www.volsu.ru

DOI: https://doi.org/10.15688/jvolsu2.2024.5.5

UDC 811.161.1:159.942
LBC 81.411.2-51

Submitted: 09.05.2024
Accepted: 20.08.2024

TEXTS OF DIFFERENT EMOTIONAL CLASSES AND THEIR TOPIC MODELING [1]

Anastasia V. Kolmogorova
HSE University, Saint Petersburg, Russia

Qiuhua Sun
Heilongjiang University, Harbin, China

Abstract. The article is devoted to studying verbalization specifics of various emotional states in the texts in the Russian language with the purpose to confirm or refute the hypothesis that texts of different emotional classes reflect the denotative situation not identically, which is reflected in thematic specifics and lexical content. The research material consisted of eight corpus texts in the Russian language, which were extracted from the public pages of the social network VKontakte. The texts were selected according to emotional hashtags that corresponded to eight basic emotions, according to H. Lövheim's model: anger, surprise, shame, enjoyment, disgust, distress, excitement, fear. The correspondence of emotion and hashtag was established in a preliminary psycholinguistic experiment. While analyzing the text collection, we used the method of computer thematic modeling to identify statistically non-random groups of words (topics). We applied the BERTopic neural network model to the collected data. As a result of the analysis, it was found that texts of 8 emotional classes contain an uneven number of topics, despite the fact that their number does not correlate directly with the amount of data: with a relatively small amount of data, there may be many topics, but in a voluminous corpus – few. The sets of words (tokens) that make up each non-random group (topic) differ in each subcorpora, reflecting the specifics of the denotative situation, which is formed under the influence of the emotional state of the speaker. The idea of diverse thematic "granularity" of texts of different emotional classes is theoretically justified.

Key words: emotions, denotative situation, topic modeling, social network texts, Russian language.

Citation. Kolmogorova A.V., Sun Qiuhua. Texts of Different Emotional Classes and Their Topic Modeling. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2024, vol. 23, no. 5, pp. 60-71. DOI: https://doi.org/10.15688/jvolsu2.2024.5.5

УДК 811.161.1:159.942
ББК 81.411.2-51

Дата поступления статьи: 09.05.2024
Дата принятия статьи: 20.08.2024

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ТЕКСТОВ РАЗНЫХ ЭМОЦИОНАЛЬНЫХ КЛАССОВ [1]

Анастасия Владимировна Колмогорова
Национальный исследовательский университет «Высшая школа экономики», г. Санкт-Петербург, Россия

Цюхуа Сунь
Хэйлундзянский университет, г. Харбин, Китай

Аннотация. Статья посвящена проблеме вербализации различных эмоциональных состояний в текстах на русском языке. Цель работы – подтвердить или опровергнуть гипотезу о том, что тексты разных эмоциональных классов неодинаково отражают денотативную ситуацию, имеют тематическую специфику и не идентичное лексическое наполнение. Материал исследования составили восемь подкорпусов текстов на русском языке, которые были извлечены из пабликов социальной сети ВКонтакте. Тексты отобраны

по эмоциональным хэштегам, которые соответствуют восьми базовым эмоциям, согласно модели Г. Лёвхейма: злость, удивление, стыд, радость, отвращение, печаль, воодушевление, унижение. Соответствие эмоции и хэштега было установлено в предварительном психолингвистическом эксперименте. Для анализа текстовой коллекции использовалась техника выделения статистически неслучайных групп слов (тем) при помощи компьютерного алгоритма – метод компьютерного тематического моделирования. К собранным данным применена нейросетевая модель BERTopic. В результате анализа было выявлено, что тексты разных эмоциональных классов содержат неодинаковое количество тем, при том, что их число не коррелирует непосредственно с объемом данных: при сравнительно небольшом объеме данных может быть много тем, а в объемном корпусе – мало. Наборы слов (токенов), составивших каждую неслучайную группу (тему), отличаются по подкорпусам, отражая специфику денотативной ситуации, формирующуюся под влиянием эмоционального состояния говорящего. Теоретическое обоснование получает идея о специфической тематической «гранулярности», характерной для текстов разных эмоциональных классов.

Ключевые слова: эмоции, денотативная ситуация, тематическое моделирование, тексты в социальных сетях, русский язык.

Цитирование. Колмогорова А. В., Сунь Цюхуа. Тематическое моделирование текстов разных эмоциональных классов // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2024. – Т. 23, № 5. – С. 60–71. – (На англ. яз.). – DOI: https://doi.org/10.15688/jvolsu2.2024.5.5

Introduction

This publication is devoted to the problem of emotional analysis of text data. We formulate the research question as follows: is there a stable correlation between the topics that receive formal expression in words with high statistical significance for the text, and the emotion that the author of the text seeks to express? The search for an answer to this question lies in three overlapping subject areas: psycholinguistics, textology, and automatic text processing. In the latter, there is great interest in the use of topic modeling models. The topic modeling of a collection of textual documents determines which topics each document belongs to and which words form each topic. To do this, each topic is described by a discrete probability distribution of words, and each document is described by a discrete probability distribution of topics. Essentially, this model performs a soft clustering of documents. In this paper we consider the results of applying the BERTopic model to topic modeling of texts retrieved from the VKontakte "Overheard" public group and published under different emotional hashtags, which were then correlated with eight classes of emotion, according to H. Lövheim's emotion model [Lövheim, 2012].

Related papers

It should be said that emotion analysis of texts is an increasingly popular research branch of the affective computing paradigm [Picard, 1997], which seeks to solve such problems as detecting different emotions expressed by humans in texts and classifying texts according to the leading emotion criterion [Blei, Ng, Jordan, 2003; Hakak et al., 2017; Li et al., 2007].

To solve both of the above tasks, expert linguistic attempts have been repeatedly made to identify some verbal and paraverbal markers of texts of a particular emotional class in order to use them as parameters fed to machine learning models as input [Kolmogorova, Kalinin, Malikova, 2019].

On the other hand, there is a well-established tradition in textual studies of identifying the emotional and semantic dominant of a text which is understood as a certain attitude, the center of interest, and therefore a certain position in all kinds of human verbal and averbal life [Shakhovsky, 2010, p. 41].

Based on the consideration of this dominant, attempts were made to psycholinguistically typologize texts into light, sad, funny, active, simple (cruel), beautiful, tired, complex and mixed [Belyanin, 2000]. It is established that for each of these types it is possible to identify a specific range of topics, a list of predicates. For example, on the material of dialect texts Y.V. Kositsina revealed that in the sad texts the external location prevails over the internal. Moreover, such texts are marked by the topics of the vicissitudes of fate and family, the sense of the inevitability of death, defenselessness of man before the laws of existence, the opposition of thematic spaces of past and present, the presence of words containing semantic components "solitude", "blindness", "gravity" [Kositsina, 2013].

Similarly, but using a neural network, we attempted to determine the groups of words that are most likely characteristic to the texts published by users of the social network VKontakte under the emotional hashtags corresponding to a particular emotion. By default, we suppose that such statistically significant words are topic cues that anchor main text themes.

Materials and methods

The research data consisted of eight subcorpora of texts from VKontakte. Their volume and the hashtags used to extract the text data are presented in Table. Each subcorpus is named according to the emotion with which it is associated.

We emphasize that the adequacy of the correlation between hashtags and emotional classes was tested experimentally in a group of 35 students, who in two experimental series were asked to correlate texts and emotional classes (160 texts were presented). We then calculated the percentage of the texts with certain hashtags that fell into a sample of one class or another. Thus, if more than 80% of the texts presented in the sample with a hashtag were associated by the informants with a particular emotion, the hashtag was considered reliable; if less than 80%, the hashtag was discarded.

As the main method, we used topic modeling by the BERTopic model, which, as research shows, is significantly superior in a number of parameters to the well-known model based on Latent Dirichlet Allocation [Sia, Dalmia, Mielke, 2020].

BERTopic generates "topics" in three stages: first, each document on the basis of the already trained model receives a vector representation, then, to conduct clustering, the dimensionality of vectors is reduced, finally, in the last stage, when the documents are already clustered, based on the standard tf-idf measure, the topics themselves are extracted from the clusters – lexemes or word forms that have the highest weight for these texts [Grootendorst, 2022]. In other words, each topic suggested by the model represents a group of words which are statistically relevant for the text or text collection. Word forms entering the same group are called its "terms".

We used the Uniform Manifold Approximation and Projection model to reduce the dimensionality of the vectors, and the Hierarchical Density-Based Spatial Clustering of Applications with Noise method to cluster word forms.

The results of the topic modeling (topics, terms and their weights, as well as similarity matrices and distances between topics) of the texts of each of the eight emotional classes were compared with each other, and the results of the comparison were interpreted.

Results and discussion

Low Degree of Topical Granularity: Anger, Distress and Excitement

Despite the rather large size, compared to the other subcorpora, the subcorpus of "angry" texts allowed the model to identify only three topics (see Fig. 1).

The first topic contains a typical gist for the posts of this group – *pissing off people who...* (бесят люди, которые...); the second reflects the negative attitude of people living in Russia towards those who live or want to live, or pretend to live abroad; finally, the third topic represents the word forms that reflect the intensity of feeling the

Subcorpora size and hashtags

Subcorpus	Size in tokens	Hashtag
Anger	131 564	#Подслушано_БЕСИТ
Disgust	45 868	#Подслушано_фуу
Distress	56 470	#Подслушано_одиночество
Enjoyment	85 117	#Подслушано_счастье
Shame	70 232	#Подслушано_стыдно
Excitement	184 074	#Подслушано_успех
Surprise	288 272	#Подслушано_наблюдения #Подслушано_иллюстрация #Подслушано_странно
Fear	230 730	#Подслушано_страшно

emotion of anger: obscenisms, interjections, invectives.

From the subcorpus of "sad" texts it was also possible to extract only three topics (Fig. 2): the first is related to the idea of the finitude of time, the second – to the idea of loneliness, and the third – of love.

Even two topics were found by the model in the subcorpus of texts, showing the excitement (Fig. 3) despite the fact that this, as well as the subcorpus of "angry" texts, is one of the largest subcorpora in the sample.

The first topic manifests the idea of time, while the second is related to effective weight loss. Interestingly, this subcorpus represented a variety of situations that evoked the emotion of elation-not just those related to long-awaited weight loss (examples 1–2), but the model apparently had difficulty to cluster them:

(1) I teach a foreign language. I was invited to lead a group of eight people in a foreign language studio; I came – and there are all men and guys. All good-looking, intelligent, educated, and all unmarried! I am ugly, fat, short, unremarkable and unattractive absolutely; I do not know what happened, but all the students come to class with a perfectly prepared homework, dressed with a needle, smelling of perfume, stretching their hands, want to answer, and the fourth student asked me out. At home there are bouquets from two other students. Either it's the magic of the German language, or I don't even know...

Я преподаю иностранный язык. Меня пригласили руководить группой из восьми человек в студии иностранных языков; я пришла – а там сплошь мужчины и парни. Все красивые, умные, образованные и все неженатые! Я некрасивая, толстая, низкорослая, ничем не примечательная и абсолютно непривлекательная; я не знаю, что случилось, но все ученики приходят на урок с идеально подготовленным домашним заданием, одетые с иголочки, пахнущие духами, протягивают руки, хотят ответить, и четвертый ученик пригласил меня на свидание. Дома меня ждут букеты от двух других студентов. То ли это волшебство немецкого языка, то ли я даже не знаю...

(2) My wife gave me a ticket to the World Cup finals. She bought it long beforehand, as a surprise. She gave me half my salary, poor thing. Now my wife,

Fig. 1. Topics and terms of texts manifesting the emotion of anger

Fig. 2. Topics and terms of texts manifesting the emotion of distress

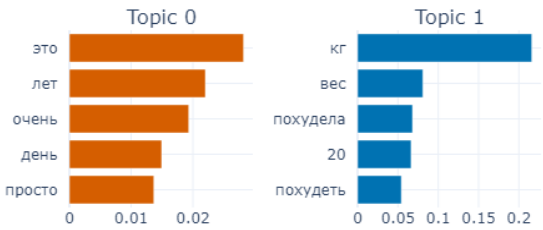

Fig. 3. Topics and terms of texts manifesting the emotion of excitement

father-in-law, and mother-in-law and I are flying to Hawaii. A fan from France bought my ticket for 100 times the price! I love my wife. And I never liked soccer.

Моя жена подарила мне билет на финал Чемпионата мира. Она купила его задолго до этого, в качестве сюрприза. Бедняжка, она отдала мне половину моей зарплаты. Сейчас мы с женой, тестем и тещей летим на Гавайи. Болельщик из Франции купил мой билет в 100 раз дороже! Я люблю свою жену. И мне никогда не нравился футбол.

Middle Degree of Topical Granularity: Disgust, Fear, Shame

The opposite of the previously analyzed case situation is observed in modeling the subcorpus of "disgusting" texts: although they are the smallest in volume, they allowed the model to extract 9 topics.

Among them predictably well stand out topics related to the physiological and bodily spheres (Fig. 4): toileting, washing the body and hair, male-female relations, women of the family, olfactory sensations, and body parts such as the mouth.

The similarity matrix, by the way, shows that this last topic (0) is very similar to all the other topics that deal with the concepts of relationship of the opposite sexes (Topics 4, 6, 7) (Fig. 5).

In the group of texts manifesting fear (Fig. 6), the model identified six topics: these are topics connected with the subject "persons for whom the fear is felt" (Topic 0 – mother), or the place where the fear is usually felt (Topic 4 – the bathroom and the toilet), or, actually, why the fear is felt, its cause (Topic 1, 2, 5) – a car accident, cancer, mental disorder.

At first glance, the weakly interpreted Topic 3 turns out to be more understandable with the help of the similarity matrix: it is very similar to Topic 0 (Fig. 7).

Thus, these two topics seem to develop the same proposition – the fear that the dearest person, the mother, may leave (*mother*, *her*, *very*), the regret after her demise (*why*, *you know*, *come back*).

The final emotion for this block is the emotion of shame. In this subcorpus, which is small enough, the model has revealed 9 topics (Fig. 8).

In their assemblage we can also identify a number of components of the situation of experiencing the shame: time (*long ago*, *not now* – Topic 8; in school childhood – Topic 2), participants of the shame situation (*husband*, *boyfriend* – Topic 4), emotional reactions (*became very ashamed* – Topic 3), internal experiences (*could not tell anyone* – Topic 7), objects with which the shame was associated – for example, money (Topic 1).

Fig. 4. Topics and terms of texts manifesting the emotion of disgust

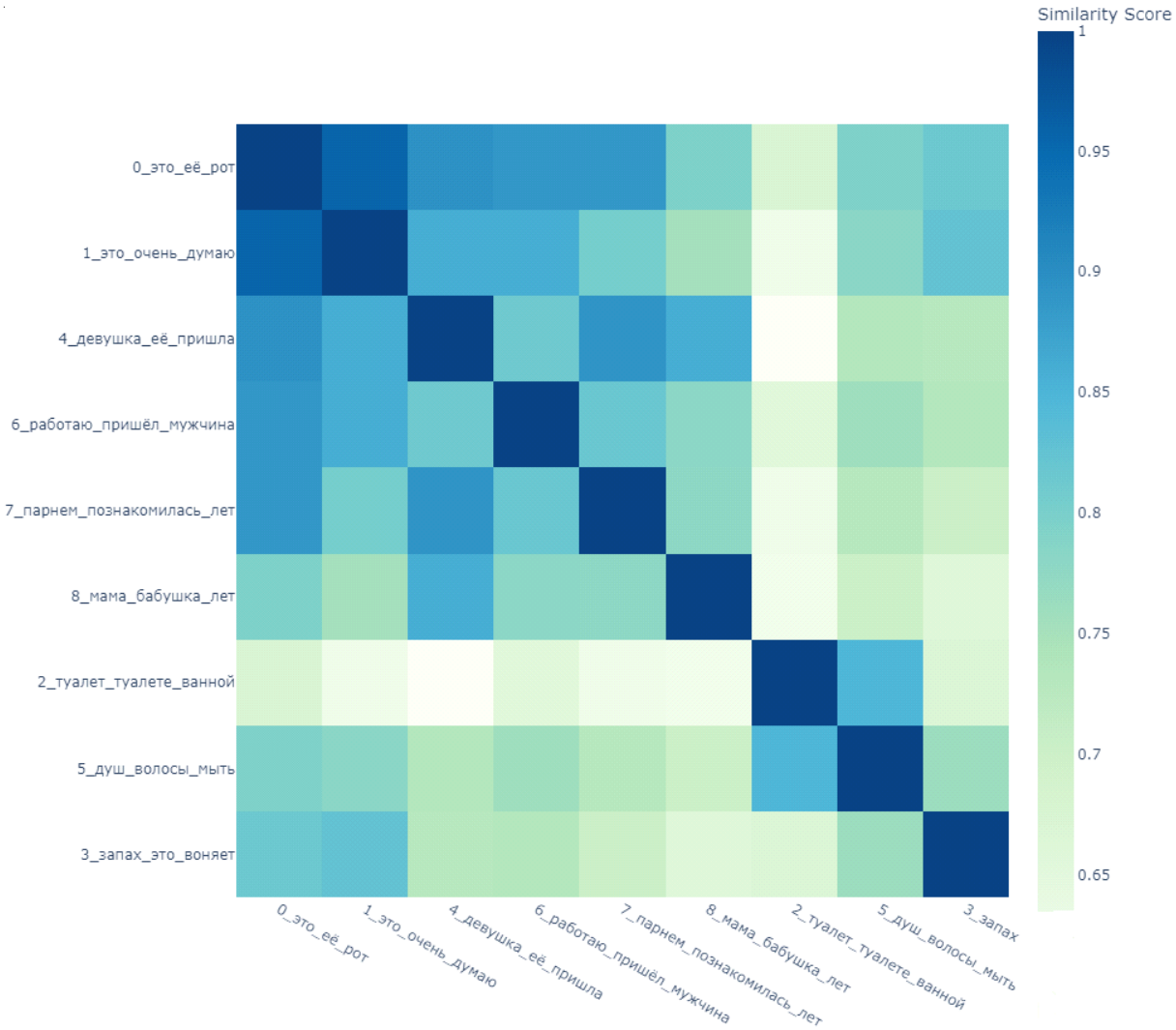

Fig. 5. Similarity matrix for topics in texts manifesting the emotion of disgust

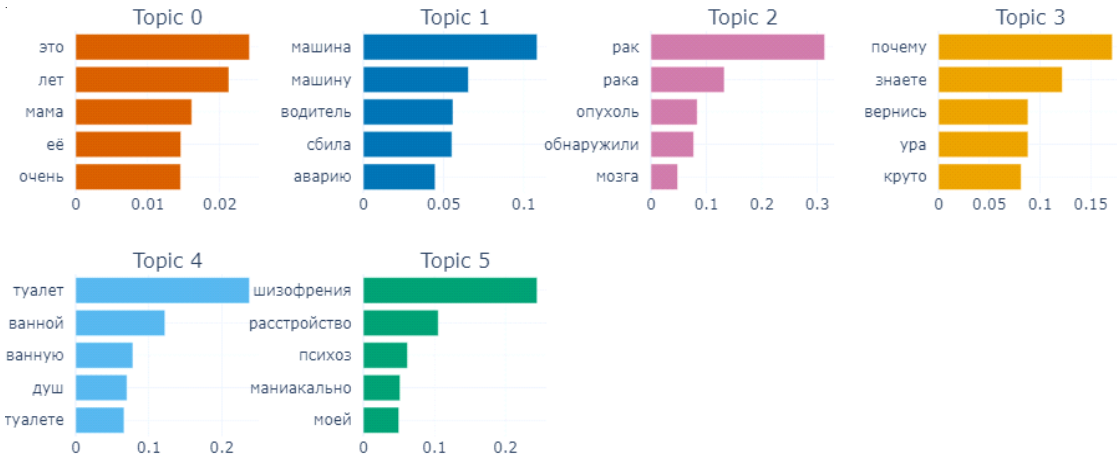

Fig. 6. Topics and terms of texts manifesting the emotion of fear

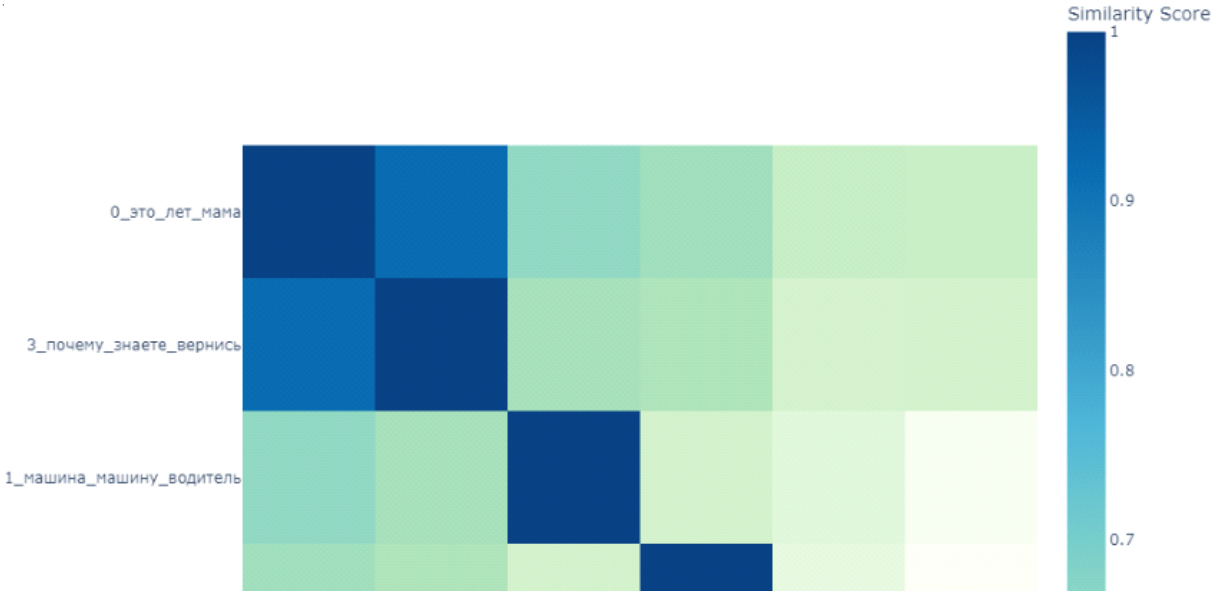

Fig. 7. Similarity matrix for topics in texts manifesting the emotion of fear

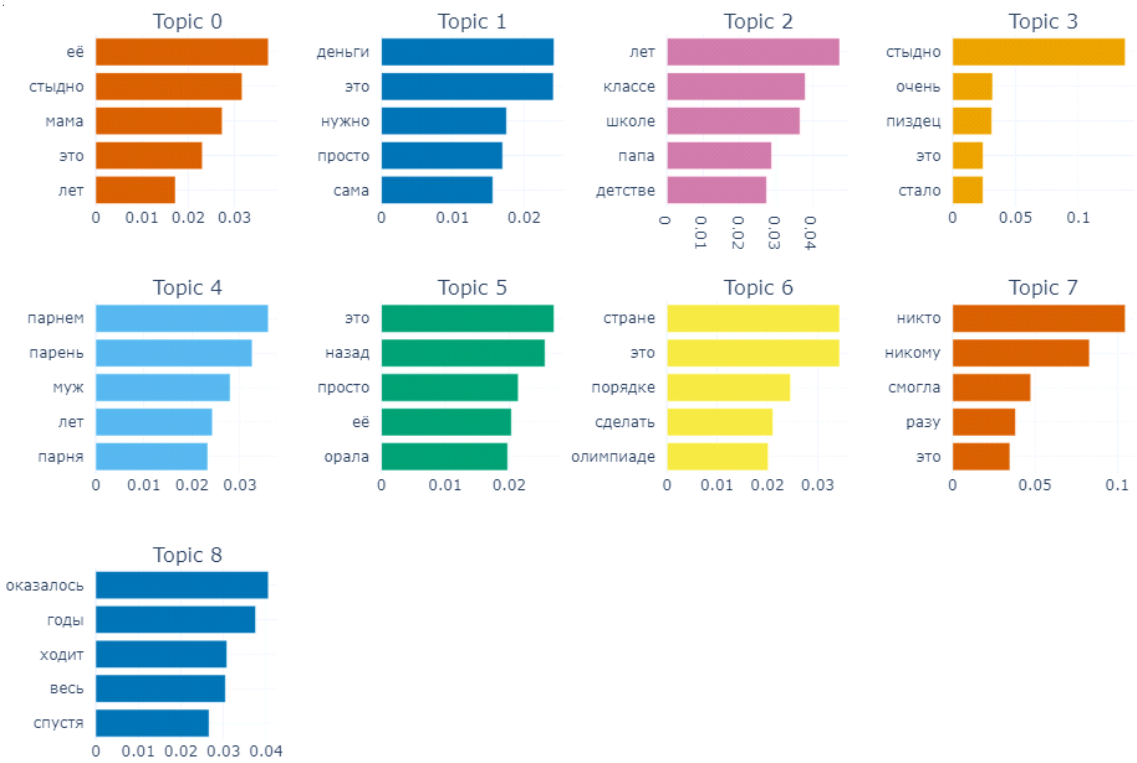

Fig. 8. Topics and terms of texts manifesting the emotion of shame

The analysis of the similarity matrix showed that Topics 5, 6, and 8 have the greatest similarity. The map of distances between topics showed that there are, in general, only two clusters: the first cluster includes Topics 5, 6, 7, and the second cluster includes all the others.

High Degree of Topic Granularity: Enjoyment and Surprise

The subcorpus of "joyful" texts was also easy to process for BERTopic model: 13 topics were identified (Fig. 9).

The topics highlighted by the model are quite harmoniously related: work (Topic 0), happiness (Topic 1, 7), family (Topic 3), love (Topic 4–5), having an apartment (Topic 6), pregnancy (Topic 8), as well as more frequent reasons for happiness – dental care (Topic 2), hair color (Topic 9), getting rid of extra pounds (Topic 10), successful surgery (Topic 11), birthday (Topic 12).

The distance between the topics (Fig. 10) shows that they form four clusters: the cluster of Topics 7, 1, 9 (happiness + hair color), the leftmost cluster – Topics 5, 11, 8 (girl, surgery, pregnant), the next – Topics 2 and 0, and the fourth – Topics 3, 4, 6, 10 (family+love+flat+weigt loss).

The similarity matrix shows that Topics 2, 5, 7, 9, 12 are the most similar, which seems to reflect the most frequent reasons for happiness: love, changes in hairstyle, birthday.

The largest number of topics was highlighted by the model on the corpus of "surprising" texts (Fig. 11).

Fig. 9. Topics and terms of texts manifesting the emotion of enjoyment

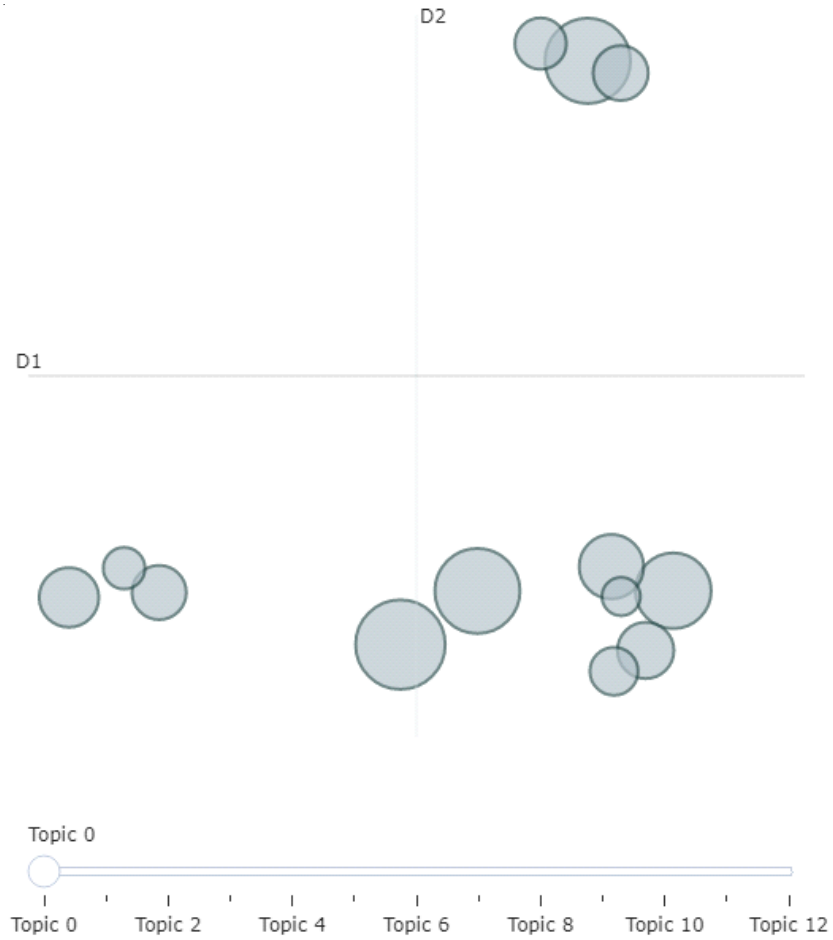

Fig. 10. Intertopic distance map in the subcorpus of texts manifesting the emotion of enjoyment

Note that compared to the previous subcorpora, these topics proved difficult to interpret, but even here we can identify ideas of family (Topic 4), relationships with the opposite sex (Topic 8, 9, 10), fear (Topic 14), shopping (Topic 12), preferences in smells and tastes (Topic 5), and own opportunities (Topic 2).

As the analysis of the similarity matrix showed, Topics 0, 1, and 2 are the most similar to each other – they are related to one proposition, which can be briefly characterized as overcoming oneself: "I thought it would not work, I cannot, but then I started doing it and everything worked out".

As for the distances between topics (Fig. 12), the model identified three clusters.

Topics 0, 3, 6, and 4 were in the first cluster; Topics 9, 8, and 10 were in the second cluster; and all the others were in the third cluster.

Conclusion

The experiment on the application of the neural network model of topic modeling to the pool of emotional texts from social networks allowed us to draw several conclusions.

Firstly, from the point of view of texts' affordances for clustering we can distinguish three groups of emotional texts: "angry", "sad" and "excited" texts cause difficulties in clustering, so they have maximum 3 topics; "shameful", "terrible" and "disgusting" are clustered relatively well – 8–9 topics are identified; finally, "happy" and "surprising" are easily clustered – 12–15 topics.

This observation provides us not so much with some technical information about the specificity of BERTopic, as with the structure and semantics of the texts of different emotional classes. Apparently, it is worth talking about the correlation between the degree of thematic granularity of the texts and the nature of the emotions they are meant to convey: low, middle and high.

Secondly, the topics selected by the model really reflect the specificity of emotional experiences and can be used in the future as

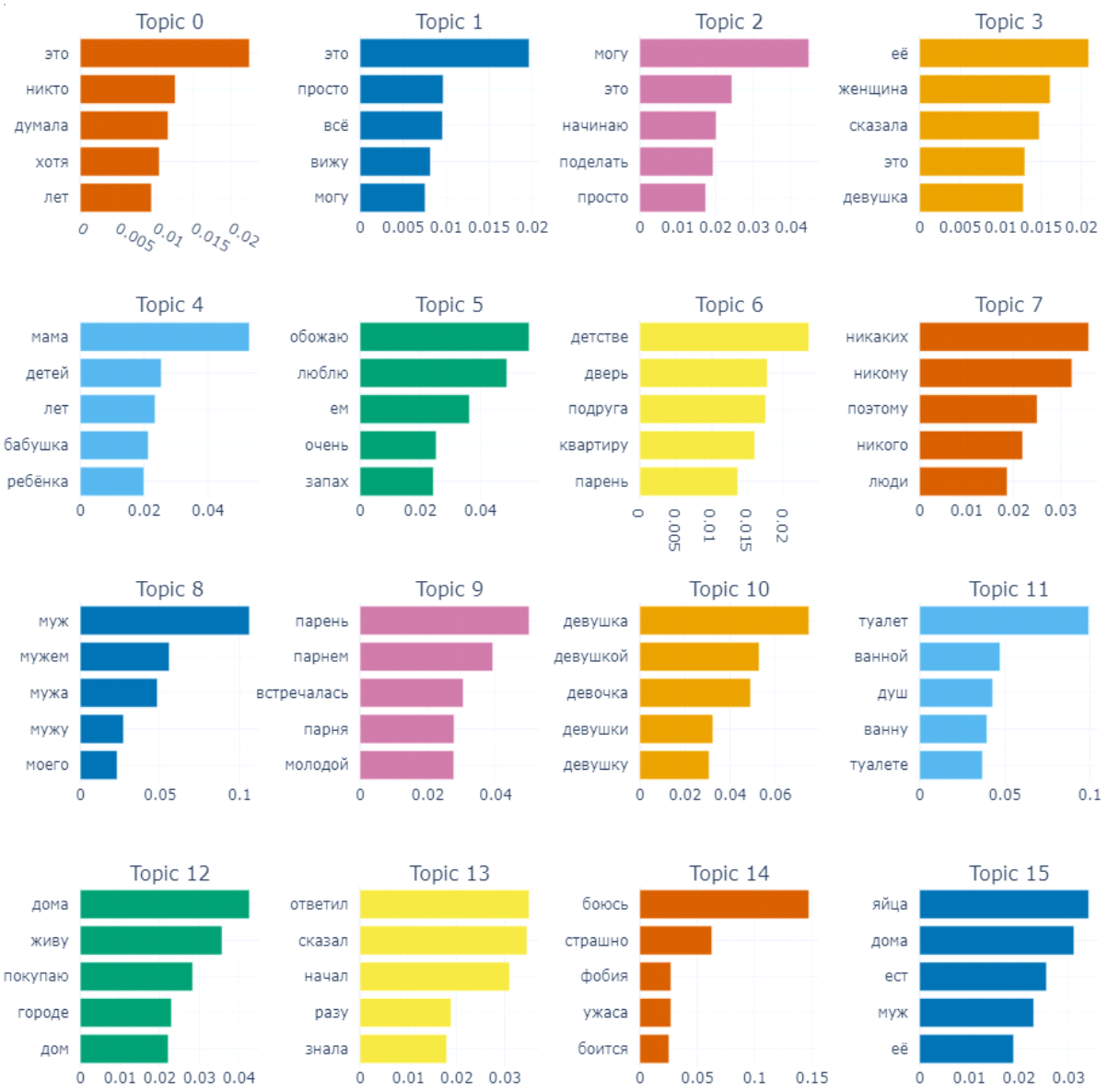

Fig. 11. Topics and terms of texts manifesting the emotion of surprise

attributes for the emotional classification, for the automatic detection of emotions in the texts of social networks.

Third, in the experiment conducted, the hypothesis that there is a correlation between the emotion expressed in the text and the nature of the thematic content was generally confirmed. Despite the fact that a number of topics are "cross-cutting" (e.g., the topic of weight loss is characteristic of both excitement and joy subcorpora, and the topic of relations with the opposite sex is characteristic of happiness, shame, and disgust text classes), the majority of topics are specific (e.g., the topic of time gone is characteristic only of shame text class, and the topic of gastronomy preferences is characteristic only of surprise class).

Thus, we can conclude that the use of the method of topic modeling is a relevant way to describe the specificity of semantics, the thematic deployment of texts of different emotional classes.

NOTE

[1] The article was prepared based on the materials of the project "Text as Big Data: Methods and Models of Working with Big Text Data", which is carried out within the framework of the Fundamental Research Program of the National Research University Higher School of Economics (HSE University) in 2024.

Fig. 12. Intertopic distance map in the subcorpus of texts manifesting the emotion of surprise

REFERENCES

Belyanin V.P., 2000. *Osnovy psikholingvisticheskoy diagnostiki (Modeli mira v literature)* [Foundations of Psycholinguistic Diagnostics (World Models in Literature)]. Moscow, Trivola Publ. 248 p.

Blei D.M., Ng A.Y., Jordan M.I., 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, no. 3, pp. 993-1022.

Grootendorst M., 2022. *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*. DOI: 10.48850/arXiv.2203.05794

Hakak N., Mohd M., Kirmani M., Mohd M., 2017. Emotion Analysis: A Survey. *Proceedings of the International Conference on Computer. Communication and Electronics*, 1–2 July, pp. 397-402. DOI: 10.1109/COMPTELIX.2017.8004002

Kolmogorova A., Kalinin A., Malikova A., 2019. Tipologiya i kombinatorika verbalnykh markerov razlichnykh emotsionalnykh tonalnostey v internet-tekstakh na russkom yazyke [Types and Combinatorics of Verbal Markers of Different Emotional Tonalities in Russian-Language Internet Texts]. *Vestnik Tomskogo gosudarstvennogo universiteta* [Tomsk State University Journal], vol. 448, pp. 48-58. DOI: 10.17223/15617793/448/6

Kositsina Yu.V., 2013. *Statiko-dinamicheskaya model tematicheskoy organizatsii monologicheskogo dialektnogo teksta: avtoref. dis. ... kand. filol. nauk* [Statical and Dynamical Model of Topical Organization of Monological Text in Dialects. Cand. philol. sci. diss.]. Kemerovo. 213 p.

Li H., Pang N., Guo S., Wang H., 2007. Research on Textual Emotion Recognition Incorporating Personality Factor. *ROBIO 2007: IEEE International Conference on Robotics and Biomimetics*, pp. 2222-2227.

Lövheim H., 2012. A New Three-Dimensional Model for Emotions and Monoamine Neuro-Transmitters. *Medical Hypotheses*, vol. 78, pp. 341-348.

Picard R., 1997. *Affective Computing*. Cambridge, The MIT Press. 306 p.

Sia S., Dalmia A., Mielke S.J., 2020. *Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics Too!* DOI: 10.85550/arXiv.abs/2004.14914

Shakhovskiy V.I., 2010. *Emotsii: dolingvistika, lingvistika, lingvokulturologiya* [Emotions: Protolinguistics, Linguistics, Lingvoculturology]. Moscow, Librokom Publ. 128 p.

Information About the Authors

Anastasia V. Kolmogorova, Doctor of Sciences (Philology), Professor, Head of the Laboratory of Language Convergence, HSE University, kanala Griboyedova Emb., 119–121, 190068 Saint Petersburg, Russia, akolmogorova@hse.ru, https://orcid.org/0000-0002-6425-2050

Qiuhua Sun, Doctor of Sciences (Philology), Professor, Head of the Department for International Cooperation, Heilongjiang University, Prosp. Xuefu, 74, Harbin, China, sunqiuhua15@163.com, https://orcid.org/0000-0002-1959-7180

Информация об авторах

Анастасия Владимировна Колмогорова, доктор филологических наук, профессор, заведующая лабораторией языковой конвергенции, Национальный исследовательский университет «Высшая школа экономики», наб. канала Грибоедова, 119–121, 190068 г. Санкт-Петербург, Россия, akolmogorova@hse.ru, https://orcid.org/0000-0002-6425-2050

Цюхуа Сунь, доктор филологических наук, профессор, руководитель департамента международного сотрудничества, Хэйлундзянский университет, просп. Сюэфу, 74, г. Харбин, Китай, sunqiuhua15@163.com, https://orcid.org/0000-0002-1959-7180