Defining Kinds of Violence in Russian Short Stories of 1900–1930: A Case of Topic Modelling With LDA and PCA

Ekaterina Gryaznova[a] and Margarita Kirina[a]

[a] National Research University Higher School of Economics, 123 Griboyedova emb., St. Petersburg, 190068, Russia

Abstract

This paper discusses the problem of defining subthemes in literary texts about violence of different kinds from the Corpus of Russian short stories of the first third of the 20th century. It considers the results of topic modelling via Latent Dirichlet Allocation (LDA), which is used to reveal various kinds of violence, and principal component analysis (PCA), which is used to compare stories by the level of 'violent lexis saturation'. The experiment based on short stories that depict violence and death demonstrates that topic modelling did not allow the detection of internal topics but did group together stories with similar plots. The LDA algorithm seems to unveil some of the semantically related episodes of texts, though it is not always sufficient for providing complete interpretation of the resulting topics. The PCA method, on the other hand, successfully distinguishes between the following themes: *death*, *execution,* and *murder*. The research has proven that literary works are, indeed, rather difficult objects for automatic theme detection. In the case of fiction, the explicitness of themes appears to be a crucial factor in success of both LDA and PCA methods. The authors suggest that for more comprehensive analysis of fictional texts, several methods should be applied at the same time.

Keywords

Computational linguistics, machine learning, text mining, violence, Russian fiction, topic modelling, principal component analysis, latent Dirichlet allocation, literary corpus, literature studies

1. Introduction

Violence is considered to be an intrinsic part of human interactions that regard periods of time when various confrontations, be they social, political or historical, take place. Indeed, it is a foundation of a majority of social conflicts. As literature is, according to some interpretations [4; 5], a reflection of human experiences, it often chooses violence as its theme. Being an intercultural phenomenon, the issue of violence is reflected in a variety of texts, however, the definition of a 'violent' text is a quite challenging task. According to Reimer, texts determined by this theme "are often assumed by critics of media and literature to be those texts that depict acts of injurious physical force" [15, p. 102]. Though the description of violent acts through the lexis is a crucial part of the narrative about violence, there are some complications: a violent act is not necessarily placed in the text in an obvious way, and it is more likely to stay hidden in the rhetorical structures of the story [17, p. 2].

As one of the literary themes, violence reoccurs in the texts of the Corpus of Russian short stories of the first third of the 20th century [10], due to the specific period of time they were written. The beginning of the 20th century in Russia was marked by a number of violent historical events, such as the Russo-Japanese War, World War I, October and February Revolutions, and the subsequent Civil War. At the same time, cruel stories include not only examples of socially-induced violent acts and

IMS 2021 - International Conference "Internet and Modern Society", June 24-26, 2021, St. Petersburg, Russia
EMAIL: esgryaznova@edu.hse.ru (A. 1); mkirina2412@gmail.com (A. 2)
ORCID: 0000-0001-9844-2664 (A. 1); 0000-0002-7381-676X (A. 2)

CEUR Workshop Proceedings (CEUR-WS.org)

their consequences (death, murder, execution, rape), but also cases of, for instance, cruelty to animals or psychological pressure towards other characters. The common feature of all these forms of violence is that they are not necessarily placed in episodes of the stories in an explicit way.

This paper aims to explore the theme of violence in Russian short story of the early 20th century. To investigate violence diversity and intensity, we provide the analysis with topic modelling, using Latent Dirichlet Allocation (LDA) which is one of the most popular and quite effective topic modelling algorithms. Then, in order to scale the explicitness of violence narrative in the stories under consideration, we test them with principal component analysis (PCA) based on the list of violent lexis that has been compiled manually. This research continues the ongoing study of automatic thematic annotation of literary texts on the basis of the Corpus of Russian short stories of 1900–1930 described in [18; 19; 22]. For that reason, it also presents a comparison of human assessment of literary texts which exploit violence as a theme and the results of theme extraction obtained via an application of computational methods.

2. Data description and preprocessing

The experiment is performed on the part of the annotated subcorpus of the Corpus of Russian short stories of the first third of the 20th century, which includes 310 texts written by 300 different authors with the total number of almost 1 000 000 words [10; 11; 12]. The thematic annotation of the subcorpus was done manually by an expert and described in [20]. As a result, the initial mark-up of 89 themes was normalized and the list of 30 tags was obtained (for details see [18]). For the present analysis a ‘violence’ subcorpus was compiled. It contains 115 texts from 115 Russian writers with the following distribution of texts into historical periods suggested for the Corpus:

- **I period:** early 20th century (1900–1913) – 41 stories;
- **II period:** World War I, October and February revolutions, and the Civil War (1917–1922) – 40 stories;
- **III period:** early Soviet period (1923–1930) – 34 stories.

The selection of these specific stories was not random – all of them are united by the tags *violence* and *death*. The topic *death* was also chosen because it, firstly, often occurs in the same stories and, secondly, death naturally presents a resolution of violent conflict. Besides, the vast majority of texts about death refer to the cases of unnatural death, mainly violent or self-violent. In addition, stories with non-violent types of death pose another interesting challenge – will LDA distinguish it as a separate group?

Given the tag’s nature, similarly to categories, it can cover multiple themes. Thus, tag *violence* includes the following thematic elements – *rape*, *cruelty*, and *blood* (3), while tag *death* unites *death from gunshot wounds* (*during the war or on the barricade*); *death from natural causes* (including *epidemic* and *thoughts about death*), *execution* (*by shooting* as well, and *fear of death*), *sudden and accidental death*, *suicide*, and *murder* (*not at war*) (6). The total number of themes suggested by the expert equals 9.

It also has to be mentioned that some of the stories not only lie in both groups presented by tags *violence* and *death*, but also are described by several themes. For instance, *The Seven Who Were Hanged* (*Rasskaz o semi poveshennykh*) by L. Andreev is one of these stories and, moreover, allegedly, the most violent text in subcorpus, as it is labelled by 4 themes in total: *execution*, *death from natural causes*, *cruelty*, and *murder* (*not at war*). A story’s thematic density, thus, varies from one to another. Another peculiarity about the thematic mark-up of the literary texts in the given corpus is that the stories can at some point of the narrative develop non-violent or non-death related themes at all. So, one story, for example *Matter* (*Materiya*) by M. Krinitskij includes not only the tag violence, but also such tags as *relations*, *love*, *sins*, and *nature*. This tendency raises the problem if a number of themes that the story carries can cause a predicament for successful detection of the ones in question.

With regards to preprocessing, the texts were tokenized and lemmatized with automatic contextual disambiguation and POS-tagging by MyStem [16]. The total number of tokens is 426 778. Then the stop words and, additionally, the specific for fiction words that indicate the direct speech, such as *skazat'* (*to say*), *govorit'* (*to speak*), *otvechat'* (*to answer*), *sprashivat'* (*to ask*), *dumat'* (*to think*), and so on, as well as the most frequent names of the characters were removed. The tidy data size equals 228 745 tokens.

3. Topic modelling with LDA

3.1. Determination of the number of topics

Topic modelling is commonly used to detect clusters of semantically connected words within various corpora [13; 14]. As thus, a topic covers a cluster of texts which share similar content. Topic modelling is widely applied to large collections of texts, mainly non-fictional, where the quantity and quality of the topics are relatively easier to determine, due to the fact that there are no that many specific and implicit themes as we find in literary works [1; 3]. One of the most popular algorithms for topic modelling is Latent Dirichlet Allocation (LDA), which is an unsupervised generative probabilistic model [2]. Commonly speaking, it represents each document in data as a mixture of random topics.

For topic modelling the LDA implementation in R package 'topicmodels' was chosen [6]. After testing different numbers of topics, it was noted that the bigger the number gets the more detailed topics the model results. On closer consideration, for the model for 20 topics, it appeared to cover mainly individual texts rather than groups and, therefore, the topics were too detailed and difficult to interpret. This problem is similar to the one described in [21]. Since "the highest coherence value does not seem to necessarily correspond to the quality of topics", it was decided to limit the number of topics [ibid., p. 65]. To better the quality of the topics and to fulfill the suggestion to experiment with a number of topics proposed in [18], it was agreed to set the number of topics that corresponds with the one deduced from the expert annotation for the chosen group of texts – 9.

3.2. Evaluation of the model with expert annotation and stories per topic distribution

Dealing with short texts and, especially ones that often include other themes as well, even though they share the same thematic tags (namely *violence* and *death*), we are still facing some difficulties. As it can be seen from Table 1, for the words of the highest weight suggested for each topic, in some cases, for instance, topics 7 and 9, it is challenging to establish semantic connection between the terms, let alone assign the name, even if they are based on the list of expert themes. That is why in order to name the topics, we took corresponding thematic elements from the expert annotation and not only their distribution among the topic but also frequency measurements. Further evaluation of the topics' quality was conducted in accordance with the expert themes, as first suggested in [18]. We also decided to look into the stories that got clustered together based on the document distribution lists. The most frequent occurrences of the themes and the stories of the highest rank for each topic are presented in the table below.

Table 1
Distribution of themes and stories per topic

Topic terms	Thematic elements	Freq. (%)	Stories of the highest rank
Topic 1 "VIOLENCE TOWARDS WOMEN"			
den' (day), *god* (year) *dusha* (soul), ***noch'* (night)**, ***vremya* (time)**, *pis'mo* (letter), *hotet'* (to want), ***zhenshchina* (woman)**, *uhodit'* (leave)	death from natural causes	28,6	*Too Late* (Pozdno) by A. Verbitskaya, *The Platform 10* (Platforma 10) by L. Charskaya, *The Rooms in Kirochnaya street* (Nomera na Kirochnoj) by F. Bogrov
	suicide	28,6	
	cruelty	14,3	
	rape	7,1	

Topic terms	Thematic elements	Freq. (%)	Stories of the highest rank
Topic 2 “NON-WAR MURDER”			
***dom* (house)**, *den'* (day), *starik* (old man), ***ubivat'* (kill)**, *delo* (matter/case), *hotet'* (want), *hod* (move), *tolpa* (croud), *vdrug* (suddenly), ***ulitsa* (street)**	murder (not at war)	26,7	*The Chess* (Shakhmaty) by Ya. Braun, *The Burning Days* (Ognennye dni) by A. Gorelov, *Riot* (Bunt) by L. Lunts
	cruelty	20,0	
	suicide	13,3	
	death from gunshot wound	13,3	
Topic 3 “DEATH AT WAR”			
zemlya (ground), *den'* (day), *belyj* (white), *stoyat'* (stand), *chjorny* (black), ***soldat* (soldier)**, *muzhik* (man), *loshad'* (horse), ***doroga* (road)**, *storona* (side)	cruelty	23,5	*The Sharashka Bureau* (Sharashkina kontora) by B. Guber, *The Earth Shakes* (Zemnoj tryas) by A. Kargopolov, *The Outhouse* (Fligel') by A. Karavaeva
	execution	17,6	
	death from gunshot wounds	17,6	
	murder (not at war)	17,6	
Topic 4 “DOMESTIC VIOLENCE”			
buryj (fulvous), *pojti* (to go), *syn* (son), *hotet'* (to want), *stojat'* (stand), ***rebjonok* (child)**, *batjushka* (*priest*), *soldat* (soldier), *golos* (voice), ***krichat'* (scream)**	cruelty	35,7	*The Fulvous* (Buryj) by M. Chernokov, *A Nightmare* (Koshmar) by Gusev-Orenburgsky, *The Barricade* (Barricada) by G. Yablochkov
	suicide	21,4	
	death from gunshot wounds	21,4	
Topic 5 “UNEXPECTED DEATH AND ILLUSIONS”			
starik (old man), *vremja* (time), *stojat'* (stand), *zemlja* (ground), *dver'* (door), ***videt'* (see)**, ***golos* (golos)**, *kazatsya* (seem), *chjornyj* (black), *voda* (water)	cruelty	21,4	*Rioters* (Buntovshchiki) by P. Semynin, *The Trophy* (Nagrada) by N. Anov, *The Forgotten Colliery* (Zabytyj rudnik), ***Two Bloods* (Dva krovnika) by L. Pasynkov**
	sudden death	21,4	
	death from natural causes	14,3	
	suicide	14,3	
	murder (not at war)	14,3	
Topic 6 “SUDDEN DEATH”			
vdrug (suddenly), *den'* (day), ***smert'* (death)**, *hotet'* (want), *slovo* (word), *volk* (wolf), *kazatsya* (seem), *nachinat'* (start), *chas* (hour), *noch'* (night)	death from gunshot wounds	25,0	***The Seven Who Were Hanged* (Rasskaz o semi poveshennykh**) **by L. Andreev**, *The Silent Valley* (Gluchaja pad') by L. Ulin, *The Wolves* (Volky) by L. Zinovyeva-Annibal
	cruelty	18,8	
	murder (not at war)	18,8	
	death from natural causes	12,5	
	execution	12,5	
Topic 7 “NATURAL DEATH”			
den' (day), *drug* (friend), *kazatsya* (seem), *vdrug* (suddenly), *stojat'* (stand), *golos* (voice), *tolpa* (crowd), *vremja* (time), ***komnata* (room)**, *tjomnyj* (dark)	death from natural causes	23,1	*In the Circus* (V cyrke) by A. Kuprin, *In the Crowd* (V tolpe) by F. Sologub, *From Another World* (Iz drugogo mira) by V. Orlovsky
	sudden death	23,1	
	execution	15,4	
	death from gun wounds	15,4	
Topic 8 “LIFE IN PRISON”			
davat' (to give), *pojti* (to go), *delo* (case), *hotet'* (to want), ***lager'* (camp)**, *den'* (day), *prihodit'* (to come), ***russkij***	natural death	33,3	*Behind the Barbed Wire* (Za koluchej provolkoj) by K. Levin, *How Ivan spent time* (Kak Ivan provel vremja) by S. Podyachev,
	death from gun wounds	22,2	

Topic terms	Thematic elements	Freq. (%)	Stories of the highest rank
(russian), *zhit'* (to live), ***sidet'* (to be seated)**			*The Bad Hat* (Neputevyj) by E. Zamyatin
Topic 9 “CRUEL DEATH”			
lipa (Lipa), *pojti* (to go), *delo* (case), *ded* (grandfather), *hotet'* (to want), *bolshoj* (big), *vyhodit'* (to exit), *dver'* (door), *vdrug* (suddenly), *zemlja* (ground)	execution	23,1	*Savel Semenych* (Savel Semenych) by K. Fedin, *In the quiet corner* (V tikhom uglu) by E. Fedorov, *Communist* (Kommunistka) by A. Tyukhanov
	murder (not at war)	23,1	
	death from natural causes	15,4	
	cruelty	15,4	
	suicide	15,4	

The words that compose the clusters do not largely differ between the topics. Though, there are a few cases where the certain words strike the most. For example, these are the nouns that describe the places where the action takes place: *dom* (*house*), *komnata* (*room*), *ulitsa* (*street*), *lager'* (*camp*). Thus, topic 8, for example, seems to have gathered stories that describe prison and labour camps. Russian word *sidet'* (*to be seated*) has a second meaning of being in prison, and the word *lager'* (*labour camp*) adds to that theme. The word *russkij* (*Russian)* indicates the topic of international relations in prisons and camps that can be found in stories from this group.

Topic 1, on the other hand, probably exploits the themes of rape or cruel behavior towards women at some point of the narrative’s development. After considering the stories of the highest rank that contribute to this topic, it would be more accurate to say that all of them present a *woman* as a central figure. However, the stories below the 3rd rank are indeed dealing with another kind of violence, namely – *death from natural causes* and *suicide* (or suicide attempt). A pattern alike is found in topic 5 and 9. It is possible that these kinds of death are not largely presented in the lexis of the stories which makes it hard to detect them.

Moreover, we deliberately did not exclude verbs from the data, though this procedure is recommended for the improvement of the model [9]. It was suggested that violence is a theme that presupposes the usage of ‘active’ lexis. For that reason, it was expected that such verbs as, for instance, *to kill*, *to murder*, *to rape*, and etc., will result as terms of the highest probability within topics. However, the most helpful for interpretation words happened to be the nouns. More interestingly, the same tendency is discovered with regards to principal component analysis which is to be discussed in the next chapter.

4. Scaling violence with PCA

4.1. Detection of violent lexis

Principal component analysis (PCA) is an unsupervised machine learning method that reduces the dimensionality without losing much of statistical information [7]. Often textual data contains variables that either strongly correlate with each other, or there is not much variation within a variable. Such variables are often quite useless for research. The PCA reduces the size of data by creating new variables that represent it, while saving only important information. It also visualizes the important correlation between variables, thus this method works well for finding dependencies in data. Compared to LDA, it does not detect deep semantic connections. That being said, the PCA can scale the explicitness of the violence narrative in the given subcorpus. The PCA algorithm that was used for this research is from the R package ‘factoextra’ [8].

A list of violent words was compiled manually with consideration of cases which are specific for the period in question: *ubit'* (*to kill*), *ubivat'* (*to kill*), *bit'* (*to beat*), *izbit'* (*to beat up*), *izbivat'* (*to beat up*), *pribit'* (*to beat to death*), *dushit'* (*to choke*)*, pridushit'* (*to choke to death*), *udushit'* (*to choke to death*), *strelyat'* (*to shoot*), *zastrelit'* (*to kill by shooting*), *rasstrelyat'* (*to kill by shooting*), *pristrelit'* (*to kill by shooting*), *rasstrelivat'* (*to kill by shooting*), *zarezat'* (*to slaughter*), *topit'* (*to drown*), *utopit'* (*to kill by drowning*), *smert'* (*death*), *nasilije* (*violence*), *nasilovat'* (*to rape*), *iznasilovat'* (*to rape*), *pytat'* (*to torture*), *pytka* (*torture*), *prikonchit'* (*to kill*), *rasstrel* (*shooting*), *kazn'* (*execution*), *krov'* (*blood*), *udarit'* (*to hit*), *udaryat*' (*to hit*), *nasilstvennyj* (*violent*), *terror* (*terror*), *terrorizirovat*' (*to terrorize*), *prigovor* (*sentence*), *viselitsa* (*gallows*).

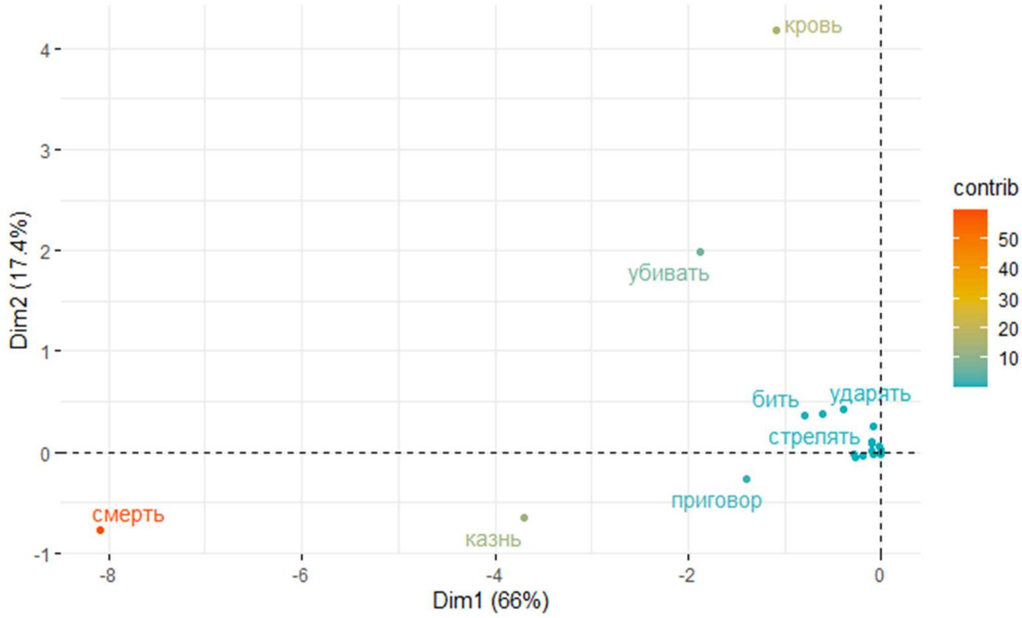

Figure 1: Distribution of violent lexis

According to Figure 1, the lemmas, which usage differs from all the other words, are *smert'* (*death*)*, krov'* (*blood*) and *kazn'* (*execution*). That means that these words appear more often in certain texts and that is what distinguishes one text from another. Other words like *bit'* (*to beat*)*, udaryat'* (*to hit*) and *strelyat'* (*to shoot*) do not excel. *Ubivat'* (*to kill*) does not contribute much to distinguishing a certain story, however it does excel. It could mean that this word is simply used more often in general, rather than it being specific to a certain story. It is expected that the stories also fall in the same pattern.

What is more, though the words that were chosen for the list are presented mainly by verbs, as it can be seen from the graph above, the most striking results, again, were obtained, except for *ubivat'* (*to kill*), by virtue of nouns. It seems that despite the proactiveness of the characters that show violent behavior, nouns contribute to the quality of both, the LDA model and the PCA, the most.

4.2. Degree of violence within a story

One of the disadvantages of the PCA is that it does not perform well on high-dimensional data; therefore, the graphs demonstrate only a few stories to make it more readable. As it can be seen from Figure 2, the most 'violent' stories are *The Seven Who Were Hanged* (Rasskaz o semi poveshennykh) by L. Andreev and *Two bloods* (Dva krovnika) by L. Pasynkov, which means they both contain more violent words than any other story. However, they differ from each other in terms of what kind of violence they describe: *The Seven Who Were Hanged* (Rasskaz o semi poveshennykh) has a strong connection with the word *smert'* (*death*), while *Two Bloods* (Dva krovnika) with the word *krov'* (*blood*) on the other side of the graph.

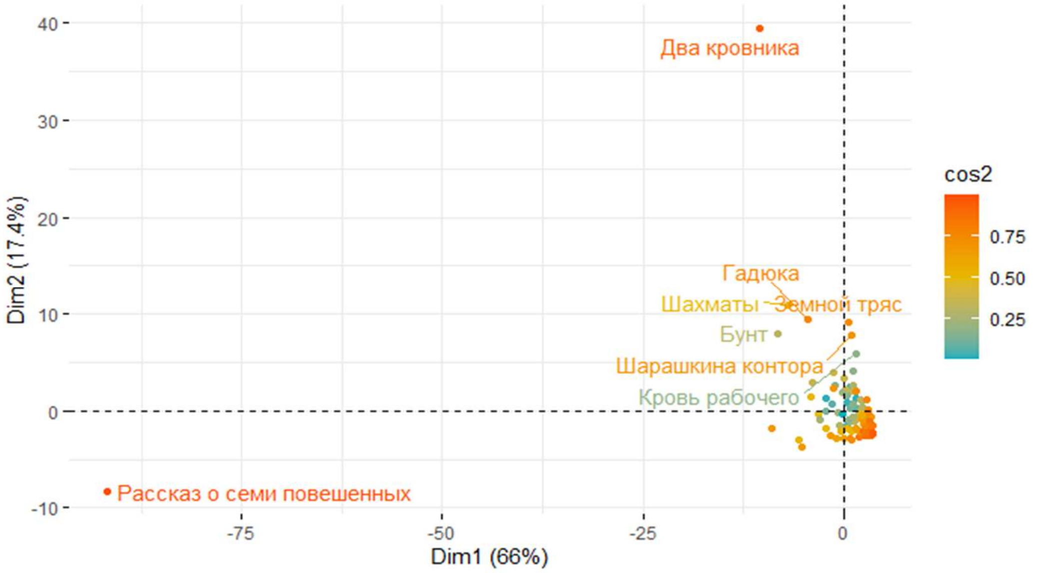

Figure 2: Degree of violent lexis in stories: a few examples

The stories *Viper* (*Gadyuka*) by A. Tolstoy and *The Sharashka Bureau* (*Sharashkina kontora*) by B. Guber also contain a lot of violent lexis, but they are not unique in the kind of lexis they contain. It is quite interesting that the story *Blood of a working man* (*Krov' rabochego*) by P. Arsky that already has a word from a dictionary in its title does not excel, which means that the word *krov'* (*blood*) does not contribute to distinguishing this story from others. Meanwhile *Two Bloods* (*Dva krovnika*) by L. Pasynkov is separated from the group, but *krov'* (*blood*) in this story also does not stand for the violent component only. As it follows from the story’s plot, it is about two brothers related by blood and their blood enemy. Though the violent episodes or discussions between characters, including, for instance, spilling blood, indeed take place, the word *blood* here may have several meanings. The examples like this one demonstrate the problem of explicitness in stories’ narratives.

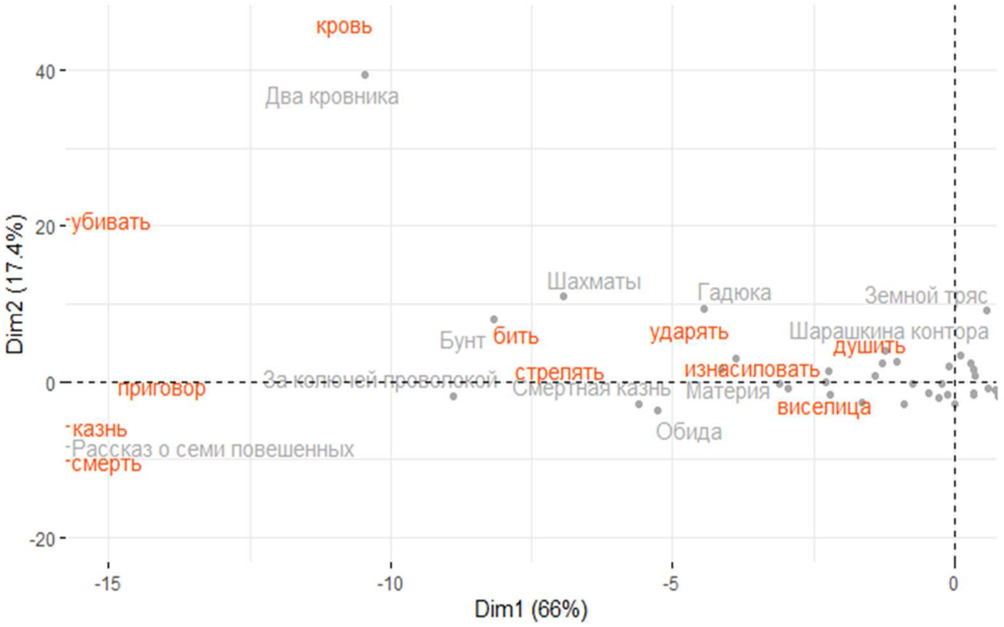

Figure 3: Correlation between lexis and stories

One thing that the PCA excelled in, as better seen in Figure 3, is in distinguishing the theme of execution (*prigovor* (*sentence*), *kazn'* (*execution*), *smert'* (*death*) and *viselitsa* (*gallows*)) from the other topics. One of them possibly being a distinctive topic of murder (*ubivat'* (*kill*), *krov'* (*blood*)), however *ubivat'* (*kill*) does not correlate with a specific story. Possibly the narrative of execution tends not to describe blood and gore, while the narrative of murder does.

To sum up the results of the PCA, this method is an efficient tool to measure the explicitness of topics in the text, in this case – violence, since it not only shows the most explicit stories, but also distinguishes death from execution and murder, though it did not separate death from violence. It did not detect any other violent acts or causes of death that are hidden in the rhetorical structures and are not that explicit.

5. Conclusion

This research has proven that, albeit the homogeneous nature of the subcorpus, the LDA and PCA algorithms are able to detect different violent acts, however, with a few restrictions in terms of their diversity. Thus, topic modelling was able to capture some common plot-related features in the stories, while the PCA allowed to distinguish two stories that excessively describe two kinds of violence.

On the whole, the analysis of the LDA model showed that the most probable words for each topic did not represent any violent acts. One possible explanation regards the data itself, namely the fact that some stories fall into different categories at the same time which may complicate the detection of themes of our interest only. Another reason for unsatisfying results is that the stories, being short, do not comprehensively cover each of the subthemes enough, so they are not vastly expressed in the lexis of the texts. On the other hand, it occurs that the LDA was able to put into one cluster the stories with the similar plot details or characters (in terms of gender or social status). For instance, topic 1 unites those stories in which woman is the main character while the depiction of the act of rape or any sexual act at all is not necessarily present. Perhaps, regarding literature, topic modelling appears to identify common structures that occur in various texts, however, they do not always constitute their themes.

Since the PCA works with variables, it performed better – the difference between death by execution and murder was detected which can be juxtaposed with the stories. Thus, the most explicitly violent stories – *Two bloods* (*Dva krovnika*) by L. Pasynkov and *The Seven Who Were Hanged* (*Rasskaz o semi poveshennyh*) by L. Andreev – tell about murder and execution respectively. What is more, a comparison of the PCA results and per-document-per-topic probabilities of the LDA shed a light on some interesting tendencies. *Two bloods* (*Dva krovnika*) by L. Pasynkov and *The Seven Who Were Hanged* (*Rasskaz o semi poveshennyh*) by L. Andreev, which were distinguished by the PCA as the most violent, are also the stories of the highest rank in topics 5 and 6 respectively of the LDA model. For any other topic they do not contribute the same, lying at the bottom of the lists. It appears that these two stories indeed differ in terms of violence representation from others.

To conclude, we suggest that applying topic modelling to the literary texts unveils some difficulties prompted by the fact that the theme in the fictional text, as a rule, is not obviously expressed in the story's lexis. In our case, comparing the expert annotation and the automatic one, the automatic one did not detect as many themes as the expert did. For that reason, when tackling fictional texts that do not differ in genre, several methods need to be applied. As this study shows, the PCA can contribute to dealing with lexis-specific themes extraction. Additionally, the results of both LDA and PCA could be properly interpreted only with the knowledge of the contents of the stories and their thematic assessment by an expert.

For future research, experimenting with various topic modelling algorithms (NMF, for instance), on the one hand, and applying supervised machine learning methods for analysis of literary works, on the other, might help to obtain better results in terms of comprehensive interpretation. Mastering the automatic theme extraction may be a step towards human-alike textual analysis, allowing studying literature with the means of computation methods in more conclusive manner.

6. Acknowledgements

The publication is prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2021 (grant # 21-04-053 'Artificial Intelligence Methods in Literature and Language Studies').

7. References

[1] R. Albalawi, T. H. Yeap, M. Benyoucef, Using topic modeling methods for short-text data: A comparative analysis, in: Frontiers in Artificial Intelligence, 3, 2020.

[2] D. M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, in: J. Mach. Learn. Res. 3(4–5), 2003, pp. 993–1022.

[3] B. Blummer, J. M. Kenton, Academic Libraries' Outreach Efforts: Identifying Themes in the Literature, in: Public Services Quarterly, Volume 15, Issue 3, 2019, pp. 179–204.

[4] J. Carroll, The extremes of conflict in literature: Violence, homicide, and war, in: The Oxford handbook of evolutionary perspectives on violence, homicide, and war, 2012.

[5] J. Carroll, Violence in literature: an evolutionary perspective, in: The evolution of violence, 2014, pp. 33–52.

[6] B. Grun, K. Hornik, Topicmodels: An {R} Package for Fitting Topic Models, in: Journal of Statistical Software, vol. 40 (13), 2011, pp. 1–30.

[7] I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, in: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374 (2065), 2016.

[8] A. Kassambara, F. Mundt, Factoextra: Extract and Visualize the Results of Multivariate Data Analyses, 2020. URL: https://CRAN.R-project.org/package=factoextra.

[9] F. Martin, M. Johnson, More efficient topic model-ling through a noun only approach, in: Proceedings of the Australasian Language Technology Association Workshop, 2015, pp. 111–115.

[10] G. Y. Martynenko, T. Y. Sherstinova, A. G. Melnik, T. I. Popova, Methodological issues related with the compilation of digital anthology of Russian short stories (the first third of the 20th century), in: Proceedings of the XXI International United Conference 'The Internet and Modern Society', IMS–2018, Computational linguistics and computational ontologies, ITMO University, St. Petersburg, Issue 2, 2018a, pp. 99–104.

[11] G. Y. Martynenko, T. Y. Sherstinova, T. I. Popova, A. G. Melnik, E.V. Zamirajlova, O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka [About Principles of the Creation of the Corpus of Russsian Short Stories of the First Third of the 20th Century], in: Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics 'TEL2018', Kazan Federal University. Kazan, 2018b, pp.180–197.

[12] G. Martynenko, T. Sherstinova, Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century, in: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552, 2020, pp. 105–120. URL: http://ceur-ws.org/Vol-2552/.

[13] O. A. Mitrofanova, Modelirovanije tematiki spe-cial'nyh tekstov na osnove algoritma LDA [Topic modeling of special texts based on LDA algorithm], in: XLII Mezhdunarodnaya filologicheskaya konferencija [XLII International philological conference], 2014.

[14] S. Nikolenko, S. Koltcov, O. Koltsova, Topic modelling for qualitative studies, in: J. Inf. Sci. 43(1), 2017, pp. 88–102.

[15] M. Reimer, Introduction: Violence and Violent Children's Texts, in: Children's Literature Association Quarterly, 22(3), 1997, pp. 102–104.

[16] I. Segalovich, V. Titov, MyStem. Yandex [Computer Software], 2011. URL: https://yandex.ru/dev/MyStem/.

[17] S. Sielke, Reading rape: The rhetoric of sexual violence in American literature and culture, 1790-1990. Princeton, 2009.

[18] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, M. Kirina, Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction, in: Advances in Computational Intelligence, MICAI 2020, Lecture Notes in Computer Science, Vol. 12469, 2020, pp. 134–151.

[19] T. Sherstinova, T. Skrebtsova, Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900-1930, in: CompLing (in print).

[20] T. G. Skrebtsova, Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories, in: CompLing, CEUR Workshop Proceedings, Vol. 2813, 2021, pp. 265-276.

[21] I. Uglanova, E. Gius, The Order of Things. A Study on Topic Modelling of Literary Texts, in: Proc. of the CHR 2020: Workshop on Computational Humanities Research, CEUR Workshop Proceedings, 2020. URL: http://ceur-ws.org/Vol-2723/long7.pdf.

[22] E. Zamiraylova, O. Mitrofanova, Dynamic topic modeling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization, in: Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019), Vol. 2552, 2019, pp. 321–339.