьме” (Взгляд.Ру, 10.08.2019). As shown in Figure 44, the topic usage corresponds with the case occurrence and process.

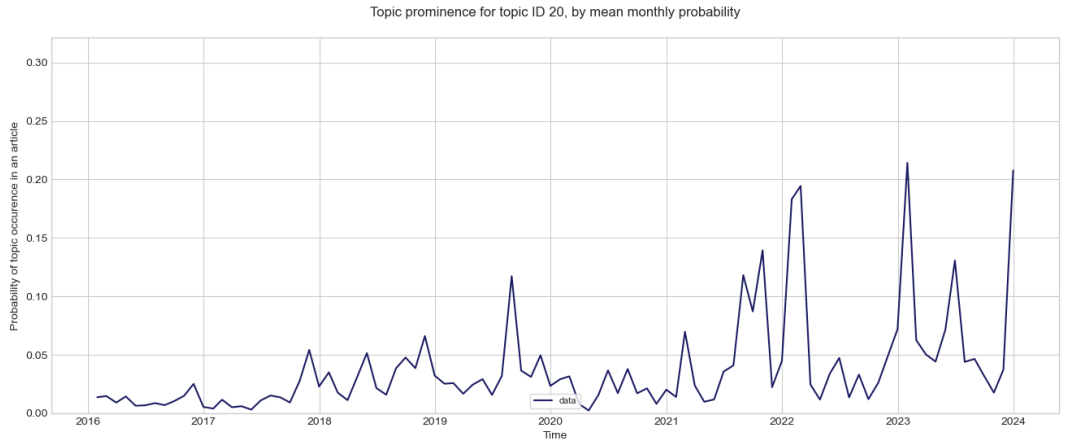

Figure 44. The dynamic of the Jeffrey Epstein case topic, by monthly mean probability

Victoria Marinova case

The “Victoria Marinova case” topic is present, having more than 0.1 probability of being discussed, in 6.74% articles (as shown in Figure 8), being the most probable topic in 0.37% articles (as shown in Figure 9). The mean value of probability for this topic is 0.004 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.36 as shown in Figure 11 – one of the highest values among all the topics. Therefore, it can be said that the Victoria Marinova case, despite low prevalence (due to the episodic coverage most likely), was very prominent in terms of the audience attention.

The “Victoria Marinova case” topic includes such words as “журналистка” and “убийство”, describing the rape and murder of the journalist Victoria Marinova. The examples of the articles highly associated with this topic are: “Глава Болгарии сделал заявление в связи с убийством журналистки Мариновой” (ИА Regnum, 15.10.2018) and “«Угроз никто не получал»: за что убили болгарскую журналистку” (Газета.Ру, 08.10.2018). As shown in Figure 45, the topic usage corresponds with the case process.

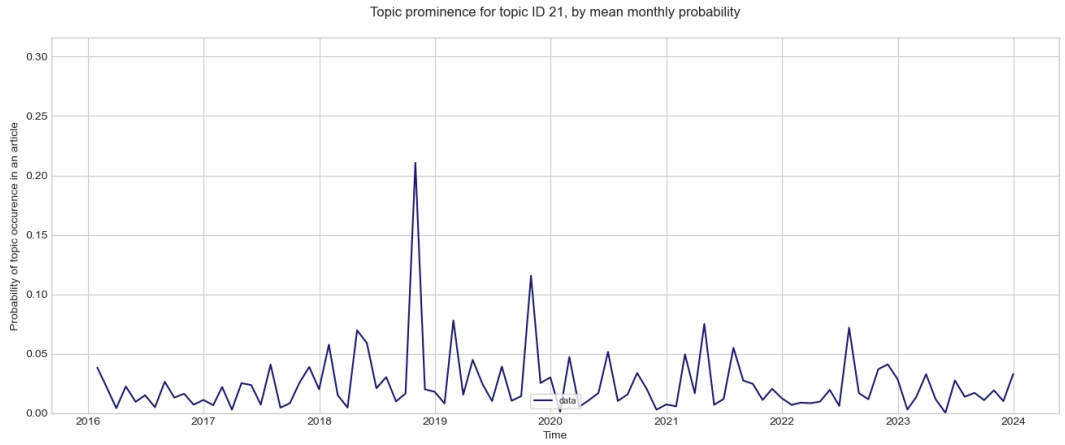

Figure 45. The dynamic of the Victoria Marinova case topic, by monthly mean probability

Rape in the Ufa police department case

The “Rape in the Ufa police department case” topic is present, having more than 0.1 probability of being discussed, in 13.99% articles (as shown in Figure 8), being the most probable topic in 1.03% articles (as shown in Figure 9). The mean value of probability for this topic is 0.021 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.25 as shown in Figure 11.

The “Rape in the Ufa police department case” topic includes such words as “изнасилование”, “полицеский”, “отдел” and “уфа”, describing the rape in the Ufa police department happened in 2018. The examples of the articles highly associated with this topic are: “Арестованы все трое подозреваемых в изнасиловании девушки-дознавателя в Уфе” (ИА Росбалт, 02.11.2018) and “Что известно о полицейских, подозреваемых в изнасиловании коллеги в Уфе?” (Аргументы и факты, 01.11.2018). As shown in Figure 46, the topic usage corresponds with the case criminal process.

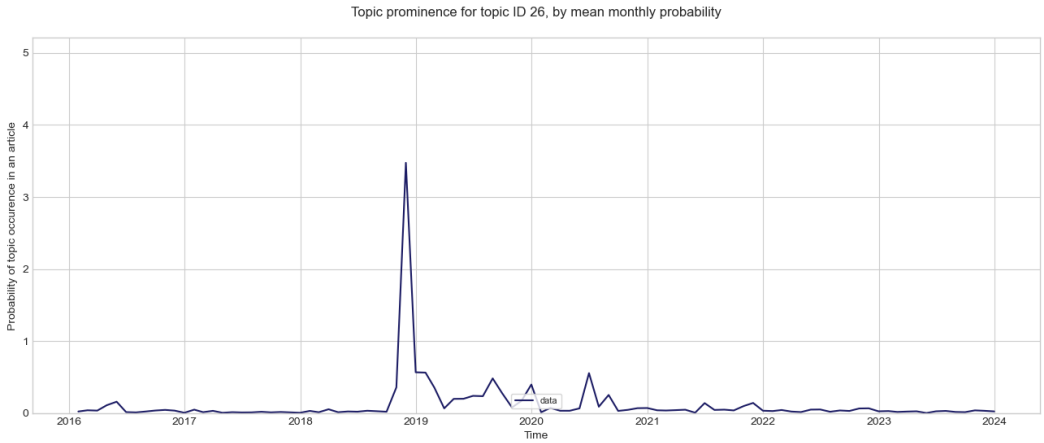

Figure 46. The dynamic of the Rape in the Ufa police department case topic, by monthly mean probability

Roman Polanski case

The “Roman Polanski case” topic is present, having more than 0.1 probability of being discussed, in 8.65% articles (as shown in Figure 8), being the most probable topic in 0.44% articles (as shown in Figure 9). The mean value of probability for this topic is 0.003 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.4 as shown in Figure 11 – the second highest value among all the topics. Therefore, it can be said that the Roman Polanski case, despite low prevalence (due to the episodic coverage most likely), was very prominent in terms of the audience attention.

The “Roman Polanski case” topic includes such words as “полански” and “женщина”, describing the Roman Polanski sexual violence allegations and case process. The examples of the articles highly associated with this topic are: “Известный телепроповедник гипнотизировал женщин и заманивал их в секс-секту. Его приговорили к тысячам лет тюрьмы” (Lenta.Ru, 28.11.2022) and “Обвинявшая Романа Полански в изнасиловании женщина попросила суд закрыть дело” (Комсомольская правда, 10.06.2017). As shown in Figure 47, the topic usage corresponds with the ongoing public media case and criminal process.

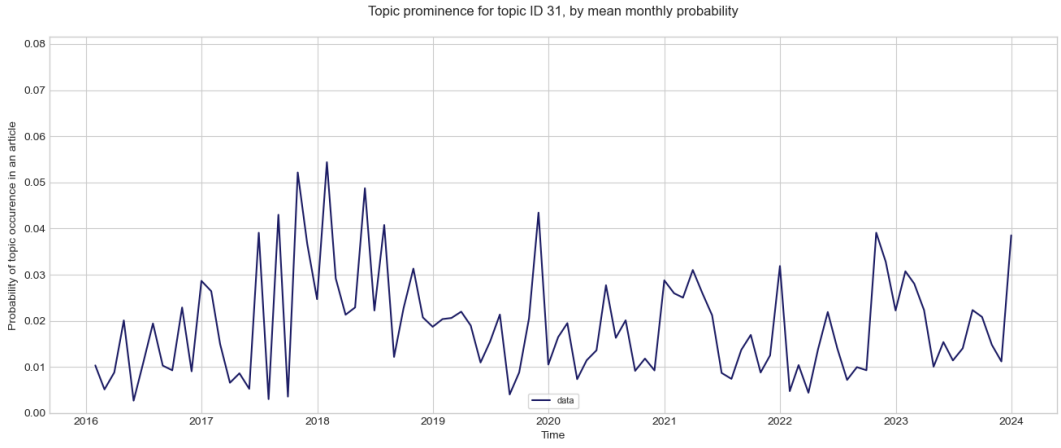

Figure 47. The dynamic of the Roman Polanski case topic, by monthly mean probability

Collectors case

The “Collectors case” topic is present, having more than 0.1 probability of being discussed, in 5.72% articles (as shown in Figure 8), being the most probable topic in 0.28% articles (as shown in Figure 9). The mean value of probability for this topic is 0.006 as shown in Figure 10. For the articles where this topic is the most probable, the mean value of the visibility index is 1.31 as shown in Figure 11.

The “Collectors case” topic includes such words as “коллектор”, “новосибирский”, “семья” and “нападение”, describing the rape of a woman by collectors that happened in 2016. The examples of the articles highly associated with this topic are: “В Новосибирской области коллекторы изнасиловали женщину и избили ее родных” (Аргументы и Факты, 05.04.2016) and “Компания «Деньги сразу» заявила о непричастности к нападению на должницу под Новосибирском” (Взгляд.Ру, 06.04.2016). As shown in Figure 48, the topic usage corresponds with the case criminal process.

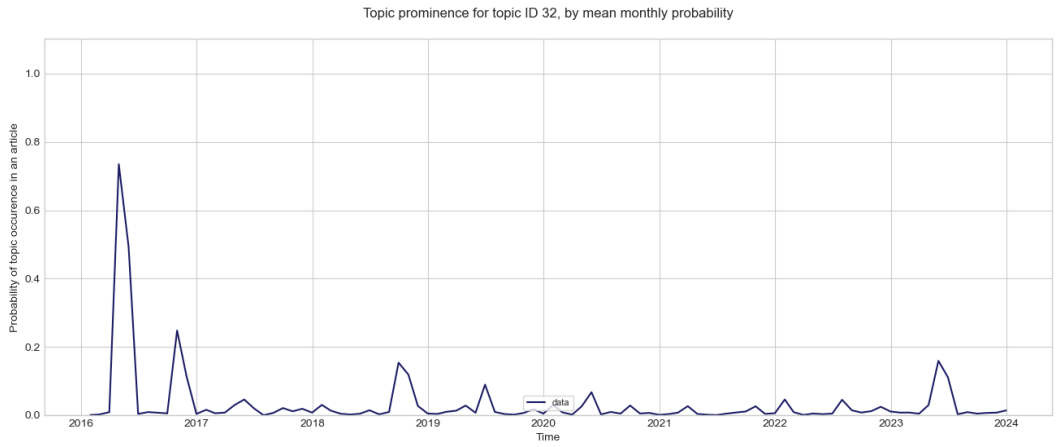

Figure 48. The dynamic of the Collectors case topic, by monthly mean probability

The themes that dominate the discussion on sexual violence

The first research objective of the study was *to identify and describe a set of dominant themes within the discussion on sexual violence against women in the Russian news media.* To solve this task, I identified 37 topics within the corpus of Russian news media articles dedicated to sexual violence, as well as explored them in terms of prevalence and visibility within the discussion. The careful interpretation of the topics retrieved through topic modeling and the assessment of their prevalence and visibility rates shows that the thematic structure of the discussion on sexual violence is rather diverse in terms of both the general themes that are found in the discussion and the variety of major public cases of sexual violence that were discussed in the Russian news media through years 2016 – 2023.

The most dominant themes within the discussion are threefold. First, the topics that are more prevalent in the corpus deal with the general narratives on sexual violence, such as descriptions of violence itself or describing the violence through the victim perspective. Second, there is a high prevalence of the topics that deal with the criminal cases against the perpetrators of violence. Third, the viewing of sexual violence as a social problem turned out to also be dominant in the discussion since the narrative about the legislative regulation of the problem is also widely present in the articles.

As for the initial assumptions regarding the first research objective, I theorized that *there will be themes within the thematic structure of the discussion on sexual violence against women that indicate the absence of agency of the victims of sexual violence (A1)* since the literature suggests that the victims are portrayed as unable to defend themselves (Hollander & Rodgers, 2014). This assumption cannot be supported based on the thematic structure obtained from the corpus. There are no topics in the discussion that would directly indicate the absence of agency of the victims of sexual violence or the absence of defense and self-protection, though it is important to note that it doesn't necessarily mean that the victims are not framed as helpless within the articles. The topic modeling method is based on the word co-occurrences, and therefore the absence of the topic that would indicate the absence of victim's agency may be based

on the fact that such framing patterns may simply not be based on a set of words used to describe a victim and include more complex narratives that may be detected only through the process of qualitative text analysis or manual coding of the articles.

Moreover, there are topics that, on the contrary, indicate the coverage of sexual violence in terms of victim's agency. The literature suggested that the victims cannot perform as active agents able to recover from violence or speak up, being damaged and traumatized forever (Alcoff & Gray, 1993). Some topics in the thematic structure of the discussion, such as the "Public statements" and "Victim perspective" topics, contradict the assumption (*A1)*. The "Public statements" topic explores the news on women's public allegations towards the perpetrators of sexual violence, therefore showing women as the central to the narrative of sexual violence and as those of demanding retributions for what happened to them. Nevertheless, this topic, in terms of framing, may be considered a part of issue-specific framing since it deals with a thematically narrow narrative and therefore is not applicable to the discussion as a whole, which is also supported by the relatively low rates of the topic prevalence. On the contrast, the "Victim perspective" topic deals with narratives about victim's experience in general and is concentrated solely on the victim's point of view. Therefore, the assumption cannot be supported based on the thematic structure of the discussion on sexual violence: the most prevalent and dominant topics within the discussion suggest that the narratives based on the victim's agency are present in the discussion on sexual violence against women.

The second assumption related to the thematic structure of the discussion on sexual violence was that *within the thematic structure of the discussion on sexual violence against women, newspapers will tend to attribute responsibility for the violence to individuals rather than the government or society*. This assumption may be supported by the high prevalence rates, as well as the wide variety of the topics exploring the details of criminal cases against the sexual violence perpetrators, indicating the individual causes consequences of the violence. However, in this case the coverage is

rather issue-specific and therefore it cannot be stated surely that descriptions of criminal trials suggest that the media blames the problem of violence itself on the criminals.

The other signs of the attribution of responsibility are shown by the prevalence of such topics as “Compensations for the victims” and “Problem regulation”. The topic that explores the narratives related to monetary compensations for the victims shows that the responsibility is placed upon different entities and not individuals exclusively. For example, within this topic, there are descriptions of the organizational entities compensating the victims that filed sexual harassment allegations which indicates the corresponding entities having the responsibility for the violence in terms of their role in the further retributions for the violence. As for the topic exploring the regulation of the problem, there are narratives about the legislative acts that government takes – or should take – in terms of dealing with the problem of sexual violence. This topic is significantly more prevalent than the other topics discussed above which allows me not to support the initial assumption and conclude that the media, along with attributing the responsibility of sexual violence to individuals, also tends to include the government into the narratives on who and what are responsible for the problem.

The third assumption regarding the thematic structure of the discussion was based on the sensationalism that is commonly present in the portrayals of sex crimes, as suggested in the literature (Benedict, 1993; Block, 2002; Hindes & Fileborn, 2020; Serisier, 2017). I assumed that *there will be a sensationalist rhetoric present in the Russian news media articles that cover sexual violence (A3).* There, the sensationalism occurs in many forms, such as the emotional coverage or the excessive detailing of the acts of sexual violence (Lemish, 2004), and with the latter, the corresponding rhetoric is certainly present in the articles on sexual violence in Russian mass media. This assumption may be supported by the “Descriptive narratives” topic which deals with the graphic descriptions of rape and other forms of sexual violence, addressing the detailed bodily features of violence. Since the “Descriptive narratives” topic is one of the most prevalent topics according to several prevalence indicators, it is fair to state that the

discussion on sexual violence is characterized by the sensationalist rhetoric in reporting cases of sexual violence as well as related crimes.

The description of the topics that are dominant in the discussion of sexual violence led to one interesting observation – in terms of the audience interest, the "Migrants as perpetrators" topic appeared to be one of the most visible ones despite the low levels of prevalence across the corpus. Literature suggests that media coverage of sexual violence tends to match many cultural stereotypes, including the Western pattern of portraying a black man as a common perpetrator of violence. Moreover, the common pattern of rape cases coverage is that the common perpetrator is a stranger rather than an acquaintance of the victim. Put in the Russian context, the high attention paid by the audience to the topic on the migrants being the perpetrators of violence, corresponds to these features of media construction of the sexual violence image, with the migrant being the ethnic and national stranger to the victim, distanced from the victim socially and fitting the common cultural stereotype.

Sensationalism in the articles covering sexual violence against women

The second research objective of this thesis was to explore the relationship between the tone of the articles and the themes that are prevalent in the discussion on sexual violence. The anticipated results of treating this objective, in a form of hypotheses, are the following: (1) the themes that explore the cases of sexual violence will be correlated to either positive or negative sentiment in the coverage of sexual violence, (2) the themes that explore sexual violence as a social problem will be correlated to the neutral sentiment in the coverage of sexual violence and (3) there will be no correlation between the other themes and level of sentiment in the articles covering sexual violence against women.

To assess how the topics prevalent in the discussion on sexual violence in Russian mass media are connected to the prevalence of different types of sentiment in articles, I use Pearson's linear correlation which is a statistical method for testing whether there is a statistically significant interchangeability of two interval, or quantitative variables. The

statistical hypothesis (H0) for all the tests of the aforementioned substantive hypotheses is that the tested *variables are not linearly correlated*; the alternative hypothesis (H1) is that *there is a linear correlation between the tested variables*.

For all the topic groups, Pearson correlation coefficients were computed separately for the three types of sentiment – positive, negative and neutral, – measured in three different ways each – by the mean value of sentence sentiment within the article, by the 3rd (75%) quantile value of sentence sentiment within the article and by the maximum value of sentence sentiment within the article. For each correlation coefficient, p-value (marked as "Sig." in the tables) was computed to evaluate the statistical significance of computed correlations to address the degree of randomness of test results obtained from the data. For the Pearson correlation coefficients that were statistically significant (at either 95% or 99% confidence levels), confidence intervals were computed to estimate the spread of true values of these coefficients.

General topic group

Table 16 shows the Pearson correlation coefficients for the General topic group and variables that indicate the prevalence of a positive sentiment in an article.

Table 16. Pearson correlation coefficient for the topics in the General topic group and positive sentiment in articles

		positive_mean	positive_75	positive_max
Rape and assault cases	Pearson Correlation	-0,011	-0,012	-0,005
	Sig. (2-tailed)	0,185	0,136	0,528
	N	15342	15342	15342
Victim perspective	Pearson Correlation	0,007	0,007	0,011
	Sig. (2-tailed)	0,361	0,371	0,193
	N	15342	15342	15342
Descriptive narratives	Pearson Correlation	0,003	0,004	0,011
	Sig. (2-tailed)	0,751	0,649	0,186
	N	15342	15342	15342
Public statements	Pearson Correlation	-0,001	-0,004	-0,002
	Sig. (2-tailed)	0,910	0,593	0,828
	N	15342	15342	15342
Sexual harassment	Pearson Correlation	0,000	0,001	0,007
	Sig. (2-tailed)	0,999	0,935	0,394
	N	15342	15342	15342
Compensations for victims	Pearson Correlation	0,004	0,002	0,007
	Sig. (2-tailed)	0,596	0,782	0,398
	N	15342	15342	15342

For all the topics in the General topic group – Rape and assault cases, Victim perspective, Descriptive narratives, Public statements, Sexual harassment, Compensations for victims – Pearson correlation coefficients are not statistically significant at any – neither 95% nor 99% – confidence levels. *The statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics in the General topic group and the degree of positive sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with positive emotional color.

As for correlations with the negative sentiment variables, the results are a bit more diverse (see Table 17 for correlation coefficients and p-value estimates).

Table 17. Pearson correlation coefficient for the topics in the General topic group and negative sentiment in articles

		negative_mean	negative_75	negative_max
Rape and assault cases	Pearson Correlation	-0,002	0,006	-0,007
	Sig. (2-tailed)	0,787	0,424	0,386
	N	15342	15342	15342
Victim perspective	Pearson Correlation	,018*	,020*	,026**
	Sig. (2-tailed)	0,024	0,011	0,001
	N	15342	15342	15342
Descriptive narratives	Pearson Correlation	0,015	,017*	0,008
	Sig. (2-tailed)	0,071	0,034	0,298
	N	15342	15342	15342
Public statements	Pearson Correlation	0,013	0,009	-0,003
	Sig. (2-tailed)	0,115	0,270	0,713
	N	15342	15342	15342
Sexual harassment	Pearson Correlation	0,006	0,004	0,006
	Sig. (2-tailed)	0,439	0,611	0,479
	N	15342	15342	15342
Compensations for victims	Pearson Correlation	0,014	0,007	0,009
	Sig. (2-tailed)	0,087	0,369	0,265
	N	15342	15342	15342
	Sig. (2-tailed)	0	0	
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

The probability of Victim perspective topic is correlated with all the variables measuring the degree of negative sentiment in the articles in a positive direction yet the correlation coefficients themselves, though statistically significant on either 95% or

99% confidence levels, are quite low. The correlation coefficient for the probability of Victim perspective topic and negative sentiment measured by mean sentence value is 0.018 (p-value < 0.05), for the probability of Victim perspective topic and negative sentiment measured by 75% quartile sentence value is 0.02 (p-value < 0.05), for the probability of Victim perspective topic and negative sentiment measured by mean sentence value is 0.026 (p-value < 0.01). There, it may be said that the degree of negative emotions in the coverage of sexual violence is higher when the theme considering victim stories and perspectives is involved in the narrative. Moreover, the degree of negative sentiment measured by the 75% quartile sentence value is higher in the articles where the probability of coverage through the descriptive narrative is higher (corr = 0.017, p-value < 0.05). There, *the statistical hypothesis H0 is not supported* for the aforementioned variables indicating that there is statistically significant linear correlation present within the variables interaction.

Table 18. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)	
			Lower	Upper
sup0 - negative_mean	0,018	0,024	0,002	0,034
sup0 - negative_75	0,02	0,011	0,005	0,036
sup0 - negative_max	0,026	0,001	0,01	0,042
sup1 - negative_75	0,017	0,034	0,001	0,033

Estimation is based on Fisher's r-to-z transformation.

As Table 18 shows, for the probability of Victim perspective topic and negative sentiment measured by mean sentence value, by 75% quartile sentence value and by maximum sentence value, the true value of Pearson correlation lies within [0,002; 0,034], [0,005; 0,036] and [0,01; 0,042] intervals respectively.

As for the other topics in the General topic group, as seen in Table 17, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Rape and assault cases, Public statements, Sexual harassment and Compensations for victims and the degree of negative sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these

topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a negative emotional color.

Table 19. Pearson correlation coefficient for the topics in the General topic group and neutral sentiment in articles

		neutral_mean	neutral_75	neutral_max
Rape and assault cases	Pearson Correlation	-0,007	0,005	0,011
	Sig. (2-tailed)	0,390	0,559	0,162
	N	15342	15342	15342
Victim perspective	Pearson Correlation	-0,011	-0,007	0,000
	Sig. (2-tailed)	0,155	0,391	0,952
	N	15342	15342	15342
Descriptive narratives	Pearson Correlation	-0,015	-0,007	0,000
	Sig. (2-tailed)	0,064	0,378	0,986
	N	15342	15342	15342
Public statements	Pearson Correlation	-0,003	-0,006	-0,008
	Sig. (2-tailed)	0,666	0,466	0,328
	N	15342	15342	15342
Sexual harassment	Pearson Correlation	-0,010	-0,012	-0,012
	Sig. (2-tailed)	0,220	0,152	0,142
	N	15342	15342	15342
Compensations for victims	Pearson Correlation	-0,008	-0,009	-0,014
	Sig. (2-tailed)	0,316	0,270	0,075
	N	15342	15342	15342

The results of testing the hypothesis for the neutral sentiment are similar here for the ones in testing the hypothesis for positive sentiment. For all the topics in the General topic group – Rape and assault cases, Victim perspective, Descriptive narratives, Public statements, Sexual harassment, Compensations for victims – Pearson correlation coefficients are not statistically significant at any – neither 95% nor 99% – confidence levels. *The statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics in the General topic group and the degree of neutral sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with neutral emotions.

Social problem topic group

Table 20 shows the Pearson correlation coefficients for the topics in the Social problem topic group and variables that indicate the prevalence of a positive sentiment in an article.

Table 20. Pearson correlation coefficient for the topics in the Social problem topic group and positive sentiment in articles

		positive_mean	positive_75	positive_max
Problem regulation	Pearson Correlation	0,006	0,004	0,003
	Sig. (2-tailed)	0,450	0,615	0,667
	N	15342	15342	15342
Reports and statistics	Pearson Correlation	,020*	,016*	0,008
	Sig. (2-tailed)	0,014	0,044	0,319
	N	15342	15342	15342
Fighting SV - organizations and funding	Pearson Correlation	0,012	0,011	-0,005
	Sig. (2-tailed)	0,139	0,173	0,511
	N	15342	15342	15342
Nobel prize for fighting SV	Pearson Correlation	0,000	-0,001	-0,004
	Sig. (2-tailed)	0,969	0,926	0,640
	N	15342	15342	15342

*. Correlation is significant at the 0.05 level (2-tailed).

As seen in Table 20, one of the topics shows significant results in terms of being correlated to the positive emotional coverage of sexual violence in the articles. The probability of Reports and statistics topic is higher in the articles where the positive sentiment of the sentences is higher, measured either by the mean value of the sentences sentiment (corr = 0.2, p-value < 0.05) or by the 75 percentile value of the sentences sentiment (corr = 0.016, p-value < 0.05). There, *the statistical hypothesis H0 is not supported* for the aforementioned pairs of variables indicating that there is statistically significant linear correlation present within the variables interchangeability. The Pearson correlations themselves, directed positively, are quite weak though.

Table 21. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)	
			Lower	Upper
@6 - positive_mean	0,02	0,014	0,004	0,036
@6 - positive_75	0,016	0,044	0	0,032

Estimation is based on Fisher's r-to-z transformation.

As Table 21 shows, for the probability of the Reports and statistics topic and positive sentiment measured by mean sentence value and by 75% percentile sentence value, the true value of Pearson correlation lies within [0,004; 0,036] and [0; 0,032] intervals respectively.

As for the other topics in the Social problem topic group, as seen in Table 20, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Problem regulation, Fighting SV - organizations and funding and Nobel prize for fighting SV and the degree of positive sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a positive emotional color.

As for the correlations of topics from the Social problem group with the variables indicating negative sentiment, there is one significantly correlating pair of values: the probability of the topic Fighting SV - organizations and funding and negative sentiment variable measured by the maximum value of sentence sentiment (corr = -0.017, p-value < 0.05). The Pearson coefficient is quite weak itself, though it shows the negative direction of variables interchangeability. *The statistical hypothesis H0 is not supported* for the aforementioned pair of variables indicating that this pair is linearly correlated.

Table 22. Pearson correlation coefficient for the topics in the Social problem topic group and negative sentiment in articles

		negative_mean	negative_75	negative_max
Problem regulation	Pearson Correlation	-0,001	-0,007	0,000
	Sig. (2-tailed)	0,936	0,360	0,952
	N	15342	15342	15342
Reports and statistics	Pearson Correlation	-0,003	-0,007	-0,002
	Sig. (2-tailed)	0,707	0,413	0,826
	N	15342	15342	15342
Fighting SV - organizations and funding	Pearson Correlation	-0,006	-0,013	-,017*
	Sig. (2-tailed)	0,459	0,104	0,040
	N	15342	15342	15342
Nobel prize for fighting SV	Pearson Correlation	-0,002	-0,004	-0,005
	Sig. (2-tailed)	0,768	0,607	0,551
	N	15342	15342	15342

*. Correlation is significant at the 0.05 level (2-tailed).

Table 22 showed that the negative sentiment measured by the maximum sentence value is lower for the articles where the prevalence of the topic Fighting SV - organizations and funding is higher. The true value of the correlation for this pair, as shown in Table 23, lies in the [-0,032; -0,0008] interval with the 0.95 probability.

Table 23. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)	
			Lower	Upper
@17 - negative_max	-0,016610906	0,039643147	-0,032426275	-0,000787221

Estimation is based on Fisher's r-to-z transformation.

As for the other topics in the Social problem topic group, as seen in Table 22, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Problem regulation, Reports and statistics and Nobel prize for fighting SV and the degree of negative sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a negative emotional color.

As displayed in Table 24, the probability of a Problem regulation topic in an article is slightly negatively correlated to the neutral tone of the publication, with sentiment measured by the 75 percentile value of the sentence sentiment (corr = -0.017, p-value < 0.05). The probability of the Fighting SV - organizations and funding topic is also slightly negatively correlated to the degree of neutrality in an article, yet measured by the maximum value of the sentence sentiment (corr = -0.019, p-value < 0.05). Therefore, it can be stated that *the statistical hypothesis H0 is not supported* for the aforementioned pairs of variables indicating that these pairs are linearly correlated.

Table 24. Pearson correlation coefficient for the topics in the Social problem topic group and neutral sentiment in articles

		neutral_mean	neutral_75	neutral_max
Problem regulation	Pearson Correlation	-0,004	-,017*	-0,012
	Sig. (2-tailed)	0,606	0,041	0,128
	N	15342	15342	15342
Reports and statistics	Pearson Correlation	-0,004	-0,010	-0,013
	Sig. (2-tailed)	0,646	0,217	0,118
	N	15342	15342	15342
Fighting SV - organizations and funding	Pearson Correlation	0,000	-0,011	-,019*
	Sig. (2-tailed)	0,964	0,184	0,018
	N	15342	15342	15342
Nobel prize for fighting SV	Pearson Correlation	0,000	-0,006	-0,001
	Sig. (2-tailed)	0,959	0,436	0,901
	N	15342	15342	15342

*. Correlation is significant at the 0.05 level (2-tailed).

Table 24 showed that the neutral sentiment measured by the 75 percentile sentence value is lower for the articles where the prevalence of the topic Problem regulation and funding is higher and also for the articles where the prevalence of the topic Fighting SV - organizations and funding is higher. The true value of the correlations for these pairs, as shown in Table 25, lie in the [-0,032; -0,0007] and the [-0,035; -0,003] 95% confidence intervals, respectively.

Table 25. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
sup2 - neutral_75	-0,016505366	0,04091662	-0,032320816	-0,000681652
@17 - neutral_max	-0,019103194	0,017971792	-0,034916532	-0,003280294

a. Estimation is based on Fisher's r-to-z transformation.

As for the other topics in the Social problem topic group, as seen in Table 24, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Reports and statistics and Nobel prize for fighting SV and the degree of neutral sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a neutral emotional color.

Criminal cases & process

Within the Criminal cases & process topic class two topics show statistically significant levels of interchangeability with the positive emotional color of the articles. Table 26 shows that the prevalence of the positive emotional coverage of sexual violence is lower when the probability of an article being about Details of criminal cases processes is higher for two ways of measuring the positive sentiment – through the mean value of sentence sentiment (corr = -0.026, p-value < 0.01) and through the 75 percentile value of sentence sentiment (corr = -0.023, p-value < 0.01). For the articles where the Perpertrators court sentences topic that includes murder cases is more probable, the positive sentiment of coverage is less likely within all the sentiment variables, regardless of having been measured through the mean value of sentence sentiment (corr = -0.03, p-value < 0.01), through the 75 percentile value of sentence sentiment (corr = -0.027, p-value < 0.01) or through the maximumvalue of sentence sentiment (corr = -0.024, p-value < 0.01). *The H0 is not supported* for all the mentioned pairs of variables, making the probabilities of Details of criminal cases processes and Perpertrators court sentences murder cases topics negatively linearly, though weakly, correlated with the corresponding degrees of positive sentiment, on the 99% confidence level.

Table 26. Pearson correlation coefficient for the topics in the Criminal cases & process topic group and positive sentiment in articles

		positive_mean	positive_75	positive_max
Perpetrators detention	Pearson Correlation	-0,014	-0,013	-0,014
	Sig. (2-tailed)	0,085	0,115	0,078
	N	15342	15342	15342
Perpertrators crime sentences	Pearson Correlation	-0,010	-0,011	-0,006
	Sig. (2-tailed)	0,198	0,155	0,468
	N	15342	15342	15342
Details of criminal cases processes	Pearson Correlation	-,026**	-,023**	-0,013
	Sig. (2-tailed)	0,001	0,004	0,112
	N	15342	15342	15342
Perpertrators court sentencing	Pearson Correlation	0,011	0,013	0,013
	Sig. (2-tailed)	0,159	0,115	0,110
	N	15342	15342	15342
Perpertrators court sentences murder cases	Pearson Correlation	-,030**	-,027**	-,024**
	Sig. (2-tailed)	0,000	0,001	0,003
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

As Table 27 shows, the true value of the Pearson correlation coefficient for the Details of criminal cases processes topic probability and variables for positive sentiment variables lies within the [-0,041; -0,01] and the [-0,039; -0,007] 95% confidence intervals respectively for both variables. The true value of the Pearson correlation coefficient between the Perpertrators court sentences murder cases topic probability and the three variables indicating the degree of positive emotion of the coverage lies within the [-0,046; -0,014], the [-0,043; -0,012] and the [-0,04; -0, 0,008] 95% confidence intervals for the variables of positive sentiment measured by the mean, 75 percentile and maximum values of sentence sentiment, respectively.

Table 27. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@28 - positive_mean	-0,025642374	0,00149115	-0,041449448	-0,009822468
@28 - positive_75	-0,023281663	0,003927979	-0,039091153	-0,007460519
@30 - positive_mean	-0,029816976	0,000220984	-0,045619344	-0,013999689
@30 - positive_75	-0,027472521	0,000666042	-0,043277599	-0,011653695
@30 - positive_max	-0,023952187	0,003007508	-0,039761009	-0,008131377

a. Estimation is based on Fisher's r-to-z transformation.

As for the other topics in the Criminal cases & process topic group, as seen in Table 26, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Perpetrators detention topic, Perpertrators crime sentences topic, Perpertrators court sentencing topic and the degree of positive sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a positive emotional color.

Table 28. Pearson correlation coefficient for the topics in the Criminal cases & process topic group and negative sentiment in articles

		negative_mean	negative_75	negative_max
Perpetrators detention	Pearson Correlation	-0,013	-0,014	-0,014
	Sig. (2-tailed)	0,119	0,084	0,086
	N	15342	15342	15342
Perpertrators crime sentences	Pearson Correlation	-0,005	-0,005	-0,002
	Sig. (2-tailed)	0,513	0,514	0,840
	N	15342	15342	15342
Details of criminal cases processes	Pearson Correlation	-0,001	0,003	-0,001
	Sig. (2-tailed)	0,881	0,740	0,934
	N	15342	15342	15342
Perpertrators court sentencing	Pearson Correlation	0,008	0,008	0,009
	Sig. (2-tailed)	0,326	0,337	0,278
	N	15342	15342	15342
Perpertrators court sentences murder cases	Pearson Correlation	-0,012	-0,001	-0,011
	Sig. (2-tailed)	0,138	0,902	0,183
	N	15342	15342	15342

Moving to the variables for negative sentiment, as displayed in Table 28, or all the topics in the Criminal cases & process topic group – Perpetrators detention, Perpertrators crime sentences, Details of criminal cases processes, Perpertrators court sentencing and Perpertrators court sentences murder cases – Pearson correlation coefficients are not statistically significant at any – neither 95% nor 99% – confidence levels. *The statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics in the Criminal cases & process topic group and the degree of negative sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with negative emotional color.

Table 29. Pearson correlation coefficient for the topics in the Criminal cases & process topic group and neutral sentiment in articles

		neutral_mean	neutral_75	neutral_max
Perpetrators detention	Pearson Correlation	0,015	0,004	-0,012
	Sig. (2-tailed)	0,072	0,622	0,149
	N	15342	15342	15342
Perpertrators crime sentences	Pearson Correlation	0,006	0,009	0,015
	Sig. (2-tailed)	0,477	0,280	0,055
	N	15342	15342	15342
Details of criminal cases processes	Pearson Correlation	,018*	,017*	0,013
	Sig. (2-tailed)	0,025	0,038	0,110
	N	15342	15342	15342
Perpertrators court sentencing	Pearson Correlation	-0,002	0,004	0,003
	Sig. (2-tailed)	0,765	0,636	0,720
	N	15342	15342	15342
Perpertrators court sentences murder cases	Pearson Correlation	,025**	,026**	,016*
	Sig. (2-tailed)	0,002	0,002	0,041
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Within the Criminal cases & process topic class two topics show statistically significant levels of interchangeability with the neutral sentiment of the articles. Table 29 shows that the prevalence of the neutral emotional coverage of sexual violence is higher when the probability of an article being about Details of criminal cases processes is higher for two ways of measuring the positive sentiment – through the mean value of sentence sentiment (corr = 0.018, p-value < 0.05) and through the 75 percentile value of sentence sentiment (corr = -0.017, p-value < 0.05). For the articles where the Perpertrators court sentences topic that includes murder cases is more probable, the neutral sentiment of coverage is more likely within all the sentiment variables, regardless of having been measured through the mean value of sentence sentiment (corr = 0.025, p-value < 0.01), through the 75 percentile value of sentence sentiment (corr = 0.026, p-value < 0.01) or through the maximum value of sentence sentiment (corr = -0.016, p-value < 0.05). *The statistical hypothesis H0 is not supported* for all the mentioned pairs of variables, making the probabilities of Details of criminal cases processes and Perpertrators court sentences murder cases topics positively linearly,

though weakly, correlated with the corresponding degrees of neutral sentiment, on the 95% or 99% confidence level, depending on the variable.

Table 30. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@28 - neutral_mean	0,018	0,025	0,002	0,034
@28 - neutral_75	0,017	0,038	0,001	0,033
@30 - neutral_mean	0,025	0,002	0,009	0,041
@30 - neutral_75	0,026	0,002	0,010	0,041
@30 - neutral_max	0,016	0,041	0,001	0,032

a. Estimation is based on Fisher's r-to-z transformation.

As Table 30 shows, the true value of the Pearson correlation coefficient for the Details of criminal cases processes topic probability and variables for neutral sentiment variables lies within the [0,002; 0,034] and the [0,001; 0,033] 95% confidence intervals respectively for both variables. The true value of the Pearson correlation coefficient between the Perpertrators court sentences murder cases topic probability and the three variables indicating the degree of positive emotion of the coverage lies within the [0,009; 0,041], the [0,010; 0,041] and the [0,001; 0,032] 95% confidence intervals for the variables of neutral sentiment measured by the mean, 75 percentile and maximum values of sentence sentiment, respectively.

As for the other topics in the Criminal cases & process topic group, as seen in Table 29, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Perpetrators detention topic, Perpertrators crime sentences topic, Perpertrators court sentencing topic and the degree of neutral sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a neutral sentiment.

Perpetrator types

Within the Perpetrator types topic group displayed in Table 31, the probability of Musicians as perpetrators topic in the article is higher when the prevalence of the

positive sentiment in the articles is higher, for positive sentiment measured by mean value of sentence sentiment (corr = 0.024, p-value < 0.01) and by 75 percentile value of sentence sentiment (corr = 0.027, p-value < 0.01). *The statistical hypothesis H0 is not supported* on the 99% confidence level, meaning that there is a statistically significant correlation between the variables.

Table 31. Pearson correlation coefficient for the topics in the Perpetrator types topic group and positive sentiment in articles

		positive_mean	positive_75	positive_max
Migrants as perpetrators	Pearson Correlation	0,004	0,001	0,014
	Sig. (2-tailed)	0,594	0,945	0,078
	N	15342	15342	15342
Priests as perpetrators	Pearson Correlation	-0,002	0,005	0,008
	Sig. (2-tailed)	0,834	0,528	0,332
	N	15342	15342	15342
Musicians as perpetrators	Pearson Correlation	,024**	,027**	0,013
	Sig. (2-tailed)	0,003	0,001	0,095
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

The true values of the Pearson correlation coefficient for the pairs of variables described above lie within [0,008; 0,039] and [0,011; 0,042] respectively for these pairs of variables as shown in Table 32.

For the other variables in the Perpetrator types group displayed in Table 31, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Migrants as perpetrators and Priests as perpetrators, and the degree of positive sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a positive sentiment.

Table 32. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@12 - positive_mean	0,024	0,003	0,008	0,039
@12 - positive_75	0,027	0,001	0,011	0,042

a. Estimation is based on Fisher's r-to-z transformation.

As for the negative sentiment, within the Perpetrator types topic group displayed in Table 33, the probability of Migrants as perpetrators topic in the article is higher when the prevalence of the negative sentiment in the articles is higher, for negative sentiment measured by mean value of sentence sentiment (corr = 0.02, p-value < 0.51) and by 75 percentile value of sentence sentiment (corr = 0.021, p-value < 0.01). *The statistical hypothesis H0 is not supported* on the 99% confidence level, meaning that there is a statistically significant correlation between the variables.

Table 33. Pearson correlation coefficient for the topics in the Perpetrator types topic group and negative sentiment in articles

		negative_mean	negative_75	negative_max
Migrants as perpetrators	Pearson Correlation	,020*	,021**	0,010
	Sig. (2-tailed)	0,013	0,009	0,209
	N	15342	15342	15342
Priests as perpetrators	Pearson Correlation	0,005	0,001	0,011
	Sig. (2-tailed)	0,511	0,946	0,169
	N	15342	15342	15342
Musicians as perpetrators	Pearson Correlation	0,008	0,007	0,008
	Sig. (2-tailed)	0,340	0,370	0,326
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

For the other variables in the Perpetrator types group displayed in Table 33, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Priests as perpetrators and Musicians as perpetrators, and the degree of negative sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a negative sentiment.

Table 34. Pearson correlation coefficient for the topics in the Perpetrator types topic group and neutral sentiment in articles

		neutral_mean	neutral_75	neutral_max
Migrants as perpetrators	Pearson Correlation	-0,009	-0,009	0,000
	Sig. (2-tailed)	0,255	0,292	0,977
	N	15342	15342	15342
Priests as perpetrators	Pearson Correlation	-0,006	-0,004	-0,007
	Sig. (2-tailed)	0,494	0,624	0,399
	N	15342	15342	15342
Musicians as perpetrators	Pearson Correlation	-0,015	-0,009	-0,005
	Sig. (2-tailed)	0,069	0,270	0,542
	N	15342	15342	15342

Finally, for all the topics in the Perpetrator types topic group – Migrants as perpetrators, Priests as perpetrators and Musicians as perpetrators – Pearson correlation coefficients are not statistically significant at any – neither 95% nor 99% – confidence levels. *The statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics in the Perpetrator types topic group and the degree of neutral sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a neutral sentiment.

Spheres of violence

Within the Spheres of violence, as Table 35 shows, there are three low Pearson linear correlation coefficients that are statistically significant: between Violence in hockey topic and two positive variables, measured as mean (corr = 0.02, p-value < 0.05) and as 75 percentile (corr = 0.021, p-value < 0.01), and also between the SV during the war in Ukraine topic and a positive sentiment measured as maximum (corr = -0.016, p-value < 0.05). *The statistical hypothesis H0 is not supported*, meaning that there is a statistically significant correlation between the variables.

Table 35. Pearson correlation coefficient for the topics in the Spheres of violence topic group and positive sentiment in articles

		positive_mean	positive_75	positive_max
Royal family scandals	Pearson Correlation	-0,009	-0,009	-0,005
	Sig. (2-tailed)	0,270	0,263	0,543
	N	15342	15342	15342
Violence in hockey	Pearson Correlation	,020*	,021**	0,004
	Sig. (2-tailed)	0,016	0,009	0,635
	N	15342	15342	15342
Violence in figure skating	Pearson Correlation	-0,003	-0,002	-0,007
	Sig. (2-tailed)	0,725	0,801	0,386
	N	15342	15342	15342
The convention of womens' rights	Pearson Correlation	-0,010	-0,014	0,002
	Sig. (2-tailed)	0,200	0,093	0,769
	N	15342	15342	15342
SV during the war in Ukraine	Pearson Correlation	0,001	0,001	-,016*
	Sig. (2-tailed)	0,873	0,864	0,050
	N	15342	15342	15342
Japanese program for women in sexual slavery	Pearson Correlation	0,003	0,005	0,007
	Sig. (2-tailed)	0,738	0,504	0,397
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The true values of the Pearson correlation coefficient lie within [0,004; 0,035] and [0,005; 0,037] 95% confidence intervals (see Table 36) for the pairs of the Violence in hockey topic and measurements of the positive sentimens, and within the [-0.032; 0] 95% confidence interval for the SV during the war in Ukraine topic and the positive sentiment variable.

For the other variables in the Spheres of violence group displayed in Table 35, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Royal family scandals, Violence in figure skating, The convention of womens' rights and Japanese program for women in sexual slavery, and the degree of positive sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a positive sentiment.

Table 36. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@2 - positive_mean	0,020	0,016	0,004	0,035
@2 - positive_75	0,021	0,009	0,005	0,037
@24 - positive_max	-0,016	0,050	-0,032	0,000

a. Estimation is based on Fisher's r-to-z transformation.

As for the negative sentiments, as Table 37 shows, there are two low Pearson linear correlation coefficients that are statistically significant: between Violence in hockey and negative sentiment measured by the 75 percentile value of sentence sentiment (corr = -0.02, p-value < 0.05) and between the The convention of womens' rights topic and a positive sentiment measured as maximum (corr = 0.016, p-value < 0.05). *The statistical hypothesis H0 is not supported* on the 95% confidence level, meaning that there is a statistically significant correlation between the variables.

Table 37. Pearson correlation coefficient for the topics in the Spheres of violence topic group and negative sentiment in articles

		negative_mean	negative_75	negative_max
Royal family scandals	Pearson Correlation	0,006	0,011	-0,001
	Sig. (2-tailed)	0,431	0,161	0,930
	N	15342	15342	15342
Violence in hockey	Pearson Correlation	-0,014	-,020*	0,002
	Sig. (2-tailed)	0,074	0,015	0,794
	N	15342	15342	15342
Violence in figure skating	Pearson Correlation	0,013	0,013	0,010
	Sig. (2-tailed)	0,114	0,119	0,207
	N	15342	15342	15342
The convention of womens' rights	Pearson Correlation	0,002	0,003	,016*
	Sig. (2-tailed)	0,767	0,698	0,047
	N	15342	15342	15342
SV during the war in Ukraine	Pearson Correlation	0,008	0,001	-0,003
	Sig. (2-tailed)	0,307	0,873	0,690
	N	15342	15342	15342
Japanese program for women in sexual slavery	Pearson Correlation	-0,005	-0,007	0,002
	Sig. (2-tailed)	0,555	0,400	0,775
	N	15342	15342	15342

*. Correlation is significant at the 0.05 level (2-tailed).

As Table 38 shows, the true values of the Pearson correlation coefficient for the pairs of variables described above lie within [-0,035; -0,004] and [0,0002; 0,032] 95% confidence intervals.

For the other variables in the Spheres of violence group displayed in Table 37, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Royal family scandals, Violence in figure skating, SV during the war in Ukraine and Japanese program for women in sexual slavery, and the degree of negative sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a negative sentiment.

Table 38. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@2 - negative_75	-0,019588878	0,015250759	-0,035401797	-0,003766153
@23 - negative_max	0,01604085	0,046941136	0,000217013	0,031856656

a. Estimation is based on Fisher's r-to-z transformation.

As for the neutral sentiments, as Table 39 shows, it can be said that the probability of the topic SV during the war in Ukraine is higher in the articles that cover sexual violence in a less neutral way, which can be stated according to negatively directed correlations between this topic and the degree of neutral sentiment, measured by the 75 percentile value for sentence sentiments (corr = -0.01, p-value < 0.05) and by the maximum value for sentence sentiments (corr = -0.02, p-value < 0.05). *The statistical hypothesis H0 is not supported* on the 95% confidence level, meaning that there is a statistically significant correlation between the variables.

Table 39. Pearson correlation coefficient for the topics in the Spheres of violence topic group and neutral sentiment in articles

		neutral_mean	neutral_75	neutral_max
Royal family scandals	Pearson Correlation	0,008	0,005	0,001
	Sig. (2-tailed)	0,337	0,565	0,931
	N	15342	15342	15342
Violence in hockey	Pearson Correlation	0,007	0,005	0,002
	Sig. (2-tailed)	0,411	0,557	0,804
	N	15342	15342	15342
Violence in figure skating	Pearson Correlation	-0,009	-0,003	0,005
	Sig. (2-tailed)	0,280	0,723	0,574
	N	15342	15342	15342
The convention of womens' rights	Pearson Correlation	0,005	0,006	0,006
	Sig. (2-tailed)	0,508	0,448	0,450
	N	15342	15342	15342
SV during the war in Ukraine	Pearson Correlation	-0,006	-,020*	-,020*
	Sig. (2-tailed)	0,471	0,014	0,014
	N	15342	15342	15342
Japanese program for women in sexual slavery	Pearson Correlation	0,005	0,008	0,016
	Sig. (2-tailed)	0,509	0,296	0,052
	N	15342	15342	15342

*. Correlation is significant at the 0.05 level (2-tailed).

As Table 40 shows, the true values of the Pearson correlation coefficient for the pairs of variables described above lie within [-0,036; -0,004] and [-0,036; -0,004] 95% confidence intervals.

Table 40. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@24 - neutral_75	-0,019902106	0,013694659	-0,03571475	-0,004079498
@24 - neutral_max	-0,019940875	0,013512154	-0,035753486	-0,004118283

a. Estimation is based on Fisher's r-to-z transformation.

For the other variables in the Spheres of violence group displayed in Table 39, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Royal family scandals, Violence in hickey, Violence in figure skating, The convention of womens' rights and Japanese program for women in sexual slavery, and the degree of neutral sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an

article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a neutral sentiment.

Cases

Table 41 shows Pearson linear correlation coefficients for the Cases topic group probabilities and the variables measuring positive sentiment in articles. There are three quite low Pearson linear correlation coefficients that are statistically significant: between Rape in the Ufa police department case topic and two negative sentiment variables, measured as mean (corr = -0.024, p-value < 0.01) and as 75 percentile (corr = -0.025, p-value < 0.01), and also between the Cristiano Ronaldo case topic and a negative sentiment measured as 75 percentile (corr = 0.016, p-value < 0.05). *The statistical hypothesis H0 is not supported*, meaning that there is a statistically significant correlation between the aforementioned variables.

Table 41. Pearson correlation coefficient for the topics in the Cases topic group and positive sentiment in articles

		positive_mean	positive_75	positive_max
Khachaturian sisters case	Pearson Correlation	0,009	0,010	0,012
	Sig. (2-tailed)	0,241	0,236	0,152
	N	15342	15342	15342
Luis Rubiales case	Pearson Correlation	0,009	0,014	0,004
	Sig. (2-tailed)	0,259	0,079	0,612
	N	15342	15342	15342
Harvey Weinstein case	Pearson Correlation	0,009	0,007	-0,004
	Sig. (2-tailed)	0,258	0,415	0,619
	N	15342	15342	15342
Cristiano Ronaldo case	Pearson Correlation	0,007	,016*	0,000
	Sig. (2-tailed)	0,360	0,044	0,962
	N	15342	15342	15342
Benjamin Mendy case	Pearson Correlation	0,010	0,015	0,002
	Sig. (2-tailed)	0,201	0,070	0,830
	N	15342	15342	15342
Gérard Depardieu case	Pearson Correlation	0,015	0,010	0,000
	Sig. (2-tailed)	0,070	0,205	0,972
	N	15342	15342	15342
Daniel Alves case	Pearson Correlation	0,005	0,002	0,015
	Sig. (2-tailed)	0,500	0,825	0,071
	N	15342	15342	15342
Skopinsky maniac case	Pearson Correlation	-0,009	-0,010	0,001
	Sig. (2-tailed)	0,279	0,226	0,944
	N	15342	15342	15342

		positive_mean	positive_75	positive_max
Jeffrey Epstein case	Pearson Correlation	-0,003	0,001	-0,007
	Sig. (2-tailed)	0,728	0,934	0,411
	N	15342	15342	15342
Victoria Marinova case	Pearson Correlation	-0,001	0,000	-0,005
	Sig. (2-tailed)	0,866	0,968	0,558
	N	15342	15342	15342
Rape in the Ufa police department case	Pearson Correlation	-,024**	-,025**	-0,003
	Sig. (2-tailed)	0,003	0,002	0,729
	N	15342	15342	15342
Roman Polanski case	Pearson Correlation	0,001	-0,001	0,003
	Sig. (2-tailed)	0,891	0,906	0,687
	N	15342	15342	15342
Collectors case	Pearson Correlation	0,003	0,004	0,000
	Sig. (2-tailed)	0,710	0,579	0,958
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

As Table 42 shows, the true value of the Pearson correlation coefficient between the probability of Cristiano Ronaldo case topic and the degree of positive sentiment in the article lies within the [0; 0,032] interval with a 95% level. The true values of the Pearson correlation coefficient between the probability of Rape in the Ufa police department case topic and the degree of positive sentiment measured by mean and 75 percentile values of sentence sentiment lie within the [-0,039; -0,008] and [-0,040; -0,009] 95% confidence intervals respectively.

Table 42. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@9 - positive_75	0,016	0,044	0,000	0,032
@26 - positive_mean	-0,024	0,003	-0,039	-0,008
@26 - positive_75	-0,025	0,002	-0,040	-0,009

a. Estimation is based on Fisher's r-to-z transformation.

For the other variables in the Cases group displayed in Table 41, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Khachaturian sisters case, Luis Rubiales case, Harvey Weinstein case, Benjamin Mendy case, Gérard Depardieu case. Daniel Alves case, Skopinsky maniac case, Jeffrey Epstein case, Victoria Marinova case, Roman Polanski case and Collectors case, and the degree of positive sentiment in the articles are not

linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a positive sentiment.

Table 43. Pearson correlation coefficient for the topics in the Cases topic group and negative sentiment in articles

		negative_mean	negative_75	negative_max
Khachaturian sisters case	Pearson Correlation	0,001	0,006	-0,004
	Sig. (2-tailed)	0,885	0,461	0,639
	N	15342	15342	15342
Luis Rubiales case	Pearson Correlation	0,002	0,001	,017*
	Sig. (2-tailed)	0,828	0,876	0,033
	N	15342	15342	15342
Harvey Weinstein case	Pearson Correlation	0,002	-0,001	-0,013
	Sig. (2-tailed)	0,822	0,925	0,118
	N	15342	15342	15342
Cristiano Ronaldo case	Pearson Correlation	-0,001	-0,007	0,002
	Sig. (2-tailed)	0,870	0,379	0,837
	N	15342	15342	15342
Benjamin Mendy case	Pearson Correlation	-0,002	-0,009	0,007
	Sig. (2-tailed)	0,847	0,291	0,384
	N	15342	15342	15342
Gérard Depardieu case	Pearson Correlation	0,010	0,001	-0,006
	Sig. (2-tailed)	0,228	0,931	0,474
	N	15342	15342	15342
Daniel Alves case	Pearson Correlation	0,005	0,003	0,005
	Sig. (2-tailed)	0,566	0,714	0,506
	N	15342	15342	15342
Skopinsky maniac case	Pearson Correlation	-0,004	0,004	-0,001
	Sig. (2-tailed)	0,664	0,647	0,929
	N	15342	15342	15342
Jeffrey Epstein case	Pearson Correlation	0,002	-0,008	0,005
	Sig. (2-tailed)	0,797	0,340	0,549
	N	15342	15342	15342
Victoria Marinova case	Pearson Correlation	0,003	0,000	0,002
	Sig. (2-tailed)	0,667	0,951	0,793
	N	15342	15342	15342
Rape in the Ufa police department case	Pearson Correlation	-0,008	-0,002	0,006
	Sig. (2-tailed)	0,313	0,810	0,455
	N	15342	15342	15342
Roman Polanski case	Pearson Correlation	0,014	0,010	0,011
	Sig. (2-tailed)	0,081	0,234	0,181
	N	15342	15342	15342
Collectors case	Pearson Correlation	-0,007	-0,008	-0,005
	Sig. (2-tailed)	0,412	0,346	0,528
	N	15342	15342	15342

As for the negative sentiment, as shown in Table 43, for all the topics in the Cases topic group – Khachaturian sisters case, Luis Rubiales case, Harvey Weinstein case,

Cristiano Ronaldo case, Benjamin Mendy case, Gérard Depardieu case. Daniel Alves case, Skopinsky maniac case, Jeffrey Epstein case, Victoria Marinova case, Rape in the Ufa police department case, Roman Polanski case and Collectors case – Pearson correlation coefficients are not statistically significant at any – neither 95% nor 99% – confidence levels. *The statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics in the Cases topic group and the degree of negative sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with negative emotional color.

Table 44. Pearson correlation coefficient for the topics in the Cases topic group and neutral sentiment in articles

		neutral_mean	neutral_75	neutral_max
Khachaturian sisters case	Pearson Correlation	-0,007	-0,004	-0,002
	Sig. (2-tailed)	0,404	0,599	0,804
	N	15342	15342	15342
Luis Rubiales case	Pearson Correlation	-0,005	-0,005	0,005
	Sig. (2-tailed)	0,555	0,498	0,534
	N	15342	15342	15342
Harvey Weinstein case	Pearson Correlation	-0,011	-,018*	-,017*
	Sig. (2-tailed)	0,175	0,026	0,033
	N	15342	15342	15342
Cristiano Ronaldo case	Pearson Correlation	0,007	0,002	0,005
	Sig. (2-tailed)	0,398	0,757	0,530
	N	15342	15342	15342
Benjamin Mendy case	Pearson Correlation	0,005	0,004	-0,002
	Sig. (2-tailed)	0,539	0,660	0,818
	N	15342	15342	15342
Gérard Depardieu case	Pearson Correlation	-0,012	-,019*	-,026**
	Sig. (2-tailed)	0,126	0,021	0,001
	N	15342	15342	15342
Daniel Alves case	Pearson Correlation	-0,005	-0,006	-0,008
	Sig. (2-tailed)	0,538	0,429	0,331
	N	15342	15342	15342
Skopinsky maniac case	Pearson Correlation	0,015	,018*	0,012
	Sig. (2-tailed)	0,071	0,027	0,126
	N	15342	15342	15342
Jeffrey Epstein case	Pearson Correlation	0,004	0,000	-0,007
	Sig. (2-tailed)	0,616	0,996	0,391
	N	15342	15342	15342
Victoria Marinova case	Pearson Correlation	-0,001	-0,003	-0,005
	Sig. (2-tailed)	0,884	0,741	0,537
	N	15342	15342	15342

		positive_mean	positive_75	positive_max
Rape in the Ufa police department case	Pearson Correlation	,022**	,032**	,025**
	Sig. (2-tailed)	0,005	0,000	0,002
	N	15342	15342	15342
Roman Polanski case	Pearson Correlation	-0,008	-0,008	-0,005
	Sig. (2-tailed)	0,341	0,308	0,537
	N	15342	15342	15342
Collectors case	Pearson Correlation	-0,011	-0,003	-0,003
	Sig. (2-tailed)	0,180	0,684	0,748
	N	15342	15342	15342

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

As for the neutral sentiment, as shown in Table 44, there are several Pearson linear correlation coefficients that are statistically significant on the 95% or 99% confidence level. These are the correlations between: probability of Harvey Weinstein case topic in an article and neutral sentiment variable measured by 75 percentile (corr = -0.018, p-value < 0.05) and maximum (corr = -0.017, p-value < 0.05) values of sentence sentiment; probability of Gérard Depardieu case topic in an article and neutral sentiment variable measured by 75 percentile (corr = -0.019, p-value < 0.05) and maximum (corr = -0.026, p-value < 0.01) values of sentence sentiment; the probability of Skopinsky maniac case topic in an article and neutral sentiment variable measured by 75 percentile value of sentence sentiment (corr = 0.018, p-value < 0.05); probability of Rape in the Ufa police department case topic in an article and all three neutral sentiment variables - measured by mean (corr = 0.022, p-value < 0.01), 75 percentile (corr = 0.032, p-value < 0.01) and maximum (corr = 0.025, p-value < 0.01) values of sentence sentiment. For these pairs of variables, *the statistical hypothesis H0 is not supported*, meaning that there is a statistically significant correlation between the aforementioned variables.

Table 45. Confidence intervals for the correlations

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a]	
			Lower	Upper
@8 - neutral_75	-0,018	0,026	-0,034	-0,002
@8 - neutral_max	-0,017	0,033	-0,033	-0,001
@14 - neutral_75	-0,019	0,021	-0,034	-0,003
@14 - neutral_max	-0,026	0,001	-0,042	-0,011
@18 - neutral_75	0,018	0,027	0,002	0,034
@26 - neutral_mean	0,022	0,005	0,007	0,038
@26 - neutral_75	0,032	0,000	0,016	0,048
@26 - neutral_max	0,025	0,002	0,009	0,041

Table 45 shows the 95% confidence intervals for the true values of Pearson linear correlation coefficients for all of the aforementioned pairs of variables. The true values of these correlations for the pairs: probability of Harvey Weinstein case topic in an article and neutral sentiment variable measured by 75 percentile and maximum values of sentence sentiment lie within the [-0,034; -0,002] and [-0,033; -0,001] 95% confidence intervals; probability of Gérard Depardieu case topic in an article and neutral sentiment variable measured by 75 percentile and maximum values of sentence sentiment lie within the [-0,034-0,003] and [-0,042; -0,011] 95% confidence intervals; the probability of Skopinsky maniac case topic in an article and neutral sentiment variable measured by 75 percentile value of sentence sentiment lie within the [0,002; 0,034] 95% confidence interval; probability of Rape in the Ufa police department case topic in an article and all three neutral sentiment variables - measured by mean, 75 percentile and maximum values of sentence sentiment lie within [0,007; 0,038], [0,016; 0,048] and [0,009; 0,041] 95% confidence intervals.

For the other variables in the Cases group displayed in Table 44, *the statistical hypothesis H0 is supported:* at any confidence level, it is fair to say that topic probabilities for the topics Khachaturian sisters case, Luis Rubiales case, Cristiano Ronaldo case, Benjamin Mendy case, Daniel Alves case, Jeffrey Epstein case, Victoria Marinova case, Roman Polanski case and Collectors case, and the degree of neutral sentiment in the articles are not linearly correlated. There, no matter the probability of encountering either of these topics in an article on sexual violence, the prevalence of these topics cannot be connected to sexual violence being covered in media with a neutral sentiment.

How topic prevalence is connected to the sensationalism in articles

The second research objective (RO2) of this thesis was *to explore the relationship between the themes present in the articles on sexual violence against women and the tone that is used by Russian news media to cover such violence.* To achieve this task, I estimated the Pearson linear correlation coefficients between the topic probabilities and

the variables indicating the degree of positive, negative and neutral sentiment in the article texts. There were several statistically (on either 95% or 99% confidence levels) significant correlation coefficients in the analysis that allow to partially support my hypotheses within the RO2.

The first hypothesis (H1) was the following: *the themes that explore the cases of sexual violence will be correlated to either positive or negative sentiment in the coverage of sexual violence.* Since the sensationalism is essentially about the emotional side of the coverage of sexual coverage and is specifically connected to the coverage of violent incidents, for testing this hypothesis the cases themselves may be understood not only as cases from the "Cases" topic group which explores the popular public cases of sexual violence, but also the topics which deal with the regular, "common" incidents of violence reported in the news. For example, the coverage of such day-to-day incidents of sexual violence are contained in such topics as "Rape and assault cases", "Descriptive narratives" from the "General" topic group, as well as in all the topics from the "Criminal cases & process" topic group where the topics deal with the sexually violent cases reported to the police.

The public cases that were significantly connected to the article sentiment variables were: the Christiano Ronaldo case, where the higher probability of encountering the case in the article is related to the higher level of article positivity; the case about the rape in the Ufa police department which is connected to the lower level of article positivity and yet the higher article neutrality in tone; the Harvey Weinstein and Gerard Depardieu cases which probabilities are higher when the articles are less neutral; the Skopinsky maniac case that was reported more neutrally. For the other cases in this topic group, it cannot be said that the cases are covered in a more or less emotional way. For the "Rape and assault cases" there were also no significant correlation coefficients with the sentiment variables and therefore the descriptions of the cases themselves are not used with a significantly present positive or negative tone. However, the "Descriptive narratives" topic was connected to the higher levels of the negative sentiment, which means that graphic images of the violence are covered with higher

levels of negative emotion. As for the criminal cases, the "Details of the criminal cases process" topic and "Perpetrator court sentences murder cases" topics are both likely to be covered in a less positive way, yet tend to correlate with a more neutral emotion. The other topics related to the criminal cases were not connected to the sentiment variables. Therefore, the existing significant connections between some topics and the level of sentiment in the articles partially supports the hypothesis on the emotional coverage of the cases of sexual violence against women.

The second hypothesis (H2) related to the second research objective (RO2) was the following: *the themes that explore sexual violence as a social problem will be correlated to the neutral sentiment in the coverage of sexual violence*. Two topic probabilities were negatively connected to the neutral sentiment – "Fighting SV – organizations and funding" and "Problem regulation" topics. Moreover, the "Fighting SV – organizations and funding" topic was associated with lower levels of negative emotions in the articles and the "Reports and statistics" topic was associated with the higher levels of positive sentiment in the articles. Therefore, the hypothesis cannot be supported based on the available data. It cannot be stated that when talking about sexual violence as a social problem, the media tends to use neutral language of the coverage, since there are various sentiments used in covering sexual violence as a social problem.

The last hypothesis within the task of finding out whether the themes regarding sexual violence against women are covered with different emotions. In this hypothesis (H3), it was anticipated that *there will be no correlation between the other themes and level of sentiment in the articles covering sexual violence against women.* Various statistically significant coefficients speak against this hypothesis: the "Victim perspective" topic is more likely to be covered with a negative sentiment, the "Musicians as perpetrators" topic is more likely to be covered with both negative and positive emotions, the topics exploring the sexual violence cases within hockey and during the armed conflict in Ukraine tend to be covered more positively, whereas the "Convention of women's rights" topic is more likely to be covered with a negative tone.

Therefore, the hypothesis is not supported since there are statistically significant interchangeabilities between the variables.

However, it is important to note that the Pearson correlation coefficients that were statistically significant are not in fact high. Most coefficients do not exceed the 0.025 value, which indicates the very low interchangeability of the tested variables. Moreover, according to the confidence intervals computed for the coefficients, the Pearson correlation coefficients are commonly fluctuating around the 0 value which indicates that the true values of the coefficients are likely to be zero in a whole population. Essentially, sensationalism, understood in terms of emotions, is a feature of coverage that requires high levels of emotional reporting. The data analyzed in this thesis, however, shows that the shifts in the emotions used in coverage of different sexual violence related themes are quite low. Therefore, though showing statistical significance, the relations between the usage of different topics and the level of emotional tone used within the coverage are practically not apparent. It is fair to state here that despite being found on the substantive level (as in detailed covering the images of the sexually violent incidents) at the assumption level in the previous subchapter, the sensational coverage in terms of emotions does not characterize the discussion on sexual violence against women happening in the Russian news media.

Although, it is possible that the discovered absence of sensationalist rhetoric is due to the methodological restrictions of this thesis. The sentiment levels were measured in the sentences and then were approximated to the article level which could significantly distort the image of analysis (though to avoid this in possibility, the maximum-value approximation was used). Therefore, the different methods for measuring the article-level sentiment may be applied to test the validity of the results that were obtained here. Moreover, the sentiment itself also may be not the most appropriate way to measure emotion. Though sentiment analysis models are trained on the large corpora and are able to catch sentiment in texts based on the wordings or word places used to express emotion, it is still possible that the emotion may be may manifest itself in a special deep structure of texts, which sentiment analysis in this case does not

reveal. This restriction could be overcome with the manual coding of the emotion in the articles for the possible further analysis conducted on smaller, and therefore available for manual coding, samples.

Temporal differences in the thematic coverage of sexual violence in Russian mass media

The third and final research objective of this thesis was to explore the temporal differences in the thematic coverage of sexual violence against women in Russian news media. To achieve this task, I ran the one-way ANOVA models to compare mean topic probabilities between the article groups based on the year of the publication. The research objective RO3 itself rose from the framing research literature which suggested the complexity of the framing process, especially regarding the temporal characteristics of frame usage in the coverage of social problems. For this part of the thesis, the analysis was conducted only on the topics from the "General" and "Social problem" topic groups since the topics in these groups are by essence not situational or episodic, such as cases, and consist of the general narratives on the sexual violence that are present in the whole discussion and which can be further interpreted in terms of framing functions.

The following parts of this subchapter contain the results of the one-way ANOVA models for all the topics from the "General" and "Social problem" topic groups, as well as some preliminary conclusions that can be made on the temporal characteristics of the thematic coverage of sexual violence against women in Russian news media.

Rape and assault cases

According to the one-way ANOVA model (results in Table 46), group means of the "Rape and assault cases" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 46. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	22,878	7	3,268	47,908	0,000
Within Groups	1046,094	15334	0,068		
Total	1068,972	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 47), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 47. Results of the test of homogeneity of variances

	Levene Statistic	df1	df2	Sig.
Based on Mean	39,071	7	15334	0,000
Based on Median	27,916	7	15334	0,000
Based on Median and with adjusted df	27,916	7	14577,789	0,000
Based on trimmed mean	39,407	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 47), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 48. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	51,793	7	6347,758	0,000
Brown-Forsythe	47,986	7	14630,609	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A6, Appendix 6.

As shown in Table A6, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the “Rape and assault cases” topic probabilities; the mean values of difference are flagged in the table. The figure 49 illustrates the mean values of topic probabilities for all the years in the study period. To sum up all of the group differences, the year 2021 is characterized by the lowest values of topic probabilities compared to all other years and the highest values of topic probabilities are observed in the years 2016, 2017 and 2019, with no statistically significant difference between these three groups. The mean topic probability for the year 2018 does not differ significantly from the mean probabilities in 2020 and 2022; the mean probability of the topic in the year 2023 is significantly higher that the mean probability in the years 2018 and 2021 and is significantly lower that in the years 2016, 2017, 2019 and so on. Since almost all the group differences are statistically significant (p-value < 0.05), the illustration of the dynamic of the mean topic probability in Figure 49 may quite accurately display these differences.

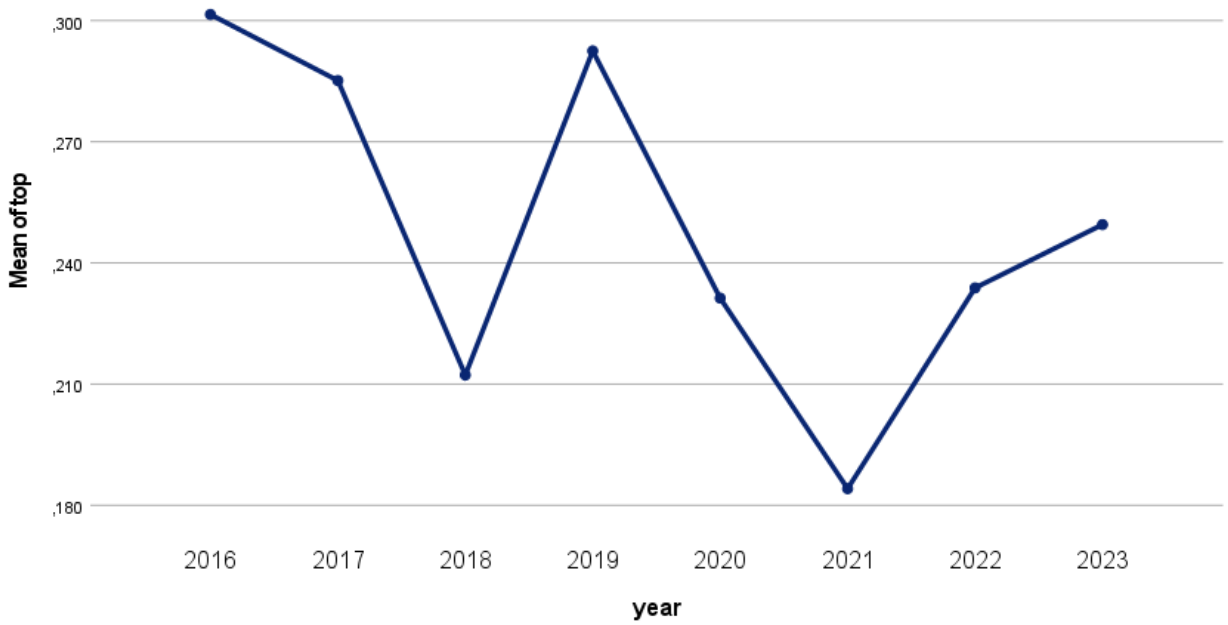

Figure 49. The dynamic of the topic, by yearly mean probability

Therefore, the group differences in the mean probabilities of the “Rape and assault cases” topic indicate that the topic, initially used in coverage of sexual violence

with about 0.3 mean probability, faced the changes in the mean probability both in a positive and negative directions throughout the study period, with the peaks of mean topic probability in the years 2016, 2017 and 2019 and the downfalls in the years 2018 and 2021.

Victim perspective

According to the one-way ANOVA model (results in Table 49), group means of the "Victim perspective" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 49. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,745	7	0,106	14,659	0,000
Within Groups	111,322	15334	0,007		
Total	112,067	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 50), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 50. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
sup0	Based on Mean	28,044	7	15334	0,000
	Based on Median	14,468	7	15334	0,000
	Based on Median and with adjusted df	14,468	7	14739,906	0,000
	Based on trimmed mean	23,263	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 51), at the 99% confidence level, it can be stated that the average values of topic

probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 51. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	12,881	7	6322,826	0,000
Brown-Forsythe	14,444	7	13798,698	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A7, Appendix 6.

As shown in Table A7, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the “Victim perspective” topic probabilities. The mean values of topic probabilities in the groups are illustrated in Figure 59. The highest mean probability value id observed in the year 2020 and is statistically higher than the mean probabilities of all other years except year 2021. Also, for example, there are statistically significant differences between the year 2016 and the years 2018, 2020 and 2021, where the latter three have the higher mean values of topic probability.

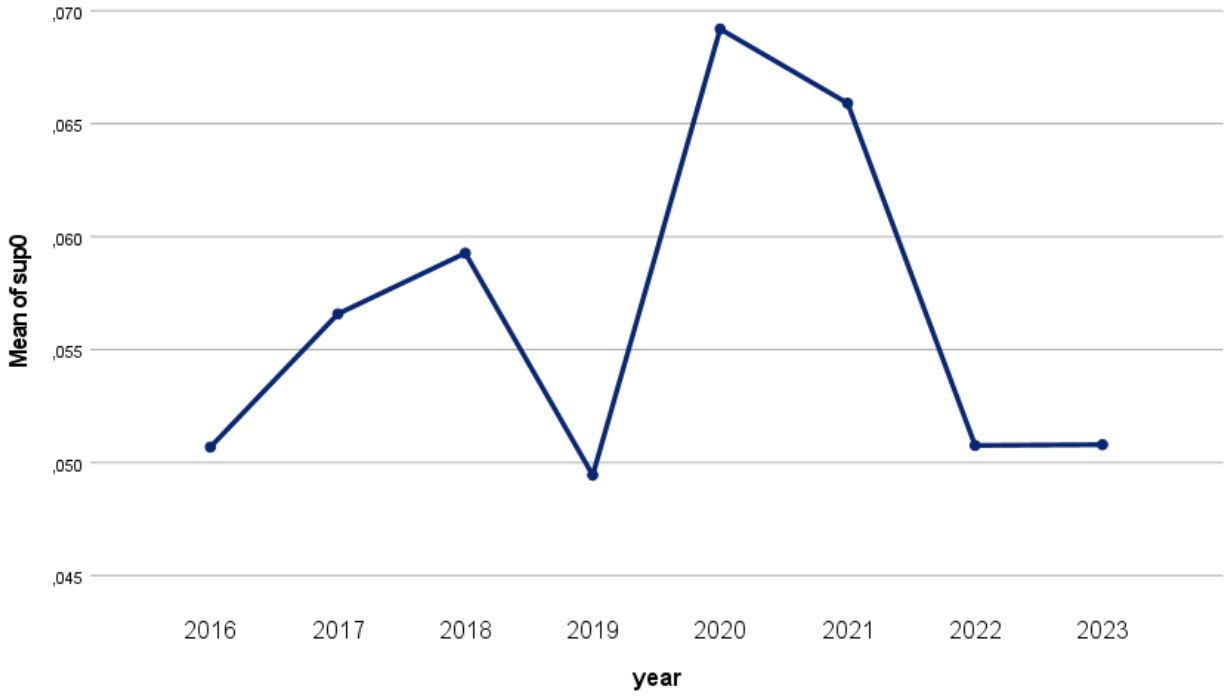

Figure 59. The dynamic of the topic, by yearly mean probability

To summarize the temporal differences in the usage of the "Vuctim perspective" topic, it is evident from the data that, having no statistically significant changes up until the year 2019, the usage of the topic became higher in the year 2020 (with the mean absolute difference of 0.02) and then again fell in the year 2022 (the mean absolute difference is 0.015).

Descriptive narratives

According to the one-way ANOVA model (results in Table 52), group means of the "Descriptive narratives" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 52. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,288	7	0,041	9,228	0,000
Within Groups	68,412	15334	0,004		
Total	68,700	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 53), the H0 hypothesis of equality of group variances is rejected at

a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 53. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
sup1	Based on Mean	7,531	7	15334	0,000
	Based on Median	5,984	7	15334	0,000
	Based on Median and with adjusted df	5,984	7	15230,269	0,000
	Based on trimmed mean	7,285	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 54), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 54. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	9,225	7	6328,387	0,000
Brown-Forsythe	9,125	7	14362,070	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A8, Appendix 6.

As shown in Table A8, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the “Descriptive narratives” topic probabilities, with at least one significantly different pair for each of the years, The figure 51 also shows the mean values of topic probability for all the groups. The most difference in terms of number of groups is shown, on the one hand, for the year 2021, which has a statistically significant higher mean of topic probability than the groups formed by the years 2016, 2018, 2019, 2022 and 2023 and, on the other hand.

For the year 2022, which has a statistically significant lower mean value of topic probability than the groups formed by the years 2017, 2018, 2020, 2021, and 2023.

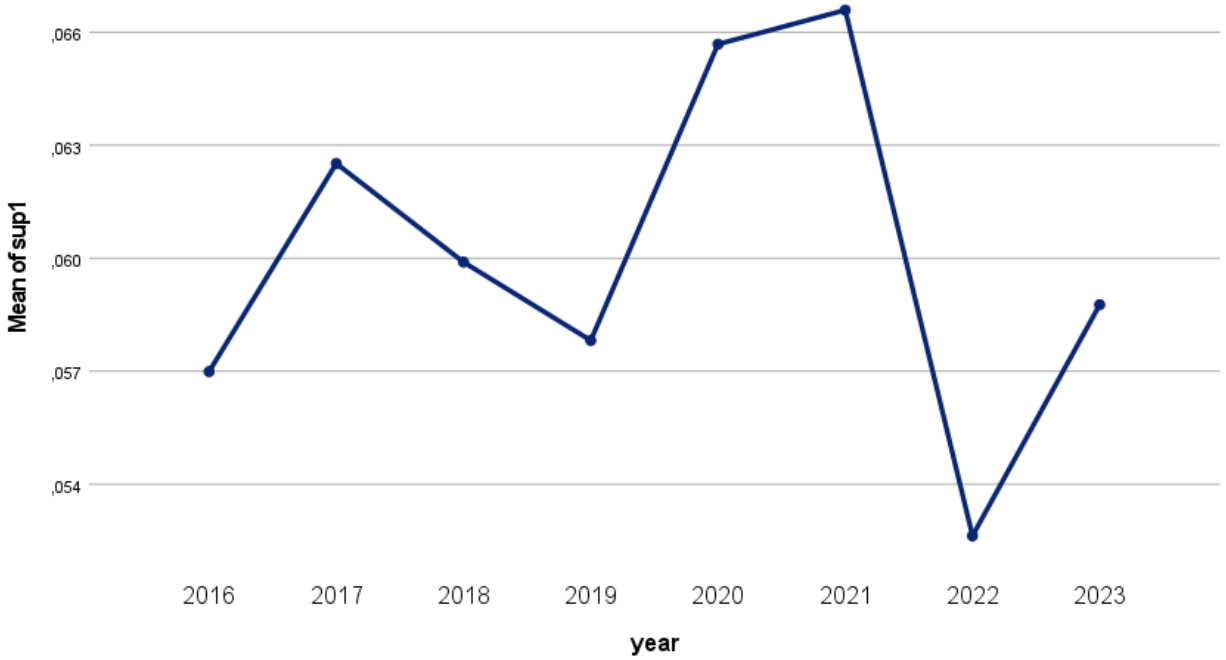

Figure 51. The dynamic of the topic, by yearly mean probability

Therefore, with the various pairs of group differences in mean value of topic probability, the "Descriptive narratives" topic usage in the articles differs from year to year with a non-linear dynamic of change.

Public statements

According to the one-way ANOVA model (results in Table 55), group means of the "Public statements" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 55. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,198	7	0,028	23,369	0,000
Within Groups	18,524	15334	0,001		
Total	18,721	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 56), the H0 hypothesis of equality of group variances is rejected at

a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 56. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
@0	Based on Mean	71,755	7	15334	0,000
	Based on Median	23,369	7	15334	0,000
	Based on Median and with adjusted df	23,369	7	12731,820	0,000
	Based on trimmed mean	46,471	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 57), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 57. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	19,070	7	6330,688	0,000
Brown-Forsythe	23,074	7	12036,236	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A9, Appendix 6.

As shown in Table A9, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the “Public statements” topic probabilities, with at least one significantly different pair for each of the years, The figure 52 also shows the mean values of topic probability for all the groups. The most difference in terms of number of groups is shown by the years 2020 and 2021 that have statistically significant mean values of topic probability that are higher than for all the other groups, with the absent difference between these years themselves. Also, the

year 2016 shows a lower level than all the other groups, except for the 2022 group, of mean topic probability value that are statistically significant (p-value < 0.05).

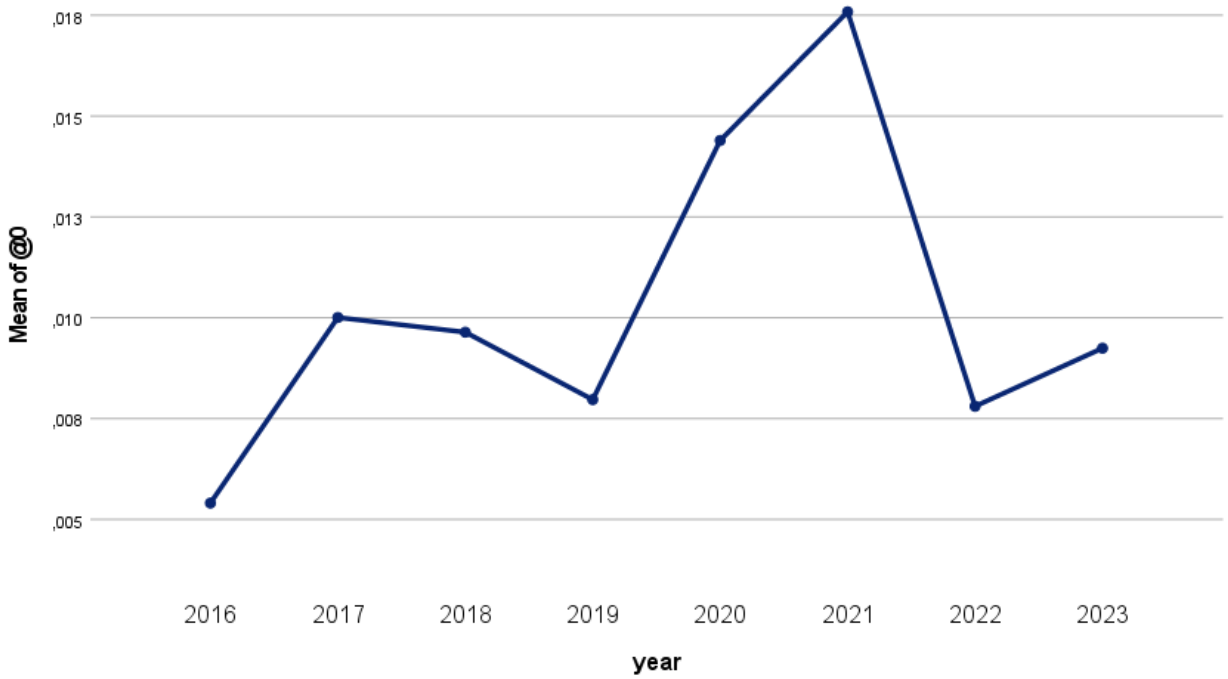

Figure 52. The dynamic of the topic, by yearly mean probability

Overall, there are many between-groups (years) differences in the usage of the "Public statements" topic that indicate that for different years, there are different probabilities of encountering this topic in an article; also, there is no linear order between the differing groups in terms of mean topic probability values, meaning that the usage of topic is higher or lower through all the study period without any one-way trends.

Sexual harassment

According to the one-way ANOVA model (results in Table 58), group means of the "Sexual harassment" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 58. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,264	7	0,038	26,212	0,000
Within Groups	22,068	15334	0,001		
Total	22,333	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 59), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 59. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
@10	Based on Mean	80,648	7	15334	0,000
	Based on Median	26,212	7	15334	0,000
	Based on Median and with adjusted df	26,212	7	11803,388	0,000
	Based on trimmed mean	51,056	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 60), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 60. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	21,773	7	6319,389	0,000
Brown-Forsythe	26,152	7	11337,175	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A10, Appendix 6.

As shown in Table A10, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the “Sexual harassment” topic probabilities, with at least four significantly different pair for each of the years, The figure 53 also shows the mean values of topic probability for all the groups.

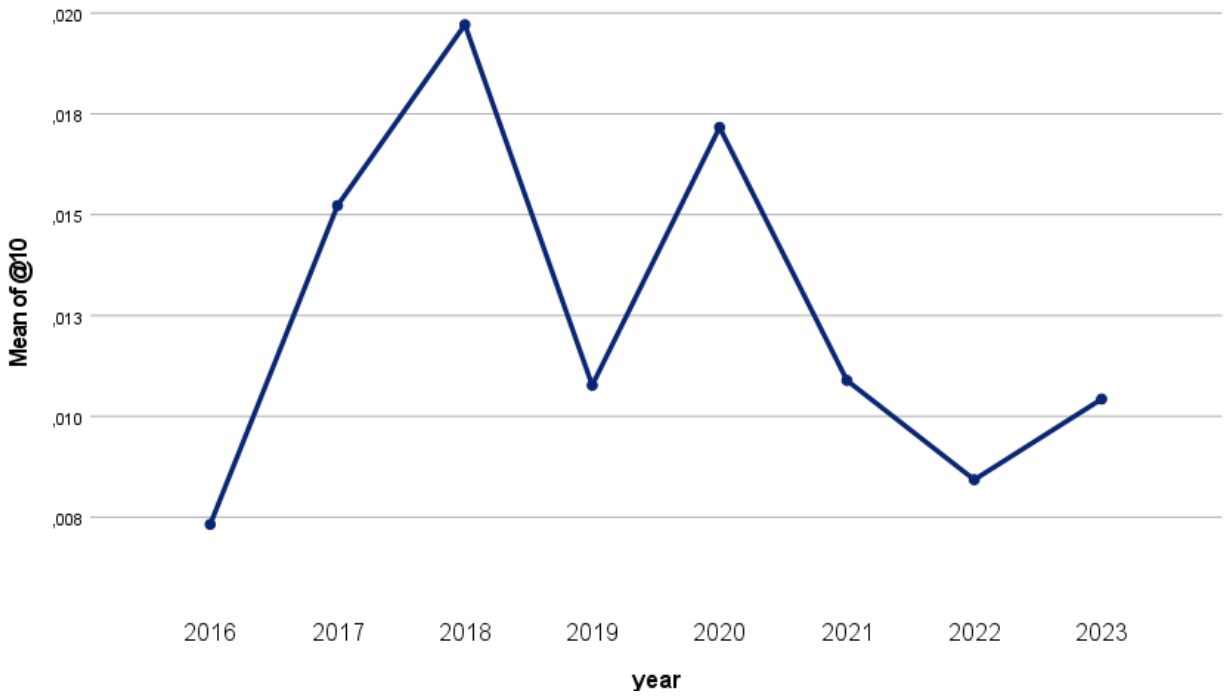

Figure 53. The dynamic of the topic, by yearly mean probability

Overall, as Figure 53 and Table A10 show, there are major and statistically significant fluctuations in the mean values of topic probability throughout the study period, with the peaks in years 2018 and 2020 (which have difference of 0.012 and 0.01 with the mean value of topic probability in the year 2016 as the lowest in terms of topic probability value in the period) and several falls in the mean values of topic probability in other years.

Compensations for victims

According to the one-way ANOVA model (results in Table 61), group means of the “Compensations for victims” topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the

Russian news media use this topic in the coverage of sexual violence against women differently.

Table 61. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,101	7	0,014	26,312	0,000
Within Groups	8,411	15334	0,001		
Total	8,512	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 62), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 62. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
@15	Based on Mean	93,781	7	15334	0,000
	Based on Median	26,312	7	15334	0,000
	Based on Median and with adjusted df	26,312	7	8151,016	0,000
	Based on trimmed mean	41,197	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 63), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 63. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	9,302	7	6329,030	0,000
Brown-Forsythe	26,500	7	8291,184	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the

unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A11, Appendix 6.

As shown in Table A11, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the "Compensations for victims" topic probabilities, yet only in pair with the year 2021 which shows a significantly higher mean topic probability compared to all the other years.

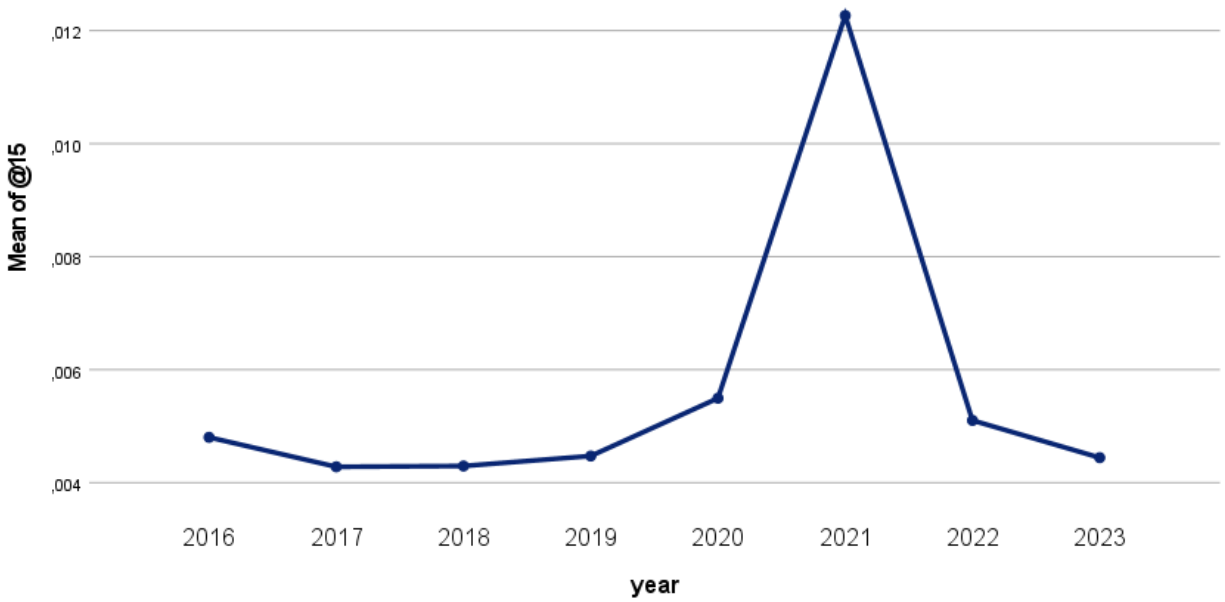

Figure 54. The dynamic of the topic, by yearly mean probability

As shown in Figure 54, the "Compensations for victims" topic is covered highly only in the year 2021 with a mean topic probability of more than 0.12 for the articles in this year. The other groups formed on the year basis, as shown in Table A11, do not significantly differ from each other in term of mean topic probability.

Problem regulation

According to the one-way ANOVA model (results in Table 64), group means of the "Problem regulation" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 64. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,847	7	0,121	12,714	0,000
Within Groups	145,963	15334	0,010		
Total	146,810	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 65), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 65. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
sup2	Based on Mean	13,710	7	15334	0,000
	Based on Median	8,748	7	15334	0,000
	Based on Median and with adjusted df	8,748	7	14935,147	0,000
	Based on trimmed mean	12,186	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 66), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 66. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	12,747	7	6334,232	0,000
Brown-Forsythe	12,618	7	14265,192	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A12, Appendix 6.

As shown in Table A12, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the "Problem regulation" topic probabilities, with at least two significantly different pairs for each of the years. The figure 55 also shows the mean values of topic probability for all the groups. The year that is most differing from other is the year 2018 – it shows the levels of mean topic probability that are significantly higher than for all the other years, except 2017 and 2021. The lowest mean values are shown by the years 2016, 2019 and 2023 that all do not differ from each other significantly.

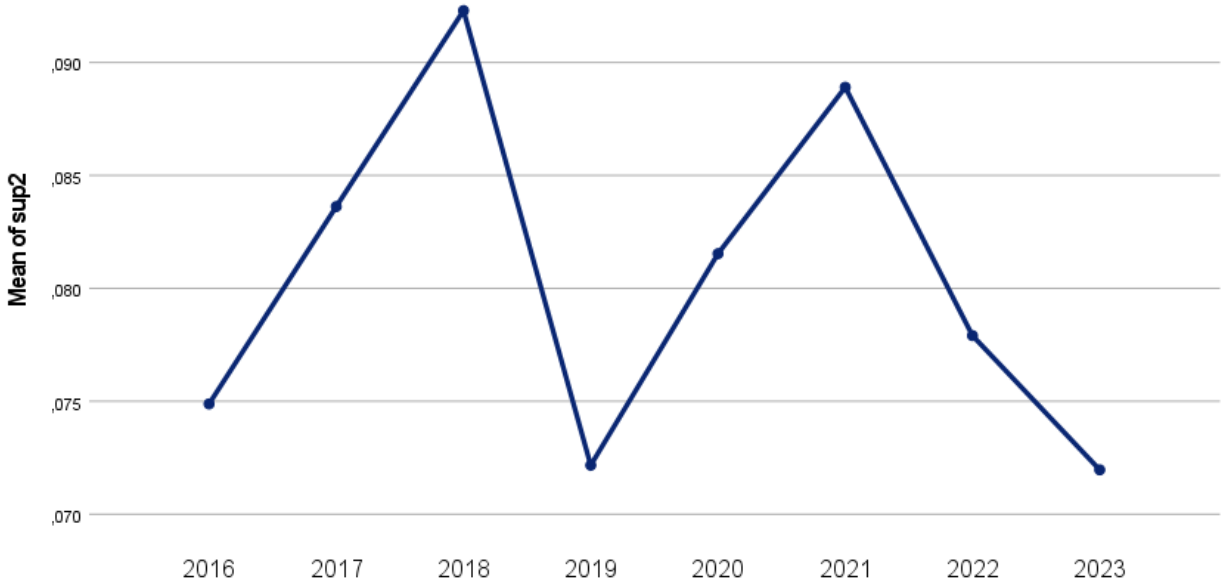

Figure 55. The dynamic of the topic, by yearly mean probability

In overall, both mean values and the results of the Games-Howell test show that the usage of the topic is different in different years, with fluctuations that have no order in terms of how the mean changes through the study period.

Reports and statistics

According to the one-way ANOVA model (results in Table 67), group means of the "Reports and statistics" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 67. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,150	7	0,021	7,910	0,000
Within Groups	41,510	15334	0,003		
Total	41,659	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 68), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 68. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
@6	Based on Mean	20,447	7	15334	0,000
	Based on Median	7,910	7	15334	0,000
	Based on Median and with adjusted df	7,910	7	14856,205	0,000
	Based on trimmed mean	15,358	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 69), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 69. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	8,337	7	6317,011	0,000
Brown-Forsythe	7,805	7	14001,683	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A13, Appendix 6.

As shown in Table A13, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the “Reports and statistics” topic probabilities, with at least one significantly different pair for each of the years. The figure 56 also shows the mean values of topic probability for all the groups. The year 2019 is statistically different from all the other years in terms of the mean topic probability – as seen in Table A13, the chances to encounter this topic in an article published in 2019 are significantly lower in comparison from the other years. Moreover, in year 2022, it is more likely to encounter this topic than in the years 2016, 2019, 2021 and 2023. The other between-group differences are not statistically significant.

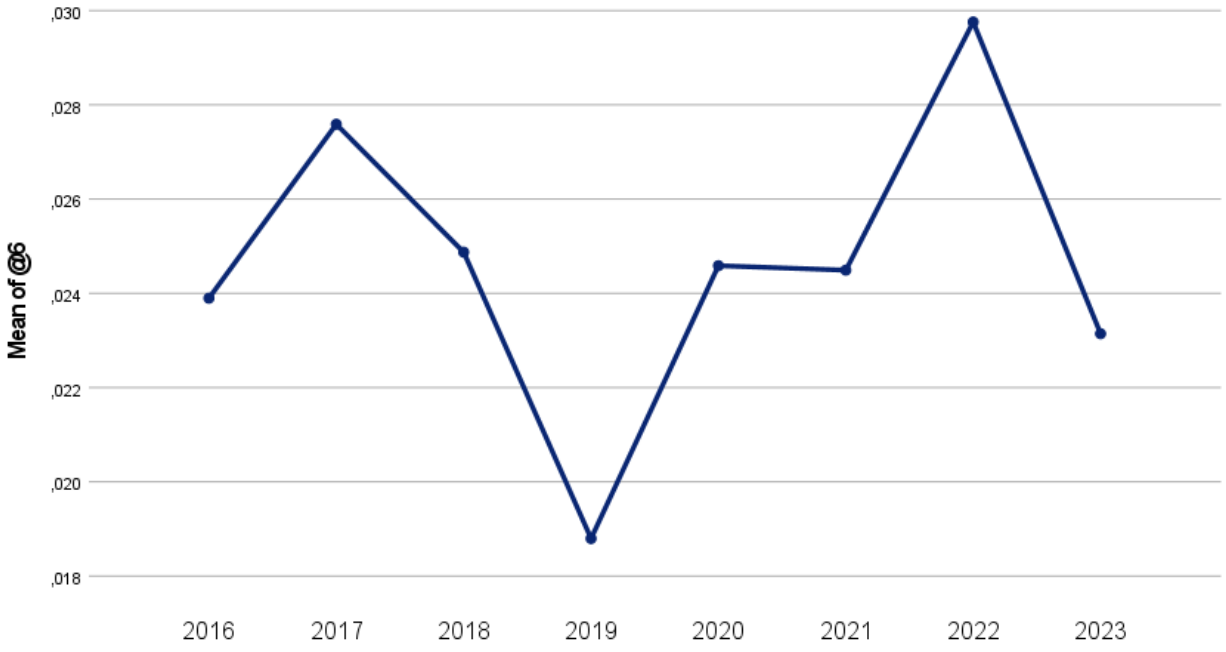

Figure 56. The dynamic of the topic, by yearly mean probability

Therefore, it is evident from the data that the usage of the “Reports and statistics” topic mostly does not depend on a year, with two exceptions in years 2019 and 2022 discussed above.

Fighting SV - organizations and funding

According to the one-way ANOVA model (results in Table 70), group means of the “Fighting SV – organizations and funding” topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in

different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 70. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,129	7	0,018	27,003	0,000
Within Groups	10,477	15334	0,001		
Total	10,607	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 71), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 71. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
@17	Based on Mean	92,515	7	15334	0,000
	Based on Median	27,003	7	15334	0,000
	Based on Median and with adjusted df	27,003	7	11970,191	0,000
	Based on trimmed mean	55,180	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 72), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 72. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	27,276	7	6333,023	0,000
Brown-Forsythe	27,305	7	12024,475	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the

unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A14, Appendix 6.

As shown in Table A14, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the "Fighting SV – organizations and funding" topic probabilities, with at least two significantly different pairs for each of the years. The figure 57 also shows the mean values of topic probability for all the groups. The year that differs most from the others in terms of mean probability value is the year 2016 – in comparison to all the other years in the period, there is a significantly lower mean of topic probability in year 2016. The year 2018 also shows significant difference (with a higher mean topic probability) with all the years except for the year 2017. The year 2019 shows higher mean topic probability than the year 2016, and the lower mean topic probability than the years 2017, 2017 and 2020. There is no statistically significant difference in the mean topic probabilities in the later years of the period which indicates that it is equally likely to encounter the "SV – organizations and funding" topic in the articles published in years 2020-2023.

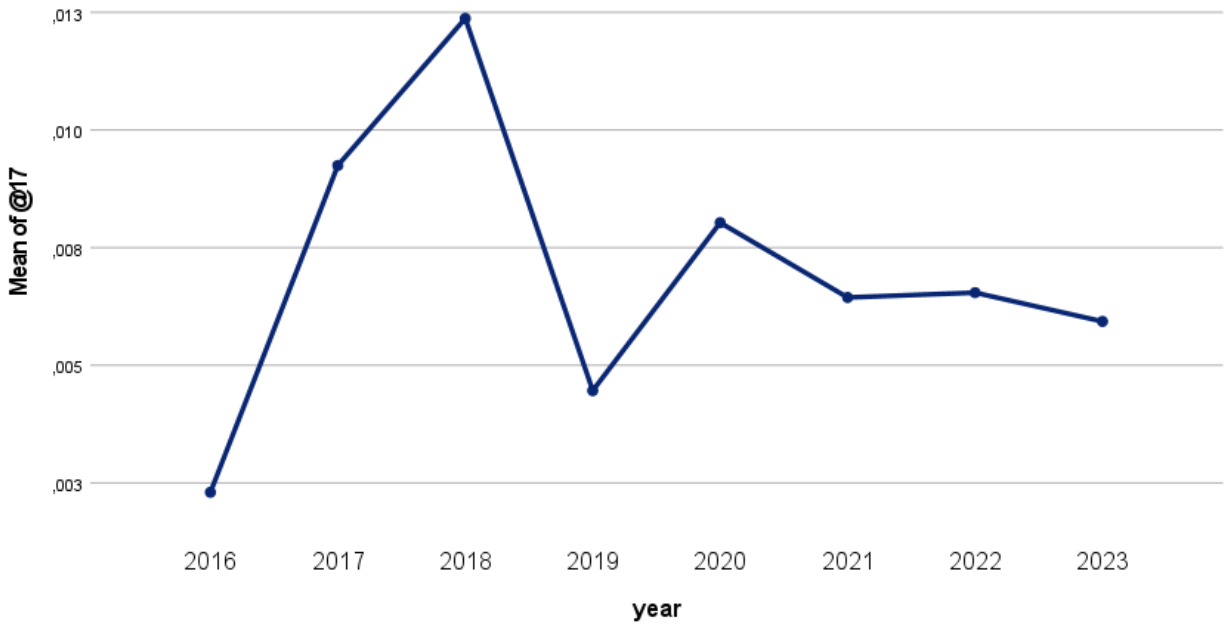

Figure 57. The dynamic of the topic, by yearly mean probability

Nobel prize for fighting SV

According to the one-way ANOVA model (results in Table 72), group means of the "Nobel prize for fighting SV" topic probabilities are not equal at the 99% confidence level (p-value < 0.001). This means that the data suggests that in different years, the Russian news media use this topic in the coverage of sexual violence against women differently.

Table 72. Results of the one-way ANOVA test

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0,100	7	0,014	18,019	0,000
Within Groups	12,196	15334	0,001		
Total	12,296	15341			

As a result of testing the one-way ANOVA model for homogeneity of variances (as shown in Table 73), the H0 hypothesis of equality of group variances is rejected at a confidence level of 99%: this indicates that the group variances are not equal (heterogenous). Since the ANOVA model works better in a situation of homogeneity, it is also necessary to use alternative (robust) Welch and Brown-Forsythe tests to compare average values of topic probabilities in groups.

Table 73. Results of the test of homogeneity of variances

		Levene Statistic	df1	df2	Sig.
@19	Based on Mean	66,070	7	15334	0,000
	Based on Median	18,019	7	15334	0,000
	Based on Median and with adjusted df	18,019	7	5493,133	0,000
	Based on trimmed mean	21,586	7	15334	0,000

According to the results of the Welch and Brown-Forsythe tests (as shown in Table 74), at the 99% confidence level, it can be stated that the average values of topic probabilities in groups are not equal. Thus, robust tests confirm the result of the one-way ANOVA under conditions of heterogeneity of the group variances.

Table 74. Results of the robust tests of equality of means

	Statistic[a]	df1	df2	Sig.
Welch	10,986	7	6327,453	0,000
Brown-Forsythe	20,218	7	6427,686	0,000

a. Asymptotically F distributed.

To detect the pairs of groups that have differences in the topic probabilities at the 95% confidence level, the Games-Howell post-hoc chosen on the premises of the unequal group variances and unequal number of cases within the groups test was applied to the data. The full results of the test are displayed in the Table A15, Appendix 6.

As shown in Table A15, there are several pairs of year groups that are characterized by the statistically significant (p-value < 0.5) differences in the "Nobel prize for fighting SV" topic probabilities, with at least one significantly different pair for each of the years. The figure 58 also shows the mean values of topic probability for all the groups. The year that differs from the others the most is the year 2018, where the mean topic probability is significantly higher than in all the other years. The year 2019, oppositely, is characterized by the lower mean topic probability than most of the years (2018-2022). The year 2023 also shows lower mean topic probability than years 2018 and 2020-2022. However, this topic should be considered situational since it mostly covers one public incident related to the fight of sexual violence as a social problem.

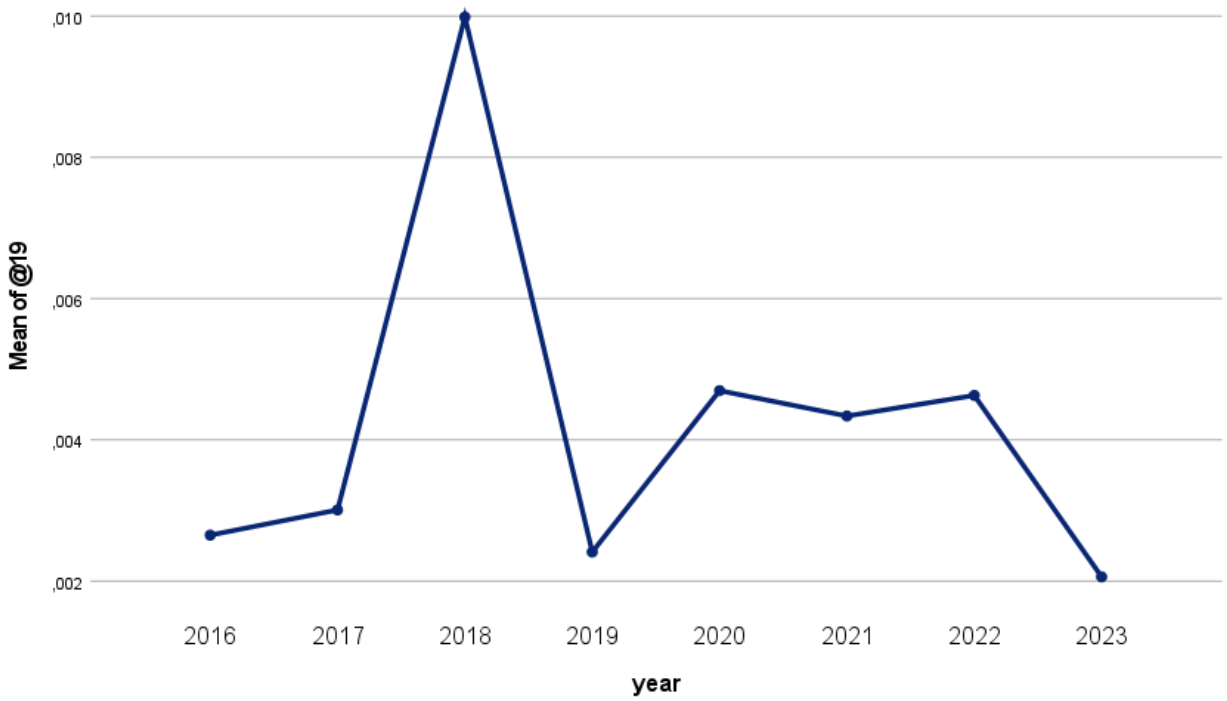

Figure 58. The dynamic of the topic, by yearly mean probability

How do the topics change in time

The last research objective of this thesis was to explore the temporal differences in the thematic coverage of sexual violence against women in Russian news media. To achieve this task, I examined the temporal differences in the coverage of sexual violence for the themes that are generally present in the discussion as a whole, which are represented by the topics from the "General" and "Social problem" topic groups. The hypothesis tested in regard of the third research objective (H4) was the following: *the general themes within the discussion on sexual violence in Russian news media will tend to be fading over time in terms of their presence in the articles*.

The topics which change in time by years was tested in this part of the thesis showed temporal patterns that could not be reduced to the simple linear notion of fading over time. Most of the topics showed higher levels of mean topic probability in some years and lower mean topic probability in other years, with the mean topic probability fluctuating over the whole study period, with the fluctuation being not random but rather statistically significant as assessing different group differences with the post-hop tests showed. Therefore, the H4 cannot be supported based on the available data.

It is important to note here that the results of the analysis are based on the year-by-year comparisons of mean topic probability values, with the whole temporal structure thus being reduced to only eight groups. This creates some methodological limitations of assessing temporal characteristics of framing since the possible trends and temporal patterns may be overlooked due to the simplification of the date variable. The division by years creates artificial boundaries in the data, and therefore, to further analyze the temporality of framing, it is worth trying other methods, for example, methods that include a sliding window rather than strict division into groups. Nevertheless, even with these study limitations the absence of the negative trend in mean topic probability is of no doubt. The main themes in the discussion on sexual violence against women remain fairly stable over time in the articles published by the Russian mass media outlets.

The results correspond with the notions in framing literature on the ongoing process of framing which is diachronic in its nature, or based on reinforcing the themes that are already present in discussion. The fading of frames through time as a notion is mainly based on the assumption that certain frames lose the need to be used in a communication text when they are completely internalized, or included in personal cognitive schemas, by the audience (Entman, Matthes & Pellicano, 2009). Therefore, the results may be interpretated in a way that the ways that sexual violence is covered is not a common way to interpret sexual violence among the Russian audience and the discussion itself is still ongoing.

However, on the other hand, the rejection of the hypothesis is also possible due to the fact that it originated from the features that characterize frames yet then applied to the topics, which are not synonymous to frames themselves. Nevertheless, it is possible to interpret the topics in framing functions, which will be assessed in the next chapter of the thesis.

DISCUSSION: FRAMES AND FRAMING FUNCTIONS

This study is made on how the Russian media framed the sexually violent cases and the discussion on sexual violence against women in general. In the process of the study, several research objectives were made on the matter, however targeted not at the frames and framing themselves, but on the thematic structure of the coverage of sexual violence, or topics that were used in the articles published in Russian news media outlets in regard of sexual violence against women. The reason for primarily targeting the thematic structure in the analysis lies within the complex structure of framing where the frame detection, as well as the interpretation of topics as frames must be made with caution. The frames themselves, being the ways of making the pieces of information salient in a text in order to shape certain interpretations, are not essentially synonymous to themes used in media texts (Matthes, 2009), though may be detected through the framing functions that are easier to operationalize in the study and lead to more valid results. The definition of frames in terms of framing functions implies that frames "*define problems* – determine what a causal agent is doing with what costs and benefits, usually measured in terms of common cultural values; *diagnose causes* – identify the forces creating the problem; *make moral judgments* – evaluate causal agents and their effects; and *suggest remedies* – offer and justify treatments for the problems and predict their likely effects" (Entman, 1993, p. 52), where one or more of these four framing functions may be yielded by a topic or a combination of topics. Such careful approach was used by Saldana Villa in her study on the framing of Chilean earthquakes (Saldana Villa, 2017) and was also adopted in my thesis. Here, I implement the inductive approach of framing research that only implies loose anticipations of the frames ad framing patters that are to be found (for examples and arguments for this approach see de Vreese, 2005; Guenther et al., 2023; Matthes, 2009; Semetko & Valkenburg, 2000), with topic modeling as a prime method to detect a thematic structure of the articles for the further interpretation in terms of framing functions. Therefore, in order to answer the research question of this thesis, or how do Russian news media frame the cases of sexual violence against women, as well the sexual violence against women in general,

the interpretative work based on the findings described in the previous chapter, has to be made.

The problem definition function of framing, based on determining what a causal agent is doing with what costs and benefits, within the subject matter of this study implies that a frame will procedurally describe the sexually violent incidents along with the consequences for both the victim and perpetrator. This framing function is yielded by several topics within the thematic structure of the discussion on sexual violence against women in Russian news media. First, this framing function may be observed in all the topics from the "Criminal cases & process" topic group. There, the topics essentially frame sexual violence, if not as a social problem, but at least as a certain event which has legal costs for the perpetrator of violence. Moreover, the whole topic group is based on attributing responsibility for the violence in single individuals, which is a signal of the attribution of the responsibility frame, one of the common five generic frames usually present in any media text on some social issue (Semetko and Valkenburg, 2000). The problem definition function may also be attributed to some of the topics within the "Social problem" topic group. The "Problem regulation" topic, for example, implies the existence on the social problem of sexual violence and addresses the government as a causal agent making some decisions on the matter of violence. The "Reports and statistics" topic also yields the problem definition function since it explores sexual violence in terms of it being the social problem that has major societal consequences. This topic is also notable as it may be interpreted in terms of thematic framing, or illustrating issues not through certain events or people (as most of the topics within the thematic structure of the corpus do) but in a more abstract way that puts them into a wider context (Iyengar, 1996). The signs of problem definition are also present in the "Public statements" and "Victim perspective" topics which present victims of violence and their voices and actions in response to the sexual violence.

The causes diagnostics function of framing is based on identifying the forces creating the problem. This function is hard to notice within the obtained thematic structure of the discussion on sexual violence since there are no separate topics, except

for “Reports and statistics” topic, that put the issue of violence in a broad social context. The topic on reports and statistics itself though doesn’t essentially diagnose causes since many articles that use this topic simply describe the reports on sexual violence made by non-profit organizations and primarily address the prevalence rates of sexual violence. However, one of the topics may yield this function in a specific way – the “Migrants as perpetrators” topic which usage implies the certain ethnic characteristics of the crime perpetrator. Though the topic itself does not diagnose any causes and mostly describes the cases, the high prevalence of the topic and the emergence of this topic in the thematic structure of the corpus as such may indicate framing the cause of sexual violence in terms of some “outside force” or “external threat”. Since framing, according to Entman (1993), is by essence making some aspects of a communication text more salient, the fact that migrants as perpetrators emerged into a separate topic (and no other social or economic characteristics of the perpetrators did not) indicates that the journalists highlight the ethnicity as a potential threat and cause for the problem of sexual violence.

Making moral judgments as a framing function is evaluating causal agents and their effects. There are no topics that directly indicate moral judgments, though some of the topics may be interpreted in terms of evaluating the agents. The topic “Compensations for victims” which in essence is about the economic retribution for the violence, may indicate the judgment of the perpetrators as the people responsible for the violence – since their actions are to be retributed. However, it is unclear how the compensations themselves are framed within the topic – it is also possible that the compensations themselves may be treated as unnecessary by the authors of the articles on this topic. Nevertheless, the framing itself, as a process, doesn’t imply the substantive component of the framed issues and therefore the “Compensations for victims” topic may be interpreted as yielding a moral judgment function despite the fact that it is unknown what these judgments exactly are. The other topic that possibly includes making judgements is “Nobel prize for fighting SV” topic since it is essentially about evaluating the agents included in the problem solving and being rewarded for it.

The *remedies suggestion* function implies that the topics yielding it offer and justify treatments for the problems and predict their likely effects. The topic that may be interpreted through this function is the "Problem regulation topic" which explores the ways through which sexual violence is or should be legally and legislatively treated. The legislation discourse itself indicates that there are remedies suggestion present in the part of the discussion on sexual violence that includes this topic. The remedies suggestion function may also be applied to the "Fighting SV – organizations and funding" and "Nobel prize for fighting SV" topics that contain the information on the people and entities dealing with the problem in some ways and are made salient in the article texts since they are algorithmically definable as separate topics within the thematic structure of the analyzed corpus. Moreover, since "Compensations for victims" topic deals with the economic retribution of violence which is simply the post-treatment of the sexual violence issue, it also may be addressed in terms of suggesting remedies through the framing process.

The "Compensations for victims" topic, which was already mentioned above in relation to the framing functions, may also be interpreted as a topic connected to the *economic consequences frame* which is one of the five generic frames commonly present in communication texts, along with the attribution of responsibility, conflict, human interest and morality frames (see Semetko and Valkenburg, 2000). Semetko and Valkenburg (2000) also found that in the sensationalist newspapers, the attribution of responsibility and conflict frames are more likely to be used, while more serious newspapers tend to use human interest frame. The *attribution of responsibility*, as was mentioned above, may be traced back to the topics within the "Criminal cases & process" group. Also, the responsibility frame is also detectable through the "Problem regulation" topic where the government is held responsible for dealing with the consequences of sexual violence, and again through the "Compensations for victims" topic since the economic remedies are made by individuals and organizations held responsible for the sexually violent cases. Some of the mentioned topics in the corpus are connected to the elevated levels of emotional rhetoric, as revealed through the linear

correlations – the “Details of the criminal cases process” and “Perpetrator court sentences murder cases” topics are both likely to be covered in a less positive way, yet, as was noted in the respective section of the analysis, the correlation coefficients, though statistically significant, are quite low and likely to be zero. Moreover, the conflict and human interest frames are not directly found in the thematic structure of the discussion, and therefore the Semetko and Valkenburg’s notions cannot be supported.

In the framing literature, it is common to differentiate between issue-specific frames and generic frames, where the *issue-specific frames* are pertinent only to specific topics or events, and *generic frames* can be identified within different thematic scopes and are not limited by the time of occurrence or different cultural contexts (De Vreese, 2005). In terms of these conceptual definitions, all the topics in the thematic structure of the discussion of sexual violence fall under the issue-specific framing type since the topics essentially are constructed of the narratives regarding the sexually violent cases of the problem of sexual violence in general. The other major distinction is made between *episodic frames,* which stand for merely an illustration of some issue, depicting it though concrete events or people, and *thematic frames*, which illustrate issues in a more abstract way yet puts them into a wider context (Iyengar, 1996). In the thematic structure of the discussion on sexual violence, only “Problem definition” and “Reports and statistics” topic fall under the category of thematic framing since they deal with providing context of the sexual violence issue.

CONCLUSION

In this thesis, I analyzed the ways through which the cases of sexual violence against women and of the discussion on sexual violence against women in general were covered by the Russian news media, as well as the interpretive frameworks, or frames, through which such coverage was carried out in the years 2016-2023. The ways of coverage of sexual violence were identified through the discovering of thematic structure of the discussion on sexual violence through topic modeling with the Pachinko allocation topic model, with the usage of sentiment analysis based on the FastText social network model for exploring the sensationalism in the articles and its relation to the discovered topics. I met several research objectives in my study formulated to achieve the goal of understanding how sexual violence against women is framed in Russian news media.

In my ***first research objective (RO1),*** I aimed to identify and describe dominant themes in the discussion of sexual violence against women in Russian news media. I analyzed 37 topics in articles from 2016-2023 and discovered a diverse thematic structure of the discussion. The most dominant themes I found included narratives on sexual violence itself, criminal cases against perpetrators, and legislative regulation addressing sexual violence as a social issue. Contrary to my initial assumption (A1), I did not find evidence that victims are portrayed as lacking agency. Instead, topics like "Public statements" and "Victim perspective" suggest that narratives acknowledging victim agency are present. Another assumption (A2) I had was that responsibility for sexual violence would be attributed to individuals rather than the government or society. This was only partially supported. While many topics focused on individual perpetrators, other topics indicated that organizational and governmental entities are also seen as responsible. Lastly, in regard for the last assumption (A3), I found support for the presence of sensationalist rhetoric in Russian news media's coverage of sexual violence, with prevalent topics featuring graphic and detailed descriptions of violence. Also, an interesting observation I made was the high visibility of the "Migrants as

perpetrators" topic, which aligns with cultural stereotypes and reflects societal biases in the portrayal of sexual violence.

The ***second research objective of my thesis (RO2)*** was to explore the relationship between the themes present in articles on sexual violence against women and the tone used by Russian news media. To achieve this, I estimated the Pearson linear correlation coefficients between the topic probabilities and the variables indicating the degree of positive, negative, and neutral sentiment in the articles. Several statistically significant correlation coefficients partially supported my hypotheses within RO2.

My first hypothesis (H1) was that themes exploring cases of sexual violence would be correlated with either positive or negative sentiment. Since sensationalism involves emotional coverage and is connected to violent incidents, I included both well-known cases and common incidents of violence in my analysis. I found that certain high-profile cases had significant correlations with sentiment variables. For instance, the Ronaldo case was linked to higher positivity, while the Ufa police department rape case correlated with lower positivity and higher neutrality. However, for other cases, there were no significant correlations. The "Descriptive narratives" topic was associated with higher negative sentiment, while topics related to criminal cases tended to be covered more neutrally. These findings partially support my hypothesis on the emotional coverage of sexual violence cases.

My second hypothesis (H2) was that themes exploring sexual violence as a social problem would be correlated with neutral sentiment. However, I found that topics like "Fighting SV – organizations and funding" and "Problem regulation" were negatively connected to neutral sentiment. The "Reports and statistics" topic was associated with higher positive sentiment. Thus, my hypothesis was not supported, as the media used various sentiments when covering sexual violence as a social problem.

The third hypothesis (H3) anticipated no correlation between other themes and sentiment levels in articles. Contrary to this, I found significant correlations: the "Victim perspective" topic was more likely to be covered negatively, while the "Musicians as perpetrators" topic was covered with both negative and positive emotions. Coverage of

sexual violence in hockey and during the Ukraine conflict tended to be more positive, whereas the "Convention of women's rights" topic had a negative tone. Therefore, my hypothesis was not supported. It's important to note that the significant Pearson correlation coefficients were generally low, indicating weak relationships. Although there were statistically significant correlations, the practical impact on the emotional tone of coverage was minimal. This suggests that, despite finding some sensationalism in specific cases, the overall discussion of sexual violence in Russian news media is not characterized by strong emotional reporting.

The last part of the analysis addressed ***the third research objective (RO3)*** and examines the temporal differences in how sexual violence against women is covered in Russian news media. However, the analysis shows that the topics fluctuate over the study period rather than simply fading away. This challenges the hypothesis (H4) that general themes would decrease in coverage over time. The method of dividing the data by years might oversimplify the temporal patterns, suggesting a need for more nuanced methods in future research. Despite these limitations, the study finds that the main themes in media coverage remain stable over time. This aligns with the idea in framing literature that certain themes persist in discussions over time.

The main research question of the study was: ***how do Russian news media frame the cases of sexual violence against women, as well the sexual violence against women in general?*** In the process on analysis rather than solely focusing on frames and framing, I examined the thematic structure of media coverage. This approach acknowledges the complexity of framing, where frames highlight certain aspects of a text to shape interpretations, which are distinct from themes used in media texts. I adopted an inductive approach to framing research, primarily using topic modeling to identify themes in articles about sexual violence against women.

To guide my analysis, I used Entman's (1993) framing functions, which include problem definition, cause diagnosis, moral judgment, and remedy suggestion. Each function can be linked to specific topics or groups of topics identified in the media coverage. In terms of problem definition, which outlines the issues and their

consequences, I found this function in topics related to criminal cases and legal processes. This signaled the attribution of responsibility frame. It also emerged in topics about social problems, government regulation, reports and statistics, public statements, and the victim's perspective. When looking at cause diagnosis, which identifies the forces behind the problem, I found it to be less apparent. The "Reports and statistics" topic partially addressed this by presenting prevalence rates of sexual violence. The "Migrants as perpetrators" topic suggested an external threat, highlighting ethnicity as a potential cause. For moral judgments, few topics directly indicated this function. The "Compensations for victims" topic implied judgment by addressing economic retribution, though the framing of compensations remained unclear. The "Nobel prize for fighting SV" topic involved evaluating those addressing the problem. Regarding remedy suggestion, this function was present in topics like "Problem regulation," which discussed legal and legislative responses, and "Fighting SV – organizations and funding," which highlighted efforts to combat sexual violence. The "Compensations for victims" topic also suggested remedies through economic retribution.

I also touched on common generic frames in media texts, such as attribution of responsibility, conflict, human interest, and morality. I found that while some topics aligned with these frames, others, like conflict and human interest, were not directly evident. The differentiation between issue-specific and generic frames was noted, with the former pertinent to specific events and the latter applicable across various contexts. I concluded that most topics in the thematic structure fell under issue-specific framing, with only a few providing a broader context through thematic framing.

There are several ***study limitations*** in this thesis. First, the corpus of the articles collected for the analysis is in high reliance to the Medialogia resources which do not store the whole scope of Russian newspapers, therefore providing the research with the potential bias in terms of article selection. For example, there were no independent Russian news media in the corpus that could potentially enrich the thematic structure of the discussion on sexual violence against women or display other patterns of framing the violence. Therefore, the possible direction for the future research is to analyze the

articles of independent Russian newspapers and make comparisons for different types of media outlets in terms of framing sexual violence.

Moreover, the use of topic modeling and sentiment analysis, while powerful, has inherent limitations. Topic modeling relies on word co-occurrences, which may overlook more nuanced or context-specific themes. The possible direction for the future research here is to try non-probabilistic methods for topic modeling, such as BERT sentence embeddings that captures the semantic structure of the texts. Sentiment analysis, particularly automated methods, might not fully capture the depth of emotions expressed in complex narratives. Here, other methods for emotion detection could be implemented, including the manual coding that could provide more accurate insights into the emotional tone of coverage. Moreover, in terms of sentence-level implementation of sentiment analysis used in this thesis, measuring sentiment at the sentence level and approximating it to the article level might introduce inaccuracies. Sentiment can be context-dependent, and therefore methods for emotion detection in the whole articles may be implemented to test correlations between the tone and the thematic structure of the discussion on sexual violence.

Finally, the thematic structure discovered through the topic modeling is rather diverse and complex which limits the model ability to discover framing patterns within the discussion, since there is a possibility, for example, to interpret topics as frames, but only in thematically narrow fields. Therefore, the possibility for the future research here is to take a more thematically narrow corpus, for example, the articles from this corpus that use one of the discovered topics and perform frame detection on them. Moreover, the reliance on automated methods for topic and sentiment analysis, without supplementary manual coding, might miss the depth and complexity of human interpretation. Manual coding of a subset of articles could validate and enrich the findings.

REFERENCES

Abrahams, N., Jewkes, R., Hoffman, M., & Laubsher, R. (2004). Sexual violence against intimate partners in Cape Town: prevalence and risk factors reported by men. *Bulletin of the world health organization*, *82*(5), 330-337.

Alcoff, L. M. (2009). Discourses of sexual violence in a global framework. *Philosophical Topics*, 123-139.

Alcoff, L., & Gray, L. (1993). Survivor discourse: Transgression or recuperation?. Signs: Journal of Women in Culture and Society, 18(2), 260-290.

Archer, J. (2002). Sex differences in physically aggressive acts between heterosexual partners: A meta-analytic review. *Aggression and violent behavior*, *7*(4), 313-351.

Armstrong, E. A., Gleckman-Krut, M., & Johnson, L. (2018). Silence, power, and inequality: An intersectional approach to sexual violence. *Annual Review of Sociology*, *44*, 99-122.

Aroustamian, C. (2020). Time's up: Recognising sexual violence as a public policy issue: A qualitative content analysis of sexual violence cases and the media. Aggression and violent behavior, 50, 101341.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances.

Baranauskas, A. J., & Drakulich, K. M. (2018). Media construction of crime revisited: Media types, consumer contexts, and frames of crime and justice. *Criminology*, *56*(4), 679-714.

Barde, B. V., & Bainwad, A. M. (2017, June). An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 745-750). IEEE.

Bateson, G. (1972). *A theory of play and fantasy* (pp. 177-93). MIT Press. Boston, MA.

Baysha, O., & Hallahan, K. (2004). Media framing of the Ukrainian political crisis, 2000–2001. *Journalism Studies*, *5*(2), 233-246.

Benedict, H. (1993). Virgin or vamp: How the press covers sex crimes. Oxford University Press, USA.

Berinsky, A. J., & Kinder, D. R. (2006). Making sense of issues through media frames: Understanding the Kosovo crisis. *The Journal of Politics*, *68*(3), 640-656.

Bhushan, K., & Singh, P. (2014). The effect of media on domestic violence norms: Evidence from India. The Economics of Peace and Security Journal, 9(1).

Blei, D. M. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, *2*(1), 8-11.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, *18*, 147.

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2003). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, *16*.

Block, S. (2002). Rape and race in colonial newspapers, 1728–1776. Journalism History, 27(4), 146-155.

Boranijašević, M. (2018). Sensationalism in Reporting on Domestic Violence. European Researcher. Series A, (9-3), 194-202.

Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2013, September). Identifying media frames and frame dynamics within and across policy issues. In New Directions in Analyzing Text as Data Workshop, London (pp. 27-28).

Brüggemann, M. (2014). Between frame setting and frame sending: How journalists contribute to news frames. Communication Theory, 24(1), 61-82.

Chaulagain, R. S. et al. (2017, November). Cloud based web scraping for big data applications. *In 2017 IEEE International Conference on Smart Cloud (SmartCloud)* (pp. 138-143). IEEE.

Cleere, C., & Lynn, S. J. (2013). Acknowledged versus unacknowledged sexual assault among college women. *Journal of interpersonal violence*, *28*(12), 2593-2611.

Cullen, P., O'Brien, A., & Corcoran, M. (2019). Reporting on domestic violence in the Irish media: An exploratory study of journalists' perceptions and practices. Media, Culture & Society, 41(6), 774-790.

Čvorović, D. S. (2022). Latent crime and police statistics – the role of sence of security in the lae enforcement work. *CRIMEN-časopis za krivične nauke*, *13*(3), 247-283.

D'angelo, P. (2002). News framing as a multiparadigmatic research program: A response to Entman. *Journal of communication*, *52*(4), 870-888.

Dartnall, E., & Jewkes, R. (2013). Sexual violence against women: the scope of the problem. *Best practice & research Clinical obstetrics & gynaecology*, *27*(1), 3-13.

De Vreese, C. H. (2005). News framing: Theory and typology. Information design journal+ document design, 13(1), 51-62.

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. Procedia Computer Science, 87, 44-49.

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

Easteal, P., Holland, K., & Judd, K. (2015, January). Enduring themes and silences in media portrayals of violence against women. In *Women's Studies International Forum* (Vol. 48, pp. 103-113). Pergamon.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, *43*(4), 51-58.

Entman, R. M. (2004). Projections of power: Framing news, public opinion, and US foreign policy. University of Chicago Press.

Entman, R. M., Matthes, J., & Pellicano, L. (2009). Nature, sources, and effects of news framing. In *The handbook of journalism studies* (pp. 195-210). Routledge.

Fisher, B. S., Daigle, L. E., Cullen, F. T., & Turner, M. G. (2003). Acknowledging sexual victimization as rape: Results from a national-level study. *Justice Quarterly*, *20*(3), 535-574.

Fomin, I., & Nadskakuła-Kaczmarczyk, O. (2022). Against Putin and Corruption, for Navalny and the "Revolution"? The Dynamics of Framing and Mobilization in the Russian Political Protests of 2017–18. *Communist and Post-Communist Studies*, *55*(1), 99-130.

Frieze, I. H. (2005). Female violence against intimate partners: An introduction. *Psychology of Women Quarterly*, *29*(3), 229-237.

Gamson, W. A., & Modigliani, A. (1987). The changing culture of affirmative action. *Research in political sociology*, *3*(1), 137-177.

Gitlin, T. (2003). *The whole world is watching: Mass media in the making and unmaking of the new left*. Univ of California Press.

Goffman E. Frame analysis: An essay on the organization of experience. – Harvard University Press, 1974.

Graham, K., Bernards, S., Abbey, A., Dumas, T. M., & Wells, S. (2017). When women do not want it: Young female bargoers' experiences with and responses to sexual harassment in social drinking contexts. *Violence against women*, *23*(12), 1419-1441.

Gruenewald, J., Chermak, S. M., & Pizarro, J. M. (2013). Covering victims in the news: What makes minority homicides newsworthy? *Justice Quarterly*, *30*(5), 755-783.

Guenther, L., Jörges, S., Mahl, D., & Brüggemann, M. (2023). Framing as a bridging concept for climate change communication: A systematic review based on 25 years of literature. *Communication Research*, 00936502221137165.

Günther, E., & Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. *International journal of communication*, *11*, 21.

Günther, E., & Quandt, T. (2018). Word counts and topic models: Automated text analysis methods for digital journalism research. In *Rethinking research methods in an age of digital journalism* (pp. 75-88). Routledge.

Günther, E., & Quandt, T. (2018). Word counts and topic models: Automated text analysis methods for digital journalism research. In *Rethinking research methods in an age of digital journalism* (pp. 75-88). Routledge.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332-359.

Hamby, S. (2014). Intimate partner and sexual violence research: Scientific progress, scientific challenges, and gender. *Trauma, Violence, & Abuse*, *15*(3), 149-158.

Harmer, E., & Lewis, S. (2020). Disbelief and counter-voices: a thematic analysis of online reader comments about sexual harassment and sexual violence against women. *Information, Communication & Society*, *25*(2), 199-216.

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020* (pp. 341-354). Springer Singapore.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, *25*(1), 114-146.

Hindes, S., & Fileborn, B. (2020). "Girl power gone wrong":# MeToo, Aziz Ansari, and media reporting of (grey area) sexual violence. Feminist Media Studies, 20(5), 639-656.

Hollander, J. A. (2001). Vulnerability and dangerousness: The construction of gender through conversation about violence. *Gender & society*, *15*(1), 83-109.

Hollander, J. A., & Rodgers, K. (2014, June). Constructing victims: The erasure of women's resistance to sexual assault. In *Sociological Forum* (Vol. 29, No. 2, pp. 342-364).

Iyengar, S. (1996). Framing responsibility for political issues. *The Annals of the American Academy of Political and Social Science*, *546*(1), 59-70.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American psychologist*, *39*(4), 341.

Kalra, G., & Bhugra, D. (2013). Sexual violence against women: Understanding cross-cultural intersections. *Indian journal of psychiatry*, *55*(3), 244-249.

Khomsah, S., Ramadhani, R. D., & Wijaya, S. (2022). The accuracy comparison between Word2Vec and FastText On sentiment analysis of hotel reviews. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 6(3), 352-358.

Kinder, D. R., & Sanders, L. M. (1990). Mimicking political debate with survey questions: The case of white opinion on affirmative action for blacks. *Social cognition*, *8*(1), 73-103.

Koltcov, S., Ignatenko, V., Terpilovskii, M., & Rosso, P. (2021). Analysis and tuning of hierarchical topic models based on Renyi entropy approach. *PeerJ Computer Science*, *7*, e608.

Kort-Butler, L. A., & Habecker, P. (2018). Framing and cultivating the story of crime: The effects of media use, victimization, and social networks on attitudes about crime. *Criminal justice review*, *43*(2), 127-146.

Kort-Butler, L. A., & Hartshorn, K. J. S. (2011). Watching the detectives: Crime programming, fear of crime, and attitudes about the criminal justice system. *The Sociological Quarterly*, *52*(1), 36-55.

Krug, E.G. et al., eds (2002). World report on violence and health. *World Health Organization,* Geneva.

Lakatos, I. (1974). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91-198). Cambridge, UK: Cambridge University Press.

Lemish, D. (2004). Exclusion and marginality: Portrayals of women in Israeli media. Women and media: International perspectives, 39-59.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning* (pp. 577-584).

Marshall, L. L. (1992a). Development of the severity of violence against women scales. *Journal of family violence*, *7*, 103-121.

Marshall, L. L. (1992b). The severity of violence against men scales. *Journal of Family Violence*, *7*, 189-203.

Matthes, J. (2009). What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & mass communication quarterly*, *86*(2), 349-367.

Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of communication*, *58*(2), 258-279.

McMurray, A. (2005). Domestic violence: conceptual and practice issues. *Contemporary Nurse*, *18*(3), 219-232.

McQuail, D. (1994). *Mass communication theory: An introduction* (2nd ed.). Sage Publications, Inc.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal,* 5(4), 1093-1113.

Miller, M. M. (1997). Frame mapping and analysis of news coverage of contentious issues. *Social science computer review*, *15*(4), 367-378.

Mimno, D., Li, W., & McCallum, A. (2007, June). Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning* (pp. 633-640).

O'hara, S. (2012). Monsters, playboys, virgins and whores: Rape myths in the news media's coverage of sexual violence. *Language and literature*, *21*(3), 247-259.

O'Hear, M. (2020). Violent crime and media coverage in one city: A statistical snapshot. *Marq. L. Rev.*, *103*, 1007.

Planty, M., Langton, L., Krebs, C., Berzofsky, M., & Smiley-McDonald, H. (2013). *Female victims of sexual violence, 1994-2010* (pp. 3-4). Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Pollak, J., & Kubrin, C. E. (2007). Crime in the news: How crimes, offenders and victims are portrayed in the media. *Journal of Criminal Justice and Popular Culture*, *14*, 59-83.

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)* (pp. 200-204). IEEE.

Ricardo, C., & Barker, G. (2008). Men, masculinities, sexual exploitation and sexual violence: A literature review and call for action. *Rio de Janeiro: Promundo and MenEngage*, 1-50.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Russell, W. (2007). Sexual violence against men and boys. *Forced Migration Review*, *27*, 22-23.

Russo, N. F., & Pirlott, A. (2006). Gender-based violence: concepts, methods, and findings. *Annals of the new york academy of sciences*, *1087*(1), 178-205.

Ryan, C., Anastario, M., & DaCunha, A. (2006). Changing coverage of domestic violence murders: A longitudinal experiment in participatory communication. Journal of interpersonal violence, 21(2), 209-228.

Saldaña Villa, M. C. (2017). Framing disaster: a topic modeling approach for the case of Chile (*Doctoral dissertation)*.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, *49*(1), 103-122.

Scheufele, D. A., & Tewksbury, D. (2007). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, *57*(1), 9-20.

Schwartz, I. L. (1991). Sexual violence against women: prevalence, consequences, societal factors, and prevention. *American journal of preventive medicine*, *7*(6), 363-373.

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of communication*, *50*(2), 93-109.

Serisier, T. (2017). Sex crimes and the media. In Oxford research encyclopedia of criminology and criminal justice.

Shahin, S. (2016). Right to be forgotten: How national identity, political orientation, and capitalist ideology structured a trans-Atlantic debate on information access and control. *Journalism & Mass Communication Quarterly*, *93*(2), 360-382.

Shoemaker, P. J., & Reese, S. D. (2014). Mediating the message in the 21st century: A media sociology perspective. *Third edition published 2014 by Routledge*.

Skjelsbaek, I. (2001). Sexual violence and war: Mapping out a complex relationship. European journal of international relations, 7(2), 211-237.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952-961).

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. Annual Review of Linguistics, 2, 325-347.

United Nations (1995). Report of the Fourth World Conference on Women. *United Nations, 4-15 September 1995*, 223 p.

Valkenburg, P. M., Peter, J., & Walther, J. B. (2016). Media effects: Theory and research. *Annual review of psychology*, *67*, 315-338.

Wahl-Jorgensen, K., & Hanitzsch, T. (2009). Introduction: On why and how we should do journalism studies. In *The handbook of journalism studies* (pp. 23-36). Routledge.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

Zalewski, M., Drumond, P., Prügl, E., & Stern, M. (2018). Introduction: sexual violence against men in global politics. In *Sexual violence against men in global politics* (pp. 1-19). Routledge.

Вахштайн, В. (2008). «Практика» vs. «фрейм»: альтернативные проекты исследования повседневного мира. *Социологическое обозрение*, *7*(1), 65-95.

Иншаков, С. М. (2009). Латентная преступность как объект исследования. *Криминология: вчера, сегодня, завтра*, (16), 107-130.

Серебренников Д., & Титаев К. (2022). Динамика преступности и виктимизации в России 2018–2021 гг. Результаты второго виктимизационного опроса: аналитический обзор. *Институт проблем правоприменения при Европейском университете в Санкт-Петербурге (Аналитические обзоры по проблемам правоприменения),* 2(2022), 34 с.

Соколов, М. М. (2022). И. Гофман, каким мы его помним. *Социологические исследования*, (6), 16-22.8

OTHER RESOURCES

Data collection and analysis materials

To systematize the materials used in the process of collecting and analyzing data, I created a repository on Github which contains the *.ipynb* files with the code for data collection and analysis, the *.txt* files containing stopwords and other filtering lists, the final dataset in *.xlsx* and *.sav* format and the pretrained FastText social network model binary file used in sentiment analysis.

The link to the repository: https://github.com/kmlapshina/bachelor-thesis

Footnote citations

This section contains a list of all sources and references that have been mentioned in footnotes (in order of citation). The citations are the following:

1. Bill № 1183390-6 “On the Prevention of Domestic Violence in the Russian Federation” (in Russian). SOZD State Duma website. URL: https://sozd.duma.gov.ru/bill/1183390-6
2. Federal Law "On Amendments to Article 116 of the Criminal Code of the Russian Federation" dated 02/07/2017 N 8-FZ URL: https://www.consultant.ru/document/cons_doc_LAW_212385/
3. Оксана Пушкина впишет харассмент в закон на фоне скандала со Слуцким. РБК, 27.02.2018. URL: https://www.rbc.ru/politics/27/02/2018/5a942e6a9a79471333d3128a
4. Депутат Госдумы разрабатывает законопроект, предусматривающий лишение свободы за харассмент. Комменсантъ, 18.06.2019. URL: https://www.kommersant.ru/doc/4004333
5. National Crime Victimization Survey (NCVS). URL: https://bjs.ojp.gov/data-collection/ncvs
6. Crime Survey for England and Wales (CSEW). URL: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingseptember2023#related-links
7. Medialogia methodology and technology description. URL: https://www.mlg.ru/about/technologies/
8. The implementation of topic models in Python with different parameters settings. URL: https://ethen8181.github.io/machine-learning/clustering/topic_model/LDA.html
9. The performance of Tomotopy library. URL: https://bab2min.github.io/tomotopy/v0.12.6/en/#performance-of-tomotopy
10. PunkSentenceTokemizer documentation. URL: https://www.nltk.org/api/nltk.tokenize.PunktSentenceTokenizer.html

APPENDIX

Appendix 1. Full list of mass media sources

Table A1. Mass media sources remaining after data preparation for automatic data collection

Newspaper name	Article count
Комсомольская правда (kp.ru)	1465
Газета.Ru	1085
Lenta.Ru	1074
ИА Regnum	864
Московский Комсомолец (mk.ru)	708
Life.ru	691
ТАСС	645
РИА Новости	619
ИА Росбалт	511
VSE42 (vse42.ru)	461
Известия (iz.ru)	420
ИА Красная весна	407
NEWS.ru	384
Российская газета (rg.ru)	338
Sports.ru	333
Фонтанка (fontanka.ru)	332
Взгляд.Ру	323
Общественная служба новостей (osnmedia.ru)	290
ФедералПресс (fedpress.ru)	276
ИА Ura.ru	274
Аргументы и Факты (aif.ru)	260
RT (russian.rt.com)	252
Чемпионат.com (championat.com)	237
Ридус (ridus.ru)	233
РБК (rbc.ru)	211
Свободная пресса (svpressa.ru)	190
Дни.Ру	182
78.ru	166
Правда.ру (pravda.ru)	163
Аргументы недели (argumenti.ru)	145
Р-Спорт	140
ИА SM-News	128
Новые Известия (newizv.ru)	123
Коммерсантъ. Новости информ. центра	120
Eadaily.com	119

47 Новостей из Ленинградской области (47news.ru)	114
Национальная Служба Новостей	102
V1.ru	90
Екатеринбург Он-лайн (e1.ru)	87
59.ru	78
Бизнес Online (business-gazeta.ru)	70
72.ru	61
Женский журнал Woman.ru	61
Коммерсантъ. Новости Online	59
ИА Татар-информ (tatar-inform.ru)	58
Комсомольская правда (msk.kp.ru)	56
Star Hit (starhit.ru)	50
Подмосковье сегодня (mosregtoday.ru)	42
Ведомости (vedomosti.ru)	38
РИАМО (riamo.ru)	36
Утро.ру (utro.ru)	35
Евро-футбол.ru (euro-football.ru)	27
Нижегородская правда (pravda-nn.ru)	26
Спорт день за днем (sportsdaily.ru)	16
Федеральные новости (fednews.ru)	15
ИА Фергана	12
ИА Stringer	11
Ведомости	7
ИА OnAir.ru	5
ИА Инфоповод	5
Подъём (pdmnews.ru)	5
ПРАЙМ	4
Завтра	1
ИА Реалист	1
Коммерсантъ (kommersant.ru)	1

Appendix 2. Scraping and parsing functions used in automatic textual data collection

```
def rg_scraper(link):
    try:
        try:
          r = requests.get(link, timeout=20)
        except requests.exceptions.RequestException as ex:
          print(ex)

        soup = BeautifulSoup(r.content, "html.parser")

        headline = soup.find('meta', property="og:title").get("content")
        description = soup.find('meta', property="og:description").get("content")
        article_body = ''

        blocks = soup.find_all('p')
        for block in blocks:
          article_body = article_body + block.get_text()

        return(headline, description, article_body)

    except (UnboundLocalError, KeyError, AttributeError, UnicodeDecodeError) as error:
      print(error)
      pass
```

Figure 81. An example of a scraping function (for 'ИА Регнум' media outlet) from the list of 68 functions used in data collection

```
def parser(links, scraper):

    headlines = []
    descriptions = []
    article_bodies = []
    urls = []

    for link in links:
      try:
        h, d, a = scraper(link)
        headlines.append(h)
        descriptions.append(d)
        article_bodies.append(a)
        urls.append(link)

      except TypeError:
        pass

    return(headlines, descriptions, article_bodies, urls)
```

Figure 82. The parsing function used in data collection

Appendix 3. The results of the LDA-based models fitting

Table A2. The results of the LDA-based models fitting

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
1	LDAModel		5	0.001		1,00E-05		-1.5778704536662338	-8.635658966522472
2	LDAModel		10	0.001		1,00E-05		-1.6150202463664542	-8.757591297037758
3	LDAModel		15	0.001		1,00E-05		-1.8202113336355197	-8.811911416973665
4	LDAModel		20	0.001		1,00E-05		-2.0356906049040804	-8.797220451121815
5	LDAModel		25	0.001		1,00E-05		-2.1065588985823833	-8.821437682624607
6	LDAModel		30	0.001		1,00E-05		-2.249233671242091	-8.840772661521282
7	LDAModel		40	0.001		1,00E-05		-2.1779757385093896	-8.852659167277995
8	LDAModel		50	0.001		1,00E-05		-2.4852130486588138	-8.883684908170544
9	LDAModel		60	0.001		1,00E-05		-2.481226781262313	-8.886854973213087
10	LDAModel		70	0.001		1,00E-05		-2.573124674579185	-8.894666750021873
11	LDAModel		5	0.001		0.001		-1.5523904881325827	-8.472779074368527
12	LDAModel		10	0.001		0.001		-1.5870279302961026	-8.544953316705541
13	LDAModel		15	0.001		0.001		-1.7994505358763024	-8.573776691891478
14	LDAModel		20	0.001		0.001		-2.0079168853863534	-8.569689985326049
15	LDAModel		25	0.001		0.001		-1.9932517938279621	-8.570793370328591
16	LDAModel		30	0.001		0.001		-2.059288428705218	-8.575843110712125
17	LDAModel		40	0.001		0.001		-2.105927015193695	-8.567319747423365
18	LDAModel		50	0.001		0.001		-2.4168429306596417	-8.580246738404442
19	LDAModel		60	0.001		0.001		-2.6366868653010775	-8.571006882090405
20	LDAModel		70	0.001		0.001		-2.560320742137253	-8.579423190354085
21	LDAModel		5	0.001		0.1		-1.5980085138098141	-8.2822601139218
22	LDAModel		10	0.001		0.1		-1.558604246140629	-8.348436362858475
23	LDAModel		15	0.001		0.1		-1.9498255277364958	-8.364008529695955
24	LDAModel		20	0.001		0.1		-2.0111475013555653	-8.388617663070569
25	LDAModel		25	0.001		0.1		-2.0627645459105803	-8.368329945181145
26	LDAModel		30	0.001		0.1		-2.275371637227306	-8.382837687644582

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
27	LDAModel		40	0.001		0.1		-2.2693267806747643	-8.421628471238396
28	LDAModel		50	0.001		0.1		-2.8746160549158053	-8.405012608432108
29	LDAModel		60	0.001		0.1		-3.344010215679805	-8.447200840021381
30	LDAModel		70	0.001		0.1		-4.806882430560782	-8.420369728178326
31	LDAModel		5	0.001		0.5		-1.6293086583982188	-8.351664762955897
32	LDAModel		10	0.001		0.5		-1.5929880197671635	-8.477491863658363
33	LDAModel		15	0.001		0.5		-2.722286827878061	-8.491011162953157
34	LDAModel		20	0.001		0.5		-4.266550997649235	-8.505240449949731
35	LDAModel		25	0.001		0.5		-6.244982635239394	-8.478182311130443
36	LDAModel		30	0.001		0.5		-6.956871311185805	-8.474591562440265
37	LDAModel		40	0.001		0.5		-7.4861658446932555	-8.51034186036514
38	LDAModel		50	0.001		0.5		-8.304049620831618	-8.501205892852658
39	LDAModel		60	0.001		0.5		-6.098122334153765	-8.566145573582116
40	LDAModel		70	0.001		0.5		-6.797351412531287	-8.525585818912928
41	LDAModel		5	0.001		1		-1.5568070781628878	-8.433136008390639
42	LDAModel		10	0.001		1		-1.490926863053502	-8.573678756177065
43	LDAModel		15	0.001		1		-5.097671259515105	-8.601549856575796
44	LDAModel		20	0.001		1		-5.802284691655415	-8.577950968236415
45	LDAModel		25	0.001		1		-7.111044227292286	-8.472078970405635
46	LDAModel		30	0.001		1		-5.586591287790214	-8.540049014176532
47	LDAModel		40	0.001		1		-3.9691006117419603	-8.531266573900137
48	LDAModel		50	0.001		1		-4.721471261986072	-8.446052296638573
49	LDAModel		60	0.001		1		-3.763458273210867	-8.515007734238637
50	LDAModel		70	0.001		1		-3.8833458617437495	-8.510968676342912
51	HLDAModel	3	881	0.001		1,00E-05	0.001	-4.95346193279525	-8.561474159845956
52	HLDAModel	3	1375	0.001		0.0001	0.001	-5.455201629823559	-8.413965696298279
53	HLDAModel	3	960	0.001		0.001	0.001	-8.328110055467949	-8.247043004052317
54	HLDAModel	3	336	0.001		0.01	0.001	-8.37014604743548	-8.137927008105041
55	HLDAModel	3	29	0.001		0.1	0.001	-8.596187348217391	-8.315819189911494
56	HLDAModel	3	12	0.001		0.2	0.001	-4.5705310998944	-8.32406233997504
57	HLDAModel	3	9	0.001		0.3	0.001	-3.7066480454047337	-8.334500160660339

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
58	HLDAModel	3	4	0.001		0.5	0.001	-2.0510542102763782	-8.358642082221667
59	HLDAModel	3	4	0.001		0.7	0.001	-1.877557132572644	-8.382657785562117
60	HLDAModel	3	4	0.001		1	0.001	-1.3551410899687797	-8.416453023365115
61	PAModel	3	5	0.001	0.001	1,00E-05		-1.693924548904438	-9.43618050362715
62	PAModel	3	10	0.001	0.001	1,00E-05		-1.6347082202769236	-9.638879753562716
63	PAModel	3	15	0.001	0.001	1,00E-05		-1.3326771745415817	-9.856381752168042
64	PAModel	3	20	0.001	0.001	1,00E-05		-2.0508409196502018	-9.972859616301053
65	PAModel	3	25	0.001	0.001	1,00E-05		-2.3216488161119693	-10.059850595342018
66	PAModel	3	30	0.001	0.001	1,00E-05		-2.05041857051901	-10.10020798154384
67	PAModel	3	40	0.001	0.001	1,00E-05		-2.8709245852078484	-10.203650821787003
68	PAModel	3	50	0.001	0.001	1,00E-05		-2.7292138563321835	-10.270324869285593
69	PAModel	3	60	0.001	0.001	1,00E-05		-2.4378883802993077	-10.349612015412774
70	PAModel	3	70	0.001	0.001	1,00E-05		-3.606005751263108	-10.399271578242335
71	PAModel	3	5	0.001	0.001	0.001		-1.6166646487113088	-9.384430013398955
72	PAModel	3	10	0.001	0.001	0.001		-1.5661175280579223	-9.63078151460463
73	PAModel	3	15	0.001	0.001	0.001		-1.4778668069053023	-9.847484753227613
74	PAModel	3	20	0.001	0.001	0.001		-1.9585855194278807	-10.017500774254884
75	PAModel	3	25	0.001	0.001	0.001		-1.7059222381691146	-10.076428814426095
76	PAModel	3	30	0.001	0.001	0.001		-2.0074187331748057	-10.158128715941883
77	PAModel	3	40	0.001	0.001	0.001		-2.8986958165704872	-10.28370112746474
78	PAModel	3	50	0.001	0.001	0.001		-2.659819610460391	-10.38921097396021
79	PAModel	3	60	0.001	0.001	0.001		-2.458298649118877	-10.42045204853434
80	PAModel	3	70	0.001	0.001	0.001		-3.747778310133531	-10.526100645710168
81	PAModel	3	5	0.001	0.001	0.1		-1.6771907954150256	-9.613188421233517
82	PAModel	3	10	0.001	0.001	0.1		-1.662356202934505	-9.962177660914612
83	PAModel	3	15	0.001	0.001	0.1		-2.4899936389468746	-10.245933864298534
84	PAModel	3	20	0.001	0.001	0.1		-1.7607672541257091	-10.422175852788355
85	PAModel	3	25	0.001	0.001	0.1		-2.1477246674774664	-10.668185347655216
86	PAModel	3	30	0.001	0.001	0.1		-1.2457179301339154	-10.646885743101738
87	PAModel	3	40	0.001	0.001	0.1		-5.62473228738892	-10.92079191714506
88	PAModel	3	50	0.001	0.001	0.1		-3.2267127485393843	-10.878281635302542

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
89	PAModel	3	60	0.001	0.001	0.1		-2.782533649185984	-10.995413725907193
90	PAModel	3	70	0.001	0.001	0.1		-10.526352692466373	-10.952424697370889
91	PAModel	3	5	0.001	0.001	0.5		-1.6790862291890896	-9.727786418109835
92	PAModel	3	10	0.001	0.001	0.5		-2.898434257701337	-9.87132383697566
93	PAModel	3	15	0.001	0.001	0.5		-1.2534100402663058	-10.287369551568238
94	PAModel	3	20	0.001	0.001	0.5		-7.139555196213538	-10.298060158207477
95	PAModel	3	25	0.001	0.001	0.5		-12.908524619286176	-10.25072623560403
96	PAModel	3	30	0.001	0.001	0.5		-7.202634977359708	-10.397910518222195
97	PAModel	3	40	0.001	0.001	0.5		-5.566617928300384	-10.222947704370734
98	PAModel	3	50	0.001	0.001	0.5		-12.105459872219773	-10.39035762128202
99	PAModel	3	60	0.001	0.001	0.5		-12.001725042006315	-10.311507450640216
100	PAModel	3	70	0.001	0.001	0.5		-5.257086650653572	-10.363284636968618
101	PAModel	3	5	0.001	0.001	1		-1.4519754249647523	-9.662305618139765
102	PAModel	3	10	0.001	0.001	1		-1.2718038364140059	-9.83193860466609
103	PAModel	3	15	0.001	0.001	1		-5.284513522300912	-9.843165804113445
104	PAModel	3	20	0.001	0.001	1		-4.630990703897102	-9.902308693910632
105	PAModel	3	25	0.001	0.001	1		-15.295230896619842	-9.909392015994927
106	PAModel	3	30	0.001	0.001	1		-5.680941877103543	-9.933470051924456
107	PAModel	3	40	0.001	0.001	1		-2.548368476620662	-9.947887727217893
108	PAModel	3	50	0.001	0.001	1		-14.042387308915403	-9.92678481138964
109	PAModel	3	60	0.001	0.001	1		-6.7450133756963915	-9.728215737649055
110	PAModel	3	70	0.001	0.001	1		-8.28718380431198	-9.92103714907374
111	HPAModel	3	5	0.001	0.001	1,00E-05		-1.7166106176095033	-9.48715533413469
112	HPAModel	3	10	0.001	0.001	1,00E-05		-1.5219461066079043	-9.737822036812457
113	HPAModel	3	15	0.001	0.001	1,00E-05		-1.494696404089651	-9.534613833575532
114	HPAModel	3	20	0.001	0.001	1,00E-05		-1.6416898678545426	-9.700785679072204
115	HPAModel	3	25	0.001	0.001	1,00E-05		-1.4628875033142756	-9.771868967415864
116	HPAModel	3	30	0.001	0.001	1,00E-05		-1.6116807532734647	-9.887664952794681
117	HPAModel	3	40	0.001	0.001	1,00E-05		-1.5920919168807321	-9.887510956223942
118	HPAModel	3	50	0.001	0.001	1,00E-05		-1.7647903945005492	-10.024461190737584
119	HPAModel	3	60	0.001	0.001	1,00E-05		-1.6841283096321342	-10.118593997705242

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
120	HPAModel	3	70	0.001	0.001	1,00E-05		-1.6348767672841433	-10.077552235306838
121	HPAModel	3	5	0.001	0.001	0.001		-1.659650335752553	-9.367138529403842
122	HPAModel	3	10	0.001	0.001	0.001		-1.3256700014820744	-9.499260256244044
123	HPAModel	3	15	0.001	0.001	0.001		-1.4540674734728352	-9.472839794542137
124	HPAModel	3	20	0.001	0.001	0.001		-1.5521654660995414	-9.498943238818065
125	HPAModel	3	25	0.001	0.001	0.001		-1.353156543145807	-9.58986869208762
126	HPAModel	3	30	0.001	0.001	0.001		-1.4776491672756806	-9.650574740016207
127	HPAModel	3	40	0.001	0.001	0.001		-1.5126929218799	-9.801733961794541
128	HPAModel	3	50	0.001	0.001	0.001		-1.2953333188568883	-9.89172213072634
129	HPAModel	3	60	0.001	0.001	0.001		-1.542748415923656	-9.89732284336646
130	HPAModel	3	70	0.001	0.001	0.001		-1.5215975512966464	-9.88632601223671
131	HPAModel	3	5	0.001	0.001	0.1		-1.3499211530822077	-9.372328034936775
132	HPAModel	3	10	0.001	0.001	0.1		-1.3406191752436871	-9.484931235550535
133	HPAModel	3	15	0.001	0.001	0.1		-1.312841538779882	-9.778869460577456
134	HPAModel	3	20	0.001	0.001	0.1		-1.3212570763814524	-9.576166683254833
135	HPAModel	3	25	0.001	0.001	0.1		-1.3072488732739878	-9.504060927078278
136	HPAModel	3	30	0.001	0.001	0.1		-1.317378126199185	-9.457305785562001
137	HPAModel	3	40	0.001	0.001	0.1		-1.3359393643498414	-9.513761825854697
138	HPAModel	3	50	0.001	0.001	0.1		-1.3053333188568883	-9.440748825795408
139	HPAModel	3	60	0.001	0.001	0.1		-1.3161443686914842	-9.448982243394065
140	HPAModel	3	70	0.001	0.001	0.1		-1.475863047313614	-9.427024042908416
141	HPAModel	3	5	0.001	0.001	0.5		-1.4564265448259954	-8.831634106379108
142	HPAModel	3	10	0.001	0.001	0.5		-1.4401268897730413	-8.552212398729345
143	HPAModel	3	15	0.001	0.001	0.5		-1.461333181924677	-8.50855539538572
144	HPAModel	3	20	0.001	0.001	0.5		-1.5216785283586853	-8.502269997216088
145	HPAModel	3	25	0.001	0.001	0.5		-1.490074614377254	-8.500120916395117
146	HPAModel	3	30	0.001	0.001	0.5		-1.48828525338341	-8.502279853990919
147	HPAModel	3	40	0.001	0.001	0.5		-1.4479190317538222	-8.504751383811394
148	HPAModel	3	50	0.001	0.001	0.5		-1.4969372158092256	-8.484627258334207
149	HPAModel	3	60	0.001	0.001	0.5		-1.5359670566605512	-8.488390213120818
150	HPAModel	3	70	0.001	0.001	0.5		-1.4884737931594385	-8.485068564802413

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
151	HPAModel	3	5	0.001	0.001	1		-1.4575600550383427	-8.475046272476133
152	HPAModel	3	10	0.001	0.001	1		-1.6772502565997864	-8.467601983112688
153	HPAModel	3	15	0.001	0.001	1		-1.5957461029113909	-8.459955650291805
154	HPAModel	3	20	0.001	0.001	1		-1.5129572788017172	-8.440166915342482
155	HPAModel	3	25	0.001	0.001	1		-1.5017607863822136	-8.441786851432079
156	HPAModel	3	30	0.001	0.001	1		-1.5075885523882306	-8.437982304782786
157	HPAModel	3	40	0.001	0.001	1		-1.4101530689873847	-8.45219408136051
158	HPAModel	3	50	0.001	0.001	1		-1.4885690643860174	-8.440185046158453
159	HPAModel	3	60	0.001	0.001	1		-1.4849931922656066	-8.446290175826835
160	HPAModel	3	70	0.001	0.001	1		-1.504292859228725	-8.435324312146854
161	CTModel		5	0.001		1,00E-05		-1.4990200278019832	-7.751767918944299
162	CTModel		10	0.001		1,00E-05		-1.806794252231184	-7.025322910526604
163	CTModel		15	0.001		1,00E-05		-2.0122373236563513	-6.894014097727763
164	CTModel		20	0.001		1,00E-05		-1.980441401305385	-6.934330040592351
165	CTModel		25	0.001		1,00E-05		-1.9829415912108233	-6.953753125050885
166	CTModel		30	0.001		1,00E-05		-2.109993554523069	-6.938384426407668
167	CTModel		40	0.001		1,00E-05		-2.0860893452838605	-7.00473608631542
168	CTModel		50	0.001		1,00E-05		-2.159752223911771	-7.070072091787518
169	CTModel		60	0.001		1,00E-05		-2.490605405946877	-6.84580868339799
170	CTModel		70	0.001		1,00E-05		-2.2991278383421943	-7.252303129972673
171	CTModel		5	0.001		0.001		-1.5460362361925717	-7.623363110301734
172	CTModel		10	0.001		0.001		-1.8397297812829705	-6.955520891459149
173	CTModel		15	0.001		0.001		-1.8943655917963227	-6.829794824151643
174	CTModel		20	0.001		0.001		-2.004919845335418	-6.783703371819858
175	CTModel		25	0.001		0.001		-2.139214724760562	-6.644929805722942
176	CTModel		30	0.001		0.001		-2.017718726885968	-6.878577778066669
177	CTModel		40	0.001		0.001		-2.3406878722230333	-6.629884587044229
178	CTModel		50	0.001		0.001		-2.2553150625393807	-6.94871047653291
179	CTModel		60	0.001		0.001		-2.328317597882233	-7.018473959235517
180	CTModel		70	0.001		0.001		-2.3122555603346813	-7.14781385720421
181	CTModel		5	0.001		0.1		-1.5506068067714363	-7.735565032513108

№	Model	Depth (levels)	Num of topics	Alpha	Subalpha	Eta	Gamma	Coherence	Log-likelihood
182	CTModel		10	0.001		0.1		-1.8800235853936855	-7.160119940629908
183	CTModel		15	0.001		0.1		-2.0538638212483025	-7.096873395261096
184	CTModel		20	0.001		0.1		-1.939229117000029	-7.116046479372132
185	CTModel		25	0.001		0.1		-2.137612399516291	-7.169483855625088
186	CTModel		30	0.001		0.1		-2.2927439582704943	-6.962219595940526
187	CTModel		40	0.001		0.1		-2.289108604916568	-7.327765861182165
188	CTModel		50	0.001		0.1		-2.4950639864588076	-7.370709393133022
189	CTModel		60	0.001		0.1		-2.5315186103822622	-7.558671386038707
190	CTModel		70	0.001		0.1		-3.0911716553796285	-7.381164215077652
191	CTModel		5	0.001		0.5		-1.4562519565059382	-8.04699631497335
192	CTModel		10	0.001		0.5		-1.789288980418672	-7.566968124770986
193	CTModel		15	0.001		0.5		-2.0629168501428174	-7.525740063531037
194	CTModel		20	0.001		0.5		-2.0104216226860574	-7.613229246283313
195	CTModel		25	0.001		0.5		-2.132439670357707	-7.736151441339753
196	CTModel		30	0.001		0.5		-2.317697953663151	-7.804722072449981
197	CTModel		40	0.001		0.5		-3.0059352050475696	-7.856750803993967
198	CTModel		50	0.001		0.5		-3.9311870318474065	-8.100579191096132
199	CTModel		60	0.001		0.5		-6.388555476043066	-7.907294962772885
200	CTModel		70	0.001		0.5		-7.9167742864937365	-8.00228658684089
201	CTModel		5	0.001		1		-1.5096120990748223	-8.22849453356763
202	CTModel		10	0.001		1		-1.8099282255655553	-7.8793611794859055
203	CTModel		15	0.001		1		-2.091198721614256	-7.885118148307272
204	CTModel		20	0.001		1		-2.034149471676556	-7.97434473674668
205	CTModel		25	0.001		1		-2.312396009384724	-8.06037764912183
206	CTModel		30	0.001		1		-2.8350772569375007	-8.131751554082872
207	CTModel		40	0.001		1		-4.863277032337969	-8.345074230060982
208	CTModel		50	0.001		1		-7.125356505003726	-8.430884835352295
209	CTModel		60	0.001		1		-7.684632845704978	-8.492313836804023
210	CTModel		70	0.001		1		-8.36196736330942	-8.54927607343145

Appendix 4. Descriptive statistics and normality tests for the topic probability variables

Table A3. Descriptive statistics for the topic probability variables

ID	Topic name	N	Minimum	Maximum	Mean	Std. Deviation
top	Rape and assault cases	15342	0	0,981	0,247	0,264
sup0	Victim perspective	15342	0	0,658	0,056	0,085
sup1	Descriptive narratives	15342	0	0,448	0,060	0,067
sup2	Problem regulation	15342	0	0,717	0,080	0,098
0	Public statements	15342	0	0,454	0,010	0,035
1	Royal family scandals	15342	0	0,451	0,005	0,027
2	Violence in hockey	15342	0	0,774	0,016	0,064
3	Khachaturian sisters case	15342	0	0,775	0,009	0,043
4	Migrants as perpetrators	15342	0	0,571	0,008	0,037
5	Luis Rubiales case	15342	0	0,719	0,009	0,044
6	Reports and statistics	15342	0	0,517	0,025	0,052
7	Priests as perpetrators	15342	0	0,620	0,006	0,037
8	Harvey Weinstein case	15342	0	0,597	0,011	0,046
9	Cristiano Ronaldo case	15342	0	0,575	0,016	0,041
10	Sexual harassment	15342	0	0,455	0,012	0,038
11	Benjamin Mendy case	15342	0	0,468	0,003	0,021
12	Musicians as perpetrators	15342	0	0,669	0,023	0,060
13	Violence in figure skating	15342	0	0,354	0,003	0,017
14	Gérard Depardieu case	15342	0	0,597	0,012	0,047
15	Compensations for victims	15342	0	0,356	0,006	0,024
16	Daniel Alves case	15342	0	0,766	0,007	0,041
17	Fighting SV - organizations and funding	15342	0	0,484	0,007	0,026
18	Skopinsky maniac case	15342	0	0,648	0,004	0,027
19	Nobel prize for fighting SV	15342	0	0,685	0,004	0,028
20	Jeffrey Epstein case	15342	0	0,471	0,007	0,029
21	Victoria Marinova case	15342	0	0,581	0,004	0,024
22	Perpetrators detention	15342	0	0,851	0,020	0,065
23	The convention of womens' rights	15342	0	0,692	0,009	0,039
24	SV during the war in Ukraine	15342	0	0,423	0,005	0,026
25	Japanese program for women in sexual slavery	15342	0	0,627	0,004	0,029
26	Rape in the Ufa police department case	15342	0	0,904	0,021	0,089
27	Perpertrators crime sentences	15342	0	0,657	0,036	0,083
28	Details of criminal cases processes	15342	0	0,730	0,057	0,100
29	Perpertrators court sentencing	15342	0	0,620	0,028	0,064
30	Perpertrators court sentences murder cases	15342	0	0,524	0,017	0,047
31	Roman Polanski case	15342	0	0,240	0,003	0,015
32	Collectors case	15342	0	0,562	0,006	0,034
Valid N (listwise)		15342				

Table A4. Results of the normality test for the topic probability variables

ID	Topic name	Kolmogorov-Smirnov[a]		
		Statistic	df	Sig.
top	Rape and assault cases	0,175	15342	0,000
sup0	Victim perspective	0,255	15342	0,000
sup1	Descriptive narratives	0,186	15342	0,000
sup2	Problem regulation	0,205	15342	0,000
@0	Public statements	0,455	15342	0,000
@1	Royal family scandals	0,500	15342	0,000
@2	Violence in hockey	0,478	15342	0,000
@3	Khachaturian sisters case	0,473	15342	0,000
@4	Migrants as perpetrators	0,477	15342	0,000
@5	Luis Rubiales case	0,485	15342	0,000
@6	Reports and statistics	0,341	15342	0,000
@7	Priests as perpetrators	0,511	15342	0,000
@8	Harvey Weinstein case	0,490	15342	0,000
@9	Cristiano Ronaldo case	0,401	15342	0,000
@10	Sexual harassment	0,437	15342	0,000
@11	Benjamin Mendy case	0,501	15342	0,000
@12	Musicians as perpetrators	0,388	15342	0,000
@13	Violence in figure skating	0,507	15342	0,000
@14	Gérard Depardieu case	0,473	15342	0,000
@15	Compensations for victims	0,485	15342	0,000
@16	Daniel Alves case	0,512	15342	0,000
@17	Fighting SV - organizations and funding	0,484	15342	0,000
@18	Skopinsky maniac case	0,493	15342	0,000
@19	Nobel prize for fighting SV	0,496	15342	0,000
@20	Jeffrey Epstein case	0,485	15342	0,000
@21	Victoria Marinova case	0,502	15342	0,000
@22	Perpetrators detention	0,400	15342	0,000
@23	The convention of womens' rights	0,460	15342	0,000
@24	SV during the war in Ukraine	0,499	15342	0,000
@25	Japanese program for women in sexual slavery	0,505	15342	0,000
@26	Rape in the Ufa police department case	0,454	15342	0,000
@27	Perpertrators crime sentences	0,345	15342	0,000
@28	Details of criminal cases processes	0,283	15342	0,000
@29	Perpertrators court sentencing	0,334	15342	0,000
@30	Perpertrators court sentences murder cases	0,432	15342	0,000
@31	Roman Polanski case	0,501	15342	0,000
@32	Collectors case	0,509	15342	0,000

a. Lilliefors Significance Correction

Appendix 5. Topic structure of the discussion on sexual violence in Russian mass media

Table A5. The final topic structure modeled by the hierarchical Pachinko allocation model, with topic interpretations and topic class structure

Topic ID	High probability words	Unique words within the top 50 words by probability	Topic name	Topic class
top	мужчина женщина изнасиловать задержать полиция	напасть злоумышленник знакомый избить надругаться	Rape and assault cases	General
sup0	мочь человек говорить история хотеть	делать почему никто думать бояться	Victim perspective	General
sup1	стать несколько тело первый убийца	тело рука поздний спустя смерть	Descriptive narratives	General
sup2	человек право должный случай проблема	социальный количество внимание цель система	Problem regulation	Social problem
0	рассказать насилие звезда актриса признаться	признаться сцена модель артистка поделиться	Public statements	General
1	принц эндрю британский туристка аллен	принц эндрю туристка аллен елизавета	Royal family scandals	Spheres of violence
2	тренер насилие игрок клуб сборная	тренер канада хоккеист отстранить состав	Violence in hockey	Spheres of violence
3	сестра отец хачатурян убийство дело михаил	сестра отец хачатурян михаил экспертиза	Khachaturian sisters case	Cases
4	полиция германия мигрант женщина человек	германия мигрант немецкий кёльн акция	Migrants as perpetrators	Perpertrator types
5	футбол лига мир федерация испания	футбол лига игра рубиалеса чемпионат	Luis Rubiales case	Cases
6	сексуальный насилие жертва сообщать случай	подвергаться правительство передавать отчёт призвать	Reports and statistics	Social problem
7	церковь священник насилие католический ватикан	церковь священник католический римский ватикан	Priests as perpetrators	Perpertrator types
8	вайнштейн продюсер домогательство актриса сексуальный	вайнштейн продюсер голливудский metoo джоля	Harvey Weinstein case	Cases
9	девушка изнасилование адвокат слово заявить	роналда юрист ложный подруга подтвердить	Cristiano Ronaldo case	Cases
10	домогательство сексуальный женщина мужчина насилие	харассмент поведение приставать интимный рабочий	Sexual harassment	General
11	папа менди манчестер сити защитник	папа менди манчестер сити бенжамена	Benjamin Mendy case	Cases

12	обвинение обвинить сексуальный год иск	выдвинуть рэпер отрицать назад исполнитель	Musicians as perpetrators	Perpertrator types
13	фото виталий год молчанов катание	виталий молчанов катание фигуристка господин	Violence in figure skating	Spheres of violence
14	актёр фильм режиссёр год актриса	депардье французский роль съёмка жерар	Gérard Depardieu case	Cases
15	жертва млн год компенсация выплатить	компенсация рубль тыс сумма размер	Compensations for victims	General
16	алвес футболист материал барселона дань	алвес барселона дань приложение разрешить	Daniel Alves case	Cases
17	миллион доллар женщина присяжный фонд	академия основатель деньга заплатить влиятельный	Fighting SV - organizations and funding	Social problem
18	маньяк год жертва преступник скопинский	мохов виктор скопинский собчак екатерина	Skopinsky maniac case	Cases
19	год премия мир нобелевский буш	премия нобелевский буш лауреат верховный	Nobel prize for fighting SV	Social problem
20	эпштейн актив содержание произведение содержимое	эпштейн актив содержание произведение содержимое	Jeffrey Epstein case	Cases
21	отель журналистка швеция полиция убийство	швеция шведский лодка болгария турист	Victoria Marinova case	Cases
22	изнасилование москва задержать женщина сообщить	помощник подозрение столичный задержание подозреваться	Perpetrators detention	Criminal cases & process
23	насилие право домашний отношение закон	гендерный долг конвенция физический равенство	The convention of womens' rights	Spheres of violence
24	украина украинский россия оружие гражданин	украина украинский мирный территория население киев	SV during the war in Ukraine	Spheres of violence
25	япония корея женщина японский кэрролл	япония корея японский кэрролл южный	Japanese program for women in sexual slavery	Spheres of violence
26	изнасилование полицейский мвд отдел дело	мвд отдел уфа уфимский уволить	Rape in the Ufa police department case	Cases
27	год суд убийство свобода колония	лишение пожизненный строгий назначить отбывать	Perpertrators crime sentences	Criminal cases & process
28	дело действие уголовный преступление отношение	совершение стража совершенный пресечение версия	Details of criminal cases processes	Criminal cases & process
29	год суд изнасилование обвинение дело	судебный судья процесс оправдать прокурор	Perpertrators court sentencing	Criminal cases & process
30	суд ст рф ук убийство дело обвинять	причинение тяжкий рассмотрение угроза умышленный	Perpertrators court sentences murder cases	Criminal cases & process
31	год роман женский полански книга	роман полански бренд леди фбр	Roman Polanski case	Cases
32	коллектор новосибирский семья нападение рубль	коллектор новосибирский вернуть кредит должница	Collectors case	Cases

Appendix 6. Post-hoc tests for the ANOVA

Table A6. Results of the Games-Howell post-hoc test, the Rape and assault cases topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	0,016	0,009	0,669	-0,012	0,045
	2018	0,089	0,008	0,000	0,065	0,114
	2019	0,009	0,009	0,965	-0,017	0,035
	2020	0,070	0,009	0,000	0,043	0,097
	2021	0,117	0,008	0,000	0,093	0,142
	2022	0,068	0,008	0,000	0,042	0,093
	2023	0,052	0,009	0,000	0,026	0,078
2017	2016	-0,016	0,009	0,669	-0,045	0,012
	2018	0,073	0,009	0,000	0,046	0,100
	2019	-0,007	0,009	0,994	-0,036	0,021
	2020	0,054	0,010	0,000	0,025	0,083
	2021	0,101	0,009	0,000	0,074	0,128
	2022	0,051	0,009	0,000	0,023	0,079
	2023	0,036	0,009	0,004	0,007	0,064
2018	2016	-0,089	0,008	0,000	-0,114	-0,065
	2017	-0,073	0,009	0,000	-0,100	-0,046
	2019	-0,080	0,008	0,000	-0,104	-0,056
	2020	-0,019	0,008	0,298	-0,044	0,006
	2021	0,028	0,007	0,004	0,006	0,051
	2022	-0,022	0,008	0,100	-0,045	0,002
	2023	-0,037	0,008	0,000	-0,062	-0,013
2019	2016	-0,009	0,009	0,965	-0,035	0,017
	2017	0,007	0,009	0,994	-0,021	0,036
	2018	0,080	0,008	0,000	0,056	0,104
	2020	0,061	0,009	0,000	0,035	0,088
	2021	0,108	0,008	0,000	0,084	0,132
	2022	0,059	0,008	0,000	0,034	0,084
	2023	0,043	0,009	0,000	0,017	0,069

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2020	2016	-0,070	0,009	0,000	-0,097	-0,043
	2017	-0,054	0,010	0,000	-0,083	-0,025
	2018	0,019	0,008	0,298	-0,006	0,044
	2019	-0,061	0,009	0,000	-0,088	-0,035
	2021	0,047	0,008	0,000	0,022	0,072
	2022	-0,003	0,009	1,000	-0,029	0,024
	2023	-0,018	0,009	0,454	-0,045	0,009
2021	2016	-0,117	0,008	0,000	-0,142	-0,093
	2017	-0,101	0,009	0,000	-0,128	-0,074
	2018	-0,028	0,007	0,004	-0,051	-0,006
	2019	-0,108	0,008	0,000	-0,132	-0,084
	2020	-0,047	0,008	0,000	-0,072	-0,022
	2022	-0,050	0,008	0,000	-0,073	-0,026
	2023	-0,065	0,008	0,000	-0,090	-0,041
2022	2016	-0,068	0,008	0,000	-0,093	-0,042
	2017	-0,051	0,009	0,000	-0,079	-0,023
	2018	0,022	0,008	0,100	-0,002	0,045
	2019	-0,059	0,008	0,000	-0,084	-0,034
	2020	0,003	0,009	1,000	-0,024	0,029
	2021	0,050	0,008	0,000	0,026	0,073
	2023	-0,016	0,008	0,576	-0,041	0,010
2023	2016	-0,052	0,009	0,000	-0,078	-0,026
	2017	-0,036	0,009	0,004	-0,064	-0,007
	2018	0,037	0,008	0,000	0,013	0,062
	2019	-0,043	0,009	0,000	-0,069	-0,017
	2020	0,018	0,009	0,454	-0,009	0,045
	2021	0,065	0,008	0,000	0,041	0,090
	2022	0,016	0,008	0,576	-0,010	0,041

. The mean difference is significant at the 0.05 level.

Table A7. Results of the Games-Howell post-hoc test, the Victim perspective topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,006	0,003	0,478	-0,015	0,003
	2018	-0,009	0,003	0,030	-0,017	0,000
	2019	0,001	0,003	1,000	-0,007	0,009
	2020	-0,019	0,003	0,000	-0,029	-0,008
	2021	-0,015	0,003	0,000	-0,024	-0,006
	2022	0,000	0,003	1,000	-0,008	0,008
	2023	0,000	0,003	1,000	-0,008	0,008
2017	2016	0,006	0,003	0,478	-0,003	0,015
	2018	-0,003	0,003	0,975	-0,011	0,005
	2019	0,007	0,003	0,119	-0,001	0,015
	2020	-0,013	0,003	0,004	-0,023	-0,002
	2021	-0,009	0,003	0,038	-0,018	0,000
	2022	0,006	0,003	0,420	-0,003	0,014
	2023	0,006	0,003	0,408	-0,003	0,014
2018	2016	0,009	0,003	0,030	0,000	0,017
	2017	0,003	0,003	0,975	-0,005	0,011
	2019	0,010	0,002	0,001	0,003	0,017
	2020	-0,010	0,003	0,031	-0,019	0,000
	2021	-0,007	0,003	0,226	-0,015	0,002
	2022	0,009	0,003	0,016	0,001	0,016
	2023	0,008	0,002	0,013	0,001	0,016
2019	2016	-0,001	0,003	1,000	-0,009	0,007
	2017	-0,007	0,003	0,119	-0,015	0,001
	2018	-0,010	0,002	0,001	-0,017	-0,003
	2020	-0,020	0,003	0,000	-0,029	-0,010
	2021	-0,016	0,003	0,000	-0,025	-0,008
	2022	-0,001	0,002	0,999	-0,009	0,006
	2023	-0,001	0,002	0,999	-0,009	0,006
2020	2016	0,019	0,003	0,000	0,008	0,029
	2017	0,013	0,003	0,004	0,002	0,023
	2018	0,010	0,003	0,031	0,000	0,019
	2019	0,020	0,003	0,000	0,010	0,029
	2021	0,003	0,003	0,978	-0,007	0,013
	2022	0,018	0,003	0,000	0,009	0,028
	2023	0,018	0,003	0,000	0,009	0,028
2021	2016	0,015	0,003	0,000	0,006	0,024
	2017	0,009	0,003	0,038	0,000	0,018
	2018	0,007	0,003	0,226	-0,002	0,015
	2019	0,016	0,003	0,000	0,008	0,025
	2020	-0,003	0,003	0,978	-0,013	0,007
	2022	0,015	0,003	0,000	0,007	0,024
	2023	0,015	0,003	0,000	0,007	0,024

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,000	0,003	1,000	-0,008	0,008
	2017	-0,006	0,003	0,420	-0,014	0,003
	2018	-0,009	0,003	0,016	-0,016	-0,001
	2019	0,001	0,002	0,999	-0,006	0,009
	2020	-0,018	0,003	0,000	-0,028	-0,009
	2021	-0,015	0,003	0,000	-0,024	-0,007
	2023	0,000	0,003	1,000	-0,008	0,008
2023	2016	0,000	0,003	1,000	-0,008	0,008
	2017	-0,006	0,003	0,408	-0,014	0,003
	2018	-0,008	0,002	0,013	-0,016	-0,001
	2019	0,001	0,002	0,999	-0,006	0,009
	2020	-0,018	0,003	0,000	-0,028	-0,009
	2021	-0,015	0,003	0,000	-0,024	-0,007
	2022	0,000	0,003	1,000	-0,008	0,008

. The mean difference is significant at the 0.05 level.

Table A8. Results of the Games-Howell post-hoc test, the Descriptive narratives topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,006	0,002	0,339	-0,013	0,002
	2018	-0,003	0,002	0,877	-0,009	0,004
	2019	-0,001	0,002	1,000	-0,007	0,006
	2020	-0,009	0,002	0,009	-0,016	-0,001
	2021	-0,010	0,002	0,001	-0,016	-0,003
	2022	0,004	0,002	0,466	-0,002	0,011
	2023	-0,002	0,002	0,992	-0,008	0,005
2017	2016	0,006	0,002	0,339	-0,002	0,013
	2018	0,003	0,002	0,949	-0,004	0,010
	2019	0,005	0,002	0,466	-0,002	0,012
	2020	-0,003	0,003	0,924	-0,011	0,005
	2021	-0,004	0,002	0,695	-0,011	0,003
	2022	0,010	0,002	0,001	0,003	0,017
	2023	0,004	0,002	0,749	-0,003	0,011
2018	2016	0,003	0,002	0,877	-0,004	0,009
	2017	-0,003	0,002	0,949	-0,010	0,004
	2019	0,002	0,002	0,964	-0,004	0,008
	2020	-0,006	0,002	0,168	-0,013	0,001
	2021	-0,007	0,002	0,026	-0,013	0,000
	2022	0,007	0,002	0,005	0,001	0,013
	2023	0,001	0,002	0,999	-0,005	0,007
2019	2016	0,001	0,002	1,000	-0,006	0,007
	2017	-0,005	0,002	0,466	-0,012	0,002
	2018	-0,002	0,002	0,964	-0,008	0,004
	2020	-0,008	0,002	0,013	-0,015	-0,001
	2021	-0,009	0,002	0,001	-0,015	-0,002
	2022	0,005	0,002	0,141	-0,001	0,011
	2023	-0,001	0,002	1,000	-0,007	0,005
2020	2016	0,009	0,002	0,009	0,001	0,016
	2017	0,003	0,003	0,924	-0,005	0,011
	2018	0,006	0,002	0,168	-0,001	0,013
	2019	0,008	0,002	0,013	0,001	0,015
	2021	-0,001	0,002	1,000	-0,008	0,006
	2022	0,013	0,002	0,000	0,006	0,020
	2023	0,007	0,002	0,051	0,000	0,014
2021	2016	0,010	0,002	0,001	0,003	0,016
	2017	0,004	0,002	0,695	-0,003	0,011
	2018	0,007	0,002	0,026	0,000	0,013
	2019	0,009	0,002	0,001	0,002	0,015
	2020	0,001	0,002	1,000	-0,006	0,008
	2022	0,014	0,002	0,000	0,008	0,020
	2023	0,008	0,002	0,005	0,001	0,014

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	-0,004	0,002	0,466	-0,011	0,002
	2017	-0,010	0,002	0,001	-0,017	-0,003
	2018	-0,007	0,002	0,005	-0,013	-0,001
	2019	-0,005	0,002	0,141	-0,011	0,001
	2020	-0,013	0,002	0,000	-0,020	-0,006
	2021	-0,014	0,002	0,000	-0,020	-0,008
	2023	-0,006	0,002	0,041	-0,012	0,000
2023	2016	0,002	0,002	0,992	-0,005	0,008
	2017	-0,004	0,002	0,749	-0,011	0,003
	2018	-0,001	0,002	0,999	-0,007	0,005
	2019	0,001	0,002	1,000	-0,005	0,007
	2020	-0,007	0,002	0,051	-0,014	0,000
	2021	-0,008	0,002	0,005	-0,014	-0,001
	2022	0,006	0,002	0,041	0,000	0,012

Table A9. Results of the Games-Howell post-hoc test, the Public statements topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,005	0,001	0,000	-0,008	-0,001
	2018	-0,004	0,001	0,000	-0,007	-0,002
	2019	-0,003	0,001	0,046	-0,005	0,000
	2020	-0,009	0,001	0,000	-0,013	-0,005
	2021	-0,012	0,001	0,000	-0,016	-0,008
	2022	-0,002	0,001	0,134	-0,005	0,000
	2023	-0,004	0,001	0,000	-0,007	-0,001
2017	2016	0,005	0,001	0,000	0,001	0,008
	2018	0,000	0,001	1,000	-0,003	0,004
	2019	0,002	0,001	0,565	-0,001	0,005
	2020	-0,004	0,001	0,038	-0,009	0,000
	2021	-0,008	0,001	0,000	-0,012	-0,003
	2022	0,002	0,001	0,519	-0,001	0,006
	2023	0,001	0,001	0,998	-0,003	0,004
2018	2016	0,004	0,001	0,000	0,002	0,007
	2017	0,000	0,001	1,000	-0,004	0,003
	2019	0,002	0,001	0,584	-0,001	0,004
	2020	-0,005	0,001	0,005	-0,009	-0,001
	2021	-0,008	0,001	0,000	-0,012	-0,004
	2022	0,002	0,001	0,541	-0,001	0,005
	2023	0,000	0,001	1,000	-0,003	0,003
2019	2016	0,003	0,001	0,046	0,000	0,005
	2017	-0,002	0,001	0,565	-0,005	0,001
	2018	-0,002	0,001	0,584	-0,004	0,001
	2020	-0,006	0,001	0,000	-0,010	-0,003
	2021	-0,010	0,001	0,000	-0,013	-0,006
	2022	0,000	0,001	1,000	-0,003	0,003
	2023	-0,001	0,001	0,867	-0,004	0,002
2020	2016	0,009	0,001	0,000	0,005	0,013
	2017	0,004	0,001	0,038	0,000	0,009
	2018	0,005	0,001	0,005	0,001	0,009
	2019	0,006	0,001	0,000	0,003	0,010
	2021	-0,003	0,002	0,438	-0,008	0,001
	2022	0,007	0,001	0,000	0,003	0,011
	2023	0,005	0,001	0,002	0,001	0,009
2021	2016	0,012	0,001	0,000	0,008	0,016
	2017	0,008	0,001	0,000	0,003	0,012
	2018	0,008	0,001	0,000	0,004	0,012
	2019	0,010	0,001	0,000	0,006	0,013
	2020	0,003	0,002	0,438	-0,001	0,008
	2022	0,010	0,001	0,000	0,006	0,014
	2023	0,008	0,001	0,000	0,004	0,012

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,002	0,001	0,134	0,000	0,005
	2017	-0,002	0,001	0,519	-0,006	0,001
	2018	-0,002	0,001	0,541	-0,005	0,001
	2019	0,000	0,001	1,000	-0,003	0,003
	2020	-0,007	0,001	0,000	-0,011	-0,003
	2021	-0,010	0,001	0,000	-0,014	-0,006
	2023	-0,001	0,001	0,824	-0,004	0,002
2023	2016	0,004	0,001	0,000	0,001	0,007
	2017	-0,001	0,001	0,998	-0,004	0,003
	2018	0,000	0,001	1,000	-0,003	0,003
	2019	0,001	0,001	0,867	-0,002	0,004
	2020	-0,005	0,001	0,002	-0,009	-0,001
	2021	-0,008	0,001	0,000	-0,012	-0,004
	2022	0,001	0,001	0,824	-0,002	0,004

. The mean difference is significant at the 0.05 level.

Table A10. Results of the Games-Howell post-hoc test, the Sexual harassment topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,008	0,001	0,000	-0,012	-0,004
	2018	-0,012	0,001	0,000	-0,016	-0,009
	2019	-0,003	0,001	0,013	-0,006	0,000
	2020	-0,010	0,001	0,000	-0,014	-0,005
	2021	-0,004	0,001	0,005	-0,006	-0,001
	2022	-0,001	0,001	0,917	-0,004	0,002
	2023	-0,003	0,001	0,022	-0,006	0,000
2017	2016	0,008	0,001	0,000	0,004	0,012
	2018	-0,004	0,002	0,058	-0,009	0,000
	2019	0,004	0,001	0,011	0,001	0,008
	2020	-0,002	0,002	0,942	-0,007	0,003
	2021	0,004	0,001	0,011	0,001	0,008
	2022	0,007	0,001	0,000	0,003	0,010
	2023	0,005	0,001	0,002	0,001	0,009
2018	2016	0,012	0,001	0,000	0,009	0,016
	2017	0,004	0,002	0,058	0,000	0,009
	2019	0,009	0,001	0,000	0,005	0,013
	2020	0,003	0,002	0,810	-0,003	0,008
	2021	0,009	0,001	0,000	0,005	0,013
	2022	0,011	0,001	0,000	0,008	0,015
	2023	0,009	0,001	0,000	0,005	0,013
2019	2016	0,003	0,001	0,013	0,000	0,006
	2017	-0,004	0,001	0,011	-0,008	-0,001
	2018	-0,009	0,001	0,000	-0,013	-0,005
	2020	-0,006	0,002	0,001	-0,011	-0,002
	2021	0,000	0,001	1,000	-0,003	0,003
	2022	0,002	0,001	0,212	-0,001	0,005
	2023	0,000	0,001	1,000	-0,003	0,003
2020	2016	0,010	0,001	0,000	0,005	0,014
	2017	0,002	0,002	0,942	-0,003	0,007
	2018	-0,003	0,002	0,810	-0,008	0,003
	2019	0,006	0,002	0,001	0,002	0,011
	2021	0,006	0,001	0,001	0,002	0,011
	2022	0,009	0,001	0,000	0,004	0,013
	2023	0,007	0,001	0,000	0,002	0,011
2021	2016	0,004	0,001	0,005	0,001	0,006
	2017	-0,004	0,001	0,011	-0,008	-0,001
	2018	-0,009	0,001	0,000	-0,013	-0,005
	2019	0,000	0,001	1,000	-0,003	0,003
	2020	-0,006	0,001	0,001	-0,011	-0,002
	2022	0,002	0,001	0,116	0,000	0,005
	2023	0,000	0,001	1,000	-0,002	0,003

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,001	0,001	0,917	-0,002	0,004
	2017	-0,007	0,001	0,000	-0,010	-0,003
	2018	-0,011	0,001	0,000	-0,015	-0,008
	2019	-0,002	0,001	0,212	-0,005	0,001
	2020	-0,009	0,001	0,000	-0,013	-0,004
	2021	-0,002	0,001	0,116	-0,005	0,000
	2023	-0,002	0,001	0,321	-0,005	0,001
2023	2016	0,003	0,001	0,022	0,000	0,006
	2017	-0,005	0,001	0,002	-0,009	-0,001
	2018	-0,009	0,001	0,000	-0,013	-0,005
	2019	0,000	0,001	1,000	-0,003	0,003
	2020	-0,007	0,001	0,000	-0,011	-0,002
	2021	0,000	0,001	1,000	-0,003	0,002
	2022	0,002	0,001	0,321	-0,001	0,005

. The mean difference is significant at the 0.05 level.

Table A11. Results of the Games-Howell post-hoc test, the Compensations for victims topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	0,001	0,001	0,994	-0,001	0,003
	2018	0,001	0,001	0,991	-0,001	0,002
	2019	0,000	0,001	0,999	-0,002	0,002
	2020	-0,001	0,001	0,978	-0,003	0,001
	2021	-0,007	0,001	0,000	-0,011	-0,004
	2022	0,000	0,001	1,000	-0,002	0,002
	2023	0,000	0,001	0,999	-0,001	0,002
2017	2016	-0,001	0,001	0,994	-0,003	0,001
	2018	0,000	0,001	1,000	-0,002	0,002
	2019	0,000	0,001	1,000	-0,002	0,002
	2020	-0,001	0,001	0,657	-0,003	0,001
	2021	-0,008	0,001	0,000	-0,011	-0,005
	2022	-0,001	0,001	0,907	-0,003	0,001
	2023	0,000	0,001	1,000	-0,002	0,002
2018	2016	-0,001	0,001	0,991	-0,002	0,001
	2017	0,000	0,001	1,000	-0,002	0,002
	2019	0,000	0,001	1,000	-0,002	0,001
	2020	-0,001	0,001	0,593	-0,003	0,001
	2021	-0,008	0,001	0,000	-0,011	-0,005
	2022	-0,001	0,001	0,878	-0,003	0,001
	2023	0,000	0,001	1,000	-0,002	0,001
2019	2016	0,000	0,001	0,999	-0,002	0,002
	2017	0,000	0,001	1,000	-0,002	0,002
	2018	0,000	0,001	1,000	-0,001	0,002
	2020	-0,001	0,001	0,768	-0,003	0,001
	2021	-0,008	0,001	0,000	-0,011	-0,005
	2022	-0,001	0,001	0,965	-0,002	0,001
	2023	0,000	0,001	1,000	-0,002	0,002
2020	2016	0,001	0,001	0,978	-0,001	0,003
	2017	0,001	0,001	0,657	-0,001	0,003
	2018	0,001	0,001	0,593	-0,001	0,003
	2019	0,001	0,001	0,768	-0,001	0,003
	2021	-0,007	0,001	0,000	-0,010	-0,003
	2022	0,000	0,001	0,999	-0,002	0,003
	2023	0,001	0,001	0,722	-0,001	0,003
2021	2016	0,007	0,001	0,000	0,004	0,011
	2017	0,008	0,001	0,000	0,005	0,011
	2018	0,008	0,001	0,000	0,005	0,011
	2019	0,008	0,001	0,000	0,005	0,011
	2020	0,007	0,001	0,000	0,003	0,010
	2022	0,007	0,001	0,000	0,004	0,010
	2023	0,008	0,001	0,000	0,005	0,011

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,000	0,001	1,000	-0,002	0,002
	2017	0,001	0,001	0,907	-0,001	0,003
	2018	0,001	0,001	0,878	-0,001	0,003
	2019	0,001	0,001	0,965	-0,001	0,002
	2020	0,000	0,001	0,999	-0,003	0,002
	2021	-0,007	0,001	0,000	-0,010	-0,004
	2023	0,001	0,001	0,950	-0,001	0,002
2023	2016	0,000	0,001	0,999	-0,002	0,001
	2017	0,000	0,001	1,000	-0,002	0,002
	2018	0,000	0,001	1,000	-0,001	0,002
	2019	0,000	0,001	1,000	-0,002	0,002
	2020	-0,001	0,001	0,722	-0,003	0,001
	2021	-0,008	0,001	0,000	-0,011	-0,005
	2022	-0,001	0,001	0,950	-0,002	0,001

. The mean difference is significant at the 0.05 level.

Table A12. Results of the Games-Howell post-hoc test, the Problem regulation topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,009	0,004	0,268	-0,020	0,003
	2018	-0,017	0,003	0,000	-0,027	-0,008
	2019	0,003	0,003	0,988	-0,007	0,012
	2020	-0,007	0,003	0,520	-0,017	0,004
	2021	-0,014	0,003	0,000	-0,024	-0,004
	2022	-0,003	0,003	0,980	-0,013	0,007
	2023	0,003	0,003	0,981	-0,006	0,012
2017	2016	0,009	0,004	0,268	-0,003	0,020
	2018	-0,009	0,004	0,234	-0,020	0,002
	2019	0,011	0,003	0,024	0,001	0,022
	2020	0,002	0,004	0,999	-0,009	0,014
	2021	-0,005	0,004	0,821	-0,016	0,006
	2022	0,006	0,004	0,744	-0,005	0,016
	2023	0,012	0,003	0,018	0,001	0,022
2018	2016	0,017	0,003	0,000	0,008	0,027
	2017	0,009	0,004	0,234	-0,002	0,020
	2019	0,020	0,003	0,000	0,011	0,029
	2020	0,011	0,003	0,023	0,001	0,021
	2021	0,003	0,003	0,954	-0,006	0,013
	2022	0,014	0,003	0,000	0,005	0,024
	2023	0,020	0,003	0,000	0,011	0,029
2019	2016	-0,003	0,003	0,988	-0,012	0,007
	2017	-0,011	0,003	0,024	-0,022	-0,001
	2018	-0,020	0,003	0,000	-0,029	-0,011
	2020	-0,009	0,003	0,063	-0,019	0,000
	2021	-0,017	0,003	0,000	-0,026	-0,008
	2022	-0,006	0,003	0,495	-0,015	0,003
	2023	0,000	0,003	1,000	-0,008	0,009
2020	2016	0,007	0,003	0,520	-0,004	0,017
	2017	-0,002	0,004	0,999	-0,014	0,009
	2018	-0,011	0,003	0,023	-0,021	-0,001
	2019	0,009	0,003	0,063	0,000	0,019
	2021	-0,007	0,003	0,318	-0,017	0,003
	2022	0,004	0,003	0,952	-0,006	0,013
	2023	0,010	0,003	0,048	0,000	0,019
2021	2016	0,014	0,003	0,000	0,004	0,024
	2017	0,005	0,004	0,821	-0,006	0,016
	2018	-0,003	0,003	0,954	-0,013	0,006
	2019	0,017	0,003	0,000	0,008	0,026
	2020	0,007	0,003	0,318	-0,003	0,017
	2022	0,011	0,003	0,006	0,002	0,020
	2023	0,017	0,003	0,000	0,008	0,026
2022	2016	0,003	0,003	0,980	-0,007	0,013

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
	2017	-0,006	0,004	0,744	-0,016	0,005
	2018	-0,014	0,003	0,000	-0,024	-0,005
	2019	0,006	0,003	0,495	-0,003	0,015
	2020	-0,004	0,003	0,952	-0,013	0,006
	2021	-0,011	0,003	0,006	-0,020	-0,002
	2023	0,006	0,003	0,429	-0,003	0,015
2023	2016	-0,003	0,003	0,981	-0,012	0,006
	2017	-0,012	0,003	0,018	-0,022	-0,001
	2018	-0,020	0,003	0,000	-0,029	-0,011
	2019	0,000	0,003	1,000	-0,009	0,008
	2020	-0,010	0,003	0,048	-0,019	0,000
	2021	-0,017	0,003	0,000	-0,026	-0,008
	2022	-0,006	0,003	0,429	-0,015	0,003

. The mean difference is significant at the 0.05 level.

Table A13. Results of the Games-Howell post-hoc test, the Reports and statistics topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,004	0,002	0,593	-0,010	0,002
	2018	-0,001	0,002	0,999	-0,006	0,004
	2019	0,005	0,002	0,034	0,000	0,010
	2020	-0,001	0,002	1,000	-0,006	0,005
	2021	-0,001	0,002	1,000	-0,006	0,005
	2022	-0,006	0,002	0,030	-0,011	0,000
	2023	0,001	0,002	1,000	-0,004	0,006
2017	2016	0,004	0,002	0,593	-0,002	0,010
	2018	0,003	0,002	0,830	-0,003	0,008
	2019	0,009	0,002	0,000	0,003	0,014
	2020	0,003	0,002	0,811	-0,003	0,009
	2021	0,003	0,002	0,741	-0,003	0,009
	2022	-0,002	0,002	0,958	-0,008	0,004
	2023	0,004	0,002	0,232	-0,001	0,010
2018	2016	0,001	0,002	0,999	-0,004	0,006
	2017	-0,003	0,002	0,830	-0,008	0,003
	2019	0,006	0,001	0,001	0,002	0,010
	2020	0,000	0,002	1,000	-0,005	0,005
	2021	0,000	0,002	1,000	-0,004	0,005
	2022	-0,005	0,002	0,069	-0,010	0,000
	2023	0,002	0,001	0,944	-0,003	0,006
2019	2016	-0,005	0,002	0,034	-0,010	0,000
	2017	-0,009	0,002	0,000	-0,014	-0,003
	2018	-0,006	0,001	0,001	-0,010	-0,002
	2020	-0,006	0,002	0,008	-0,011	-0,001
	2021	-0,006	0,002	0,004	-0,010	-0,001
	2022	-0,011	0,002	0,000	-0,016	-0,006
	2023	-0,004	0,001	0,040	-0,009	0,000
2020	2016	0,001	0,002	1,000	-0,005	0,006
	2017	-0,003	0,002	0,811	-0,009	0,003
	2018	0,000	0,002	1,000	-0,005	0,005
	2019	0,006	0,002	0,008	0,001	0,011
	2021	0,000	0,002	1,000	-0,005	0,005
	2022	-0,005	0,002	0,090	-0,011	0,000
	2023	0,001	0,002	0,989	-0,004	0,007
2021	2016	0,001	0,002	1,000	-0,005	0,006
	2017	-0,003	0,002	0,741	-0,009	0,003
	2018	0,000	0,002	1,000	-0,005	0,004
	2019	0,006	0,002	0,004	0,001	0,010
	2020	0,000	0,002	1,000	-0,005	0,005
	2022	-0,005	0,002	0,049	-0,011	0,000
	2023	0,001	0,002	0,989	-0,003	0,006

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,006	0,002	0,030	0,000	0,011
	2017	0,002	0,002	0,958	-0,004	0,008
	2018	0,005	0,002	0,069	0,000	0,010
	2019	0,011	0,002	0,000	0,006	0,016
	2020	0,005	0,002	0,090	0,000	0,011
	2021	0,005	0,002	0,049	0,000	0,011
	2023	0,007	0,002	0,002	0,002	0,012
2023	2016	-0,001	0,002	1,000	-0,006	0,004
	2017	-0,004	0,002	0,232	-0,010	0,001
	2018	-0,002	0,001	0,944	-0,006	0,003
	2019	0,004	0,001	0,040	0,000	0,009
	2020	-0,001	0,002	0,989	-0,007	0,004
	2021	-0,001	0,002	0,989	-0,006	0,003
	2022	-0,007	0,002	0,002	-0,012	-0,002

. The mean difference is significant at the 0.05 level.

Table A14. Results of the Games-Howell post-hoc test, the Fighting SV – organizations and funding topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	-0,007	0,001	0,000	-0,010	-0,004
	2018	-0,010	0,001	0,000	-0,013	-0,007
	2019	-0,002	0,001	0,005	-0,004	0,000
	2020	-0,006	0,001	0,000	-0,008	-0,003
	2021	-0,004	0,001	0,000	-0,006	-0,002
	2022	-0,004	0,001	0,000	-0,006	-0,002
	2023	-0,004	0,001	0,000	-0,005	-0,002
2017	2016	0,007	0,001	0,000	0,004	0,010
	2018	-0,003	0,001	0,071	-0,006	0,000
	2019	0,005	0,001	0,000	0,002	0,007
	2020	0,001	0,001	0,952	-0,002	0,004
	2021	0,003	0,001	0,044	0,000	0,006
	2022	0,003	0,001	0,068	0,000	0,006
	2023	0,003	0,001	0,003	0,001	0,006
2018	2016	0,010	0,001	0,000	0,007	0,013
	2017	0,003	0,001	0,071	0,000	0,006
	2019	0,008	0,001	0,000	0,005	0,011
	2020	0,004	0,001	0,002	0,001	0,008
	2021	0,006	0,001	0,000	0,003	0,009
	2022	0,006	0,001	0,000	0,003	0,009
	2023	0,006	0,001	0,000	0,004	0,009
2019	2016	0,002	0,001	0,005	0,000	0,004
	2017	-0,005	0,001	0,000	-0,007	-0,002
	2018	-0,008	0,001	0,000	-0,011	-0,005
	2020	-0,004	0,001	0,002	-0,006	-0,001
	2021	-0,002	0,001	0,066	-0,004	0,000
	2022	-0,002	0,001	0,053	-0,004	0,000
	2023	-0,001	0,001	0,225	-0,003	0,000
2020	2016	0,006	0,001	0,000	0,003	0,008
	2017	-0,001	0,001	0,952	-0,004	0,002
	2018	-0,004	0,001	0,002	-0,008	-0,001
	2019	0,004	0,001	0,002	0,001	0,006
	2021	0,002	0,001	0,704	-0,001	0,004
	2022	0,001	0,001	0,782	-0,001	0,004
	2023	0,002	0,001	0,278	-0,001	0,005
2021	2016	0,004	0,001	0,000	0,002	0,006
	2017	-0,003	0,001	0,044	-0,006	0,000
	2018	-0,006	0,001	0,000	-0,009	-0,003
	2019	0,002	0,001	0,066	0,000	0,004
	2020	-0,002	0,001	0,704	-0,004	0,001
	2022	0,000	0,001	1,000	-0,002	0,002
	2023	0,001	0,001	0,996	-0,002	0,003

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,004	0,001	0,000	0,002	0,006
	2017	-0,003	0,001	0,068	-0,006	0,000
	2018	-0,006	0,001	0,000	-0,009	-0,003
	2019	0,002	0,001	0,053	0,000	0,004
	2020	-0,001	0,001	0,782	-0,004	0,001
	2021	0,000	0,001	1,000	-0,002	0,002
	2023	0,001	0,001	0,989	-0,002	0,003
2023	2016	0,004	0,001	0,000	0,002	0,005
	2017	-0,003	0,001	0,003	-0,006	-0,001
	2018	-0,006	0,001	0,000	-0,009	-0,004
	2019	0,001	0,001	0,225	0,000	0,003
	2020	-0,002	0,001	0,278	-0,005	0,001
	2021	-0,001	0,001	0,996	-0,003	0,002
	2022	-0,001	0,001	0,989	-0,003	0,002

. The mean difference is significant at the 0.05 level.

Table A15. Results of the Games-Howell post-hoc test, the Nobel prize for fighting SV topic

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2016	2017	0,000	0,001	0,999	-0,002	0,001
	2018	-0,007	0,001	0,000	-0,011	-0,004
	2019	0,000	0,001	1,000	-0,001	0,002
	2020	-0,002	0,001	0,065	-0,004	0,000
	2021	-0,002	0,001	0,170	-0,004	0,000
	2022	-0,002	0,001	0,043	-0,004	0,000
	2023	0,001	0,000	0,926	-0,001	0,002
2017	2016	0,000	0,001	0,999	-0,001	0,002
	2018	-0,007	0,001	0,000	-0,011	-0,003
	2019	0,001	0,001	0,950	-0,001	0,002
	2020	-0,002	0,001	0,235	-0,004	0,000
	2021	-0,001	0,001	0,477	-0,003	0,001
	2022	-0,002	0,001	0,190	-0,004	0,000
	2023	0,001	0,000	0,524	-0,001	0,002
2018	2016	0,007	0,001	0,000	0,004	0,011
	2017	0,007	0,001	0,000	0,003	0,011
	2019	0,008	0,001	0,000	0,004	0,011
	2020	0,005	0,001	0,002	0,001	0,009
	2021	0,006	0,001	0,000	0,002	0,010
	2022	0,005	0,001	0,001	0,001	0,009
	2023	0,008	0,001	0,000	0,004	0,012
2019	2016	0,000	0,001	1,000	-0,002	0,001
	2017	-0,001	0,001	0,950	-0,002	0,001
	2018	-0,008	0,001	0,000	-0,011	-0,004
	2020	-0,002	0,001	0,012	-0,004	0,000
	2021	-0,002	0,001	0,038	-0,004	0,000
	2022	-0,002	0,001	0,005	-0,004	0,000
	2023	0,000	0,000	0,991	-0,001	0,002
2020	2016	0,002	0,001	0,065	0,000	0,004
	2017	0,002	0,001	0,235	0,000	0,004
	2018	-0,005	0,001	0,002	-0,009	-0,001
	2019	0,002	0,001	0,012	0,000	0,004
	2021	0,000	0,001	1,000	-0,002	0,003
	2022	0,000	0,001	1,000	-0,002	0,002
	2023	0,003	0,001	0,001	0,001	0,005
2021	2016	0,002	0,001	0,170	0,000	0,004
	2017	0,001	0,001	0,477	-0,001	0,003
	2018	-0,006	0,001	0,000	-0,010	-0,002
	2019	0,002	0,001	0,038	0,000	0,004
	2020	0,000	0,001	1,000	-0,003	0,002
	2022	0,000	0,001	1,000	-0,002	0,002
	2023	0,002	0,001	0,002	0,001	0,004

(I) year		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
2022	2016	0,002	0,001	0,043	0,000	0,004
	2017	0,002	0,001	0,190	0,000	0,004
	2018	-0,005	0,001	0,001	-0,009	-0,001
	2019	0,002	0,001	0,005	0,000	0,004
	2020	0,000	0,001	1,000	-0,002	0,002
	2021	0,000	0,001	1,000	-0,002	0,002
	2023	0,003	0,001	0,000	0,001	0,004
2023	2016	-0,001	0,000	0,926	-0,002	0,001
	2017	-0,001	0,000	0,524	-0,002	0,001
	2018	-0,008	0,001	0,000	-0,012	-0,004
	2019	0,000	0,000	0,991	-0,002	0,001
	2020	-0,003	0,001	0,001	-0,005	-0,001
	2021	-0,002	0,001	0,002	-0,004	-0,001
	2022	-0,003	0,001	0,000	-0,004	-0,001

. The mean difference is significant at the 0.05 level.