Clustering Thematic Information in Social Media

Mikhail Ulizko [1,2], Aleksey Artamonov [2], Julia Fomina [2], Evgeniy Antonov [2], Rufina Tukumbetova [1,2]

[1] *Plekhanov Russian University of Economics, Stremyannyy Pereulok, 36, 115093 Moscow, Russia*
[2] *National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashira Hwy, 31, 115409 Moscow, Russia*

Abstract

The constant growth in the number of users of the Internet and the improvement in technical capabilities of communications allow the use of various tools for the rapid notification of the population about the events occurring in the world. Depending on the type of source, models of information dissemination differ. When analyzing the information it is necessary to determine the relationship of signal distribution channels, determination of the primary source, etc.

The article examines the dissemination of information messages in open networks using messages on religious topics using visual analytics. The paper specifies the ways to identify the messages of the required topic, as well as the visualization of the content. For topic modelling Latent Dirichlet Allocation (LDA) is used. The applicability of various dimensionality reduction and clustering algorithms for the interpretation of clustering results is considered. The developed methods can be scaled to analyze information events in different thematic areas.

Keywords

Cluster analysis, Latent Dirichlet Allocation, K-Means, DBSCAN, HDBSCAN, dimensionality reduction, t-SNE, Principal Component Analysis, religion, Telegram.

1. Introduction

Under the informatization, modern society is becoming more involved in the events taking place in the world. Information signals (individual events), disseminated in various ways, can have both positive and negative impacts on the population [1].

The amount of generated data provided to users is increasing daily that has several features [2]. First, an user is not able to consider all the important events occurring in the world due to the volume of information. Second, the information space divides users by the content they receive, based on a smart news feed.

All of the above points to the need to create tools for analyzing the information posted online. However, under present-day conditions it is not enough to analyze individual messages and posts - the information signal often consists of the chain of interrelated events, such as reactions to the published news and its redirection via other channels.

It should be emphasized that it is not possible to consider the entire information space (the Internet) for several reasons (limited research time and costs for data collection, storage and analysis), so the instant messaging platform Telegram was chosen as an information channel. This platform has become particularly popular among Russian users in 2022.

As an example, the paper considers the analysis of the information by messages related to religions. The objectives of the study are to divide the information flow into clusters and analyze them.

GraphiCon 2022: 32nd International Conference on Computer Graphics and Vision, September 19-22, 2022,
Ryazan State Radio Engineering University named after V.F. Utkin, Ryazan, Russia
EMAIL: mulizko@kaf65.ru (M. Ulizko); aartamonov@kaf65.ru (A. Artamonov); ufomina@kaf65.ru (Ju. Fomina); eantonov@kaf65.ru (E. Antonov), rrtukumbetova@kaf65.ru (R. Tukumbetova)
ORCID: 0000-0003-2608-8330 (M. Ulizko); 0000-0002-9140-5526 (A. Artamonov); 0000-0002-7315-9704 (Ju. Fomina); 0000-0003-1498-9131 (E. Antonov); 0000-0002-1976-1390 (R. Tukumbetova)
© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Literature review

The task of spreading information signals in the network is relevant and has been described by various scientists. Some researchers distinguish key nodes [3], while others describe the nature of propagation itself or try to give predictions of possible signal propagation [4, 5]. In addition, a recent study [3] showed a way to statistically analyze signals collected directly from the Telegram network.

These studies are more related to social research and describe the problem in general, in detachment from the content of the information posted. In fact, content is also important.

Given that the basis of information messages in Telegram is text, there is an evident need to analyze text messages. For this purpose methods of natural text processing are usually used [6-10]. However, considering a single message in the context of determining the state of the subject area as a whole is impractical, therefore it is necessary to consider analyzing a large number of documents relevant to this search.

On the other hand, since more data give more accurate results, as much data as possible should be obtained. However, it may be related to different topics, so there is also a necessity to classify individual texts according to these topics.

Clustering algorithms (k-means, DBSCAN) can divide texts by topics, complementing them with text preprocessing [11][12]. The authors recommend word embedding [13], using a bag of words and then the reduction of the dimensionality to two-dimensional [14]. The authors also highlight the applicability of Latent Dirichlet allocation (or LDA) [15] for separating texts into different topics [16][17].

Analysis of the dissemination of information signals within an individual cluster may contain the following subtasks: the identification of "opinion leaders," the audience reached, the average time to deliver information to users, etc. However, many of such subtasks can be reduced to statistical analysis (number of subscribers, ratio of own publications to reposts, etc.), which can be implemented by statistical and software methods.

3. Methodology

3.1. Data Extraction

The information signals disseminated on the Internet affect the general public. There is a particular need to analyze both the way individual newsworthy events are disseminated and the content itself.

The information field is heterogeneous, with information dissemination patterns varying according to the medium of dissemination. We distinguished the following major media:

- online news outlet (Yandex.News, Rambler.News, RIA News, etc.),
- social networks (Vkontakte, etc.),
- microblogging Twitter,
- video hostings (YouTube, RuTube, etc.),
- instant messaging platforms (Telegram).

Each type of these channels uses different ways of generating content (mind control techniques). The ways of distribution can also be described differently, so it is necessary to consider each type of channels separately. In this paper, the Telegram instant messaging platform is chosen for consideration. The choice is related to two main factors:

1. users' activity (achieved by securing the messenger and supporting anonymity) allows them to participate in the information agenda.
2. open data (the ability to collect data) allows us to collect statistical data (number of subscribers in the channel, dates and times of publications, etc.) and content (text of messages).

The data model is an information message containing fields:

- channel information/description,
- identifier (Id) of the message in the channel,
- date and time of publication,
- text (if available),

- source (if available).

After defining the information field and presenting the data model, it is necessary to narrow the subject search, because there are more than 300 thousand channels in Telegram and more than 500 million people are registered. To specify the subject field, we chose 16 channels related to religion from which messages are collected over 2022 (25885 messages in total).

3.2. Preprocessing and data visualization

As mentioned earlier, the information environment is heterogeneous, so one of the primary tasks is to develop and implement methods of classification/clustering of information (even within a single source Telegram or its individual channels) to solve the problem of identifying distribution channels, opinion leaders, etc. Different algorithms can be used for this purpose. Since the input data are text, we will use the following sequence of actions for data clustering:

1. Matrix representation of a document set based on the term's frequency.
2. Topic modeling using LDA [15].
3. Data clustering (algorithms K-means, DBSCAN и HDBSCAN) [18][19].
4. Data visualization in the plain (algorithms t-SNE и PCA for dimensionality reduction) [20].

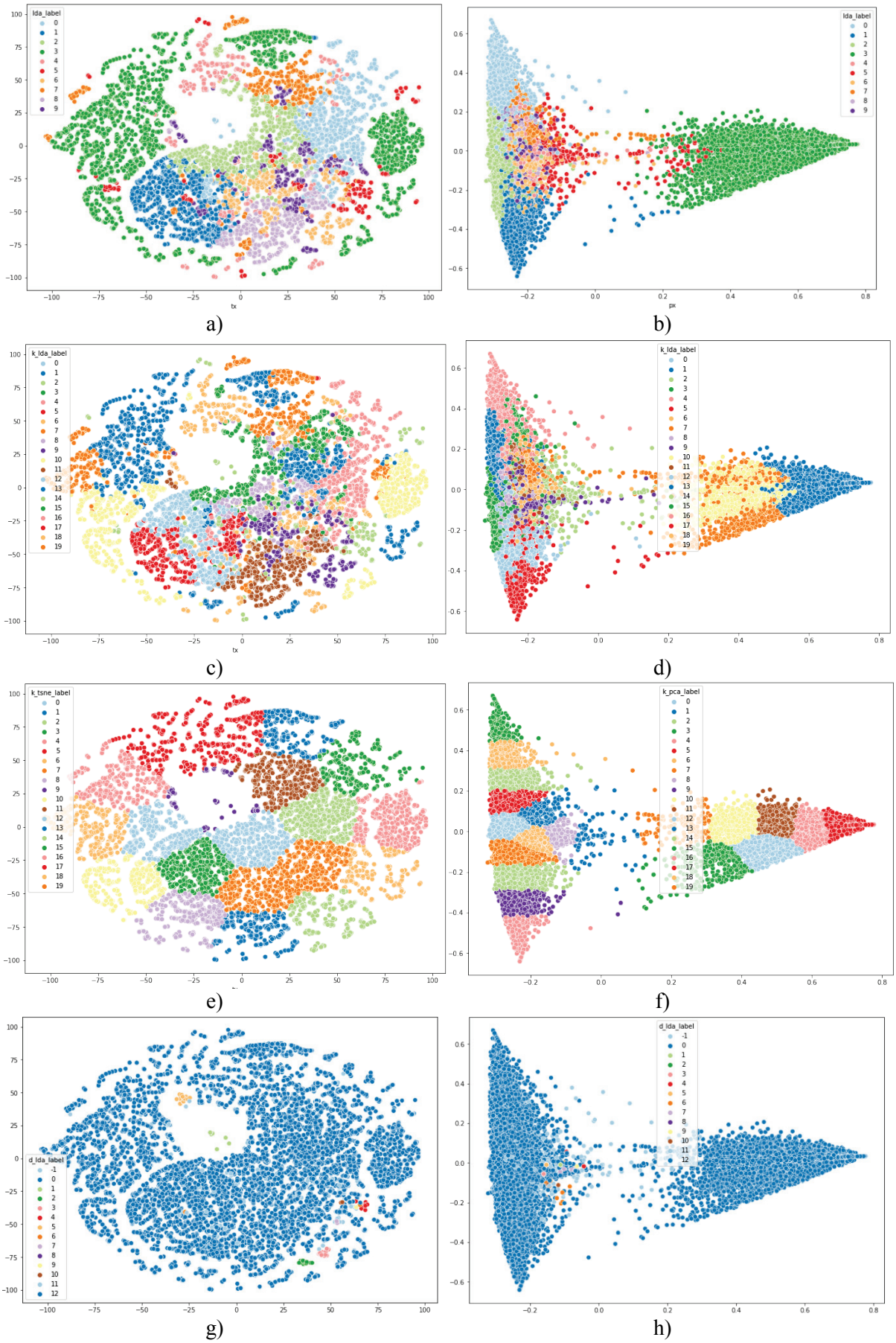
The LDA algorithm determines the probability for each found topic that the text Ti belongs to the topic J (Table 1) (values that are maximal for the line are highlighted in yellow). The following approximation is used: the topic (cluster) of text Ti is that topic J, which probability (Pij) for this text is maximal.

Table 1
LDA probabilities for 5 texts

Text/Topic	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Text 1	0.02	0.02	0.42	0.02	0.02	0.02	0.02	0.02	0.42	0.02
Text 2	0.01	0.14	0.64	0.01	0.01	0.01	0.12	0.01	0.01	0.06
Text 3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Text 4	0.01	0.01	0.13	0.01	0.01	0.01	0.19	0.01	0.39	0.25
Text 5	0.34	0.01	0.01	0.01	0.01	0.07	0.01	0.01	0.01	0.54

After preprocessing the text and LDA markup we obtain a possible partitioning into clusters. In advance, we set 10 topics for LDA and 20 clusters for K-means algorithm. Since different combinations of algorithms give different results, let us consider all possible clustering results (Figure 1):

1. LDA topics after t-SNE transformation (Figure 1a).
2. LDA topics after PCA transformation (Figure 1b).
3. K-means clustering to LDA probabilities and t-SNE transformation (Figure 1c).
4. K-means clustering to LDA probabilities and PCA transformation (Figure 1d).
5. T-SNE transformation and K-means clustering to LDA topic projections (Figure 1e).
6. PCA transformation and K-means clustering to LDA topic projections (Figure 1f).
7. DBSCAN clustering to LDA probabilities and t-SNE transformation (Figure 1g).
8. DBSCAN clustering to LDA probabilities and PCA transformation (Figure 1h).
9. T-SNE transformation and DBSCAN clustering to LDA topic projections (Figure 1i).
10. PCA transformation and DBSCAN clustering to LDA topic projections (Figure 1j).
11. HDBSCAN clustering to LDA probabilities and t-SNE transformation (Figure 1k).
12. HDBSCAN clustering to LDA probabilities and PCA transformation (Figure 1l).
13. T-SNE transformation and HDBSCAN clustering to LDA topic projections (Figure 1m).
14. PCA transformation and HDBSCAN clustering to LDA topic projections (Figure 1n).


lda_label
tx
px
k_lda_label
k_tsne_label
k_pca_label
d_lda_label
a)
b)
c)
d)
e)
f)
g)
h)

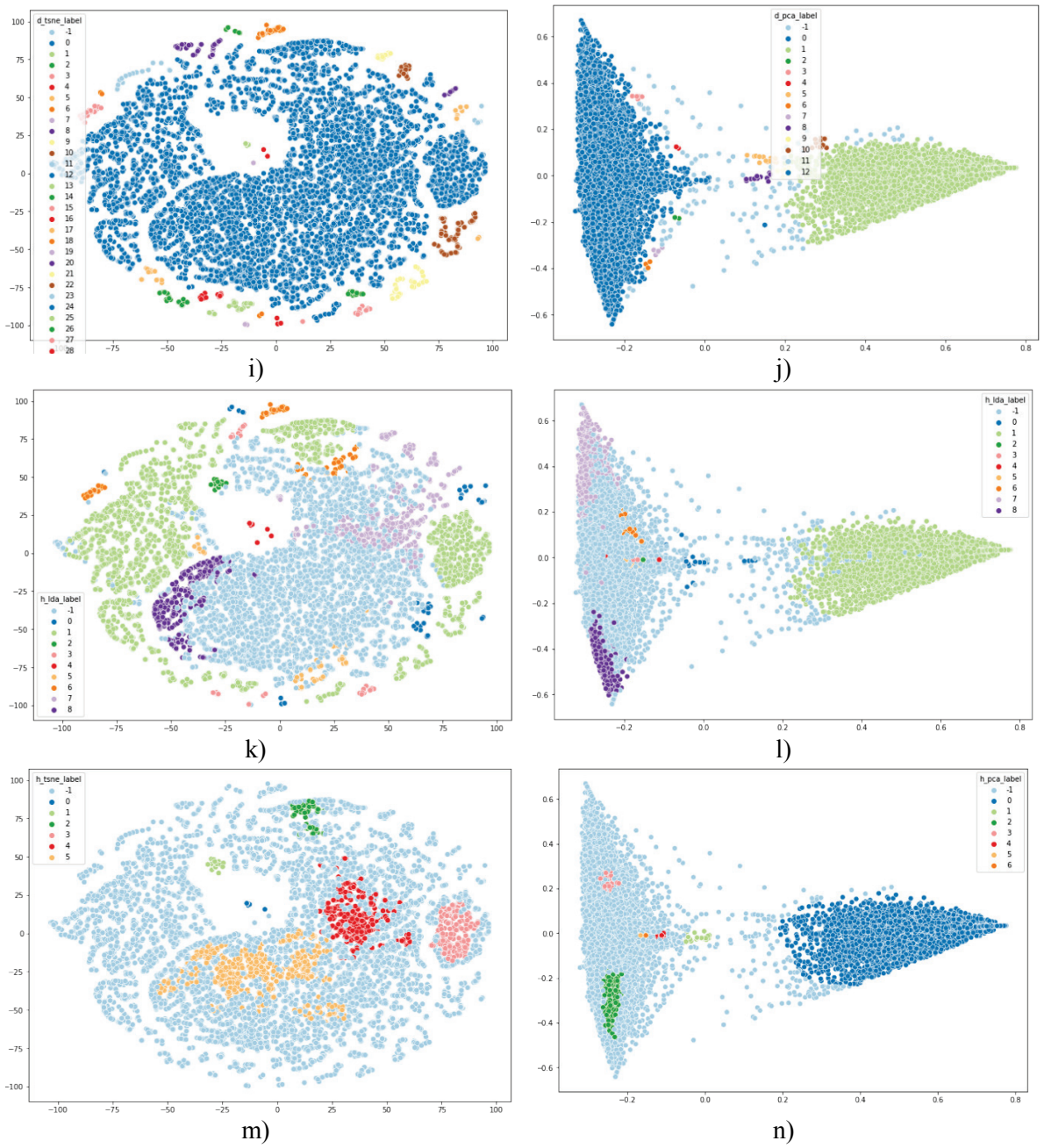


Figure 1: Clustering results

The scatter plots show a significant difference un the clustering results. However, it is evident that the LDA partitioning by topics is improving when changing to two-dimensional space (Figure 1a and Figure 1b). At the same time the K-means clustering results (Figure 1c and Figure 1d) are most similar to the LDA topics, which is due to the pre-specified number of clusters. In contrast, the DBSCAN (Figure 1g and Figure 1h) and HDSCAN (Figure 1k and Figure 1l) clustering results do not resemble the LDA results and isolate a single cluster (if we exclude the noisy "-1" cluster).

These results are confirmed by pivot tables (Figure 2, Figure 3, Figure 4) (the abscissa axis shows the cluster number of the clustering algorithm, the ordinate axis shows the LDA topic number). For the K-means clustering (plotted for LDA topic probabilities) (Figure 2), there are several one-to-one correspondences between clusters. For example, K-means cluster 6 corresponds to LDA topic 4, cluster 14 to topic 5, cluster 9 to topic 6, etc.

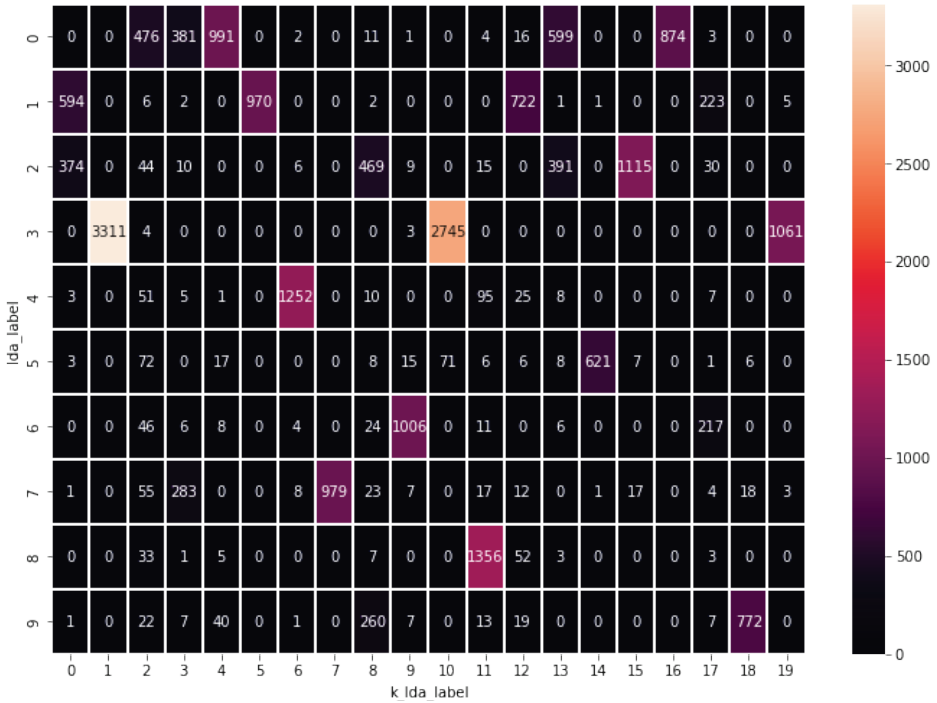


Figure 2: Pivot Table LDA-Kmeans

On the other hand, the DBSCAN (Figure 3) and HDBSCAN (Figure 4) algorithms arrange most of the data into no more than 2 clusters, with the largest cluster of algorithms uniformly distributed across the LDA topics.

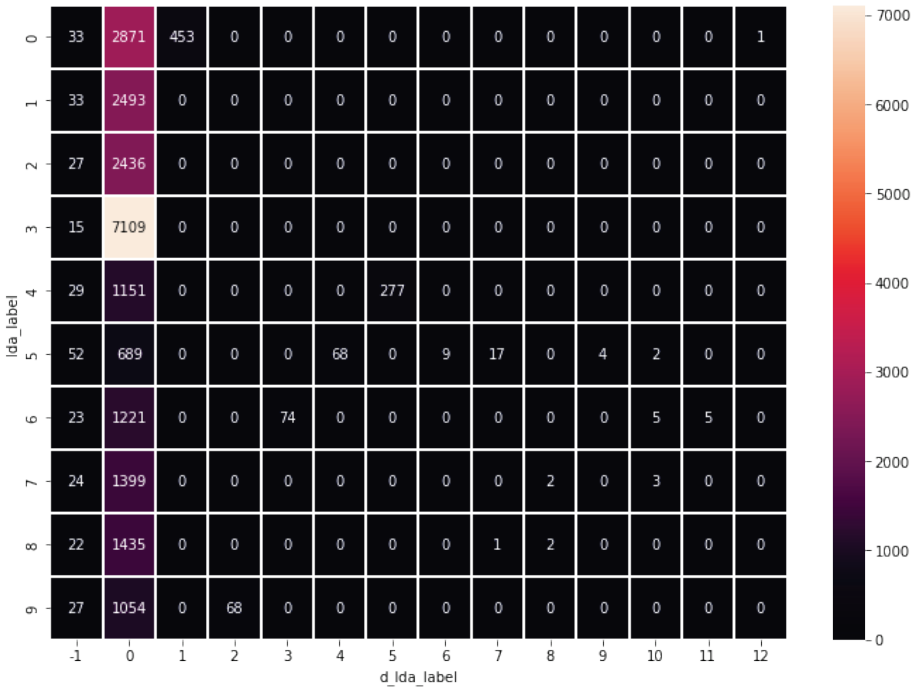


Figure 3: Pivot table LDA-DBSCAN

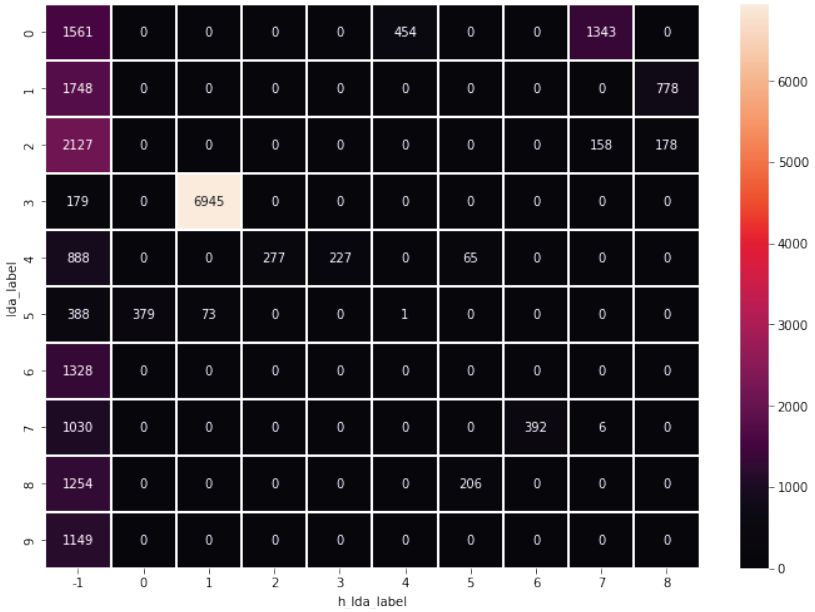


Figure 4: Pivot table LDA-HDBSCAN

Since the results of clustering differ from each other, it is necessary to compare the obtained clusters by their content. The tag cloud is taken for evaluation. An example of the tag cloud for 2 LDA topics is shown below (Figure 5), which proves the possibility of the identification of individual topics using LDA. Despite the difference in the number of clusters among clustering algorithms, each cluster can be assigned to one of the 3 topics:

- texts related to religion,
- Ukrainian texts related to religion,
- other.

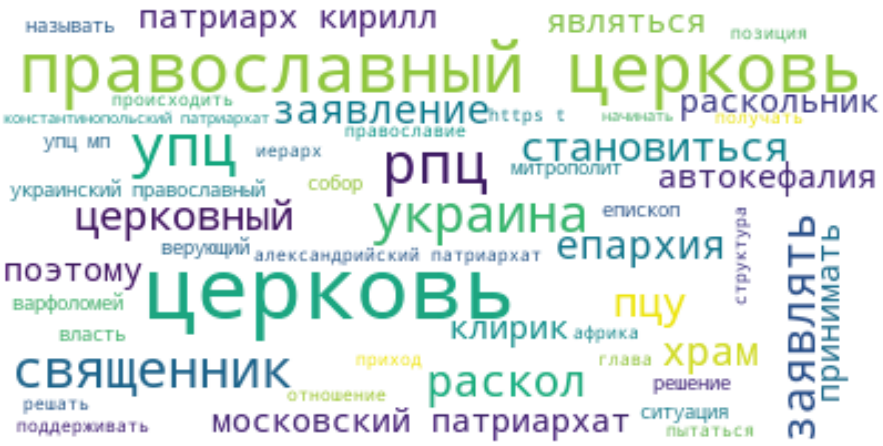


Figure 5: Tag cloud for LDA topic "2"

On the other hand, let us put forward the hypothesis that the channels publish information only on a certain topic. This hypothesis is supported by common sense and the name of the channels. To test the hypothesis let us build a pivot table (Figure 6). The abscissa axis is the Telegram Channel ID, the ordinate axis is the LDA subject. Separately, three fields are added: "text" - number of messages (own post or repost), "repost" - number of reposts, and "total" - total number of messages in the channel.

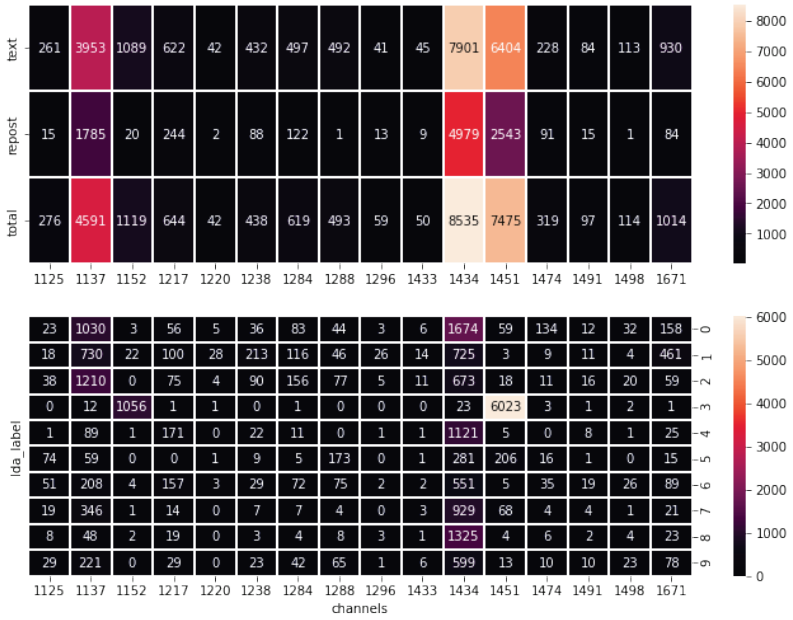


Figure 6: Pivot table Channels-LDA

The following peculiarities should be noted. First, there are channels that write on several topics (channels 1137 and 1434), and topics that several channels write on (topics 0, 1, 2). We will also distinguish channels that often refer to other channels (1137, 1434, and 1451). Finally, there is a narrowly focused channel - channel 1451 publishes more than 90% of its posts on one topic.

4. Results and Discussion

Nowadays natural language processing is an evolving and complex field. One of the difficulties is the interpretation of the results. The paper revealed several features which are significant for analysis.

First of all, in the study we fixed the number of LDA topics using the specified assumption. However, the true number of topics can be determined based on the LDA probabilities. To accomplish this, it is necessary to find topics that give similar probabilities on most examples, and then combine them.

Secondly, the result of K-means clustering in LDA probabilities space finds more approximations to the original topics than the DBSCAN and HDBSCAN algorithms. This is due to the equality of the number of topics and K-means clusters and the complexity of adjusting the parameters of the other two algorithms (DBSCAN and HDBSCAN). However, the advantage of the DBSCAN and HDBSCAN algorithms is the automatic detection of topics - one main cluster is defined for the original data.

Also it should be taken into consideration that clustering by any of the proposed methods can be performed both in the original feature space, and in the space obtained as a result of the t-SNE and PCA algorithms. Despite the difference in the clustering results (Figure 1), in each case individual clusters belong to exactly one topic.

The experiment was performed on data on a single topic, which was confirmed by the results of clustering using DBSCAN and HDBSCAN algorithms. It is also possible to conduct an experiment when the original data describe several topics. For this purpose, we added texts on "machine learning" and "politics" to the dataset.

Since visualization of clusters does not give the understanding of the partitioning by clusters, a pivot table for the new data is presented (Figure 7). Added channels have some features:

- channel "1262" writes mainly on 1 topic (machine learning),
- channel "1429" writes on various topics (politics),
- channel "1713" is not representative.

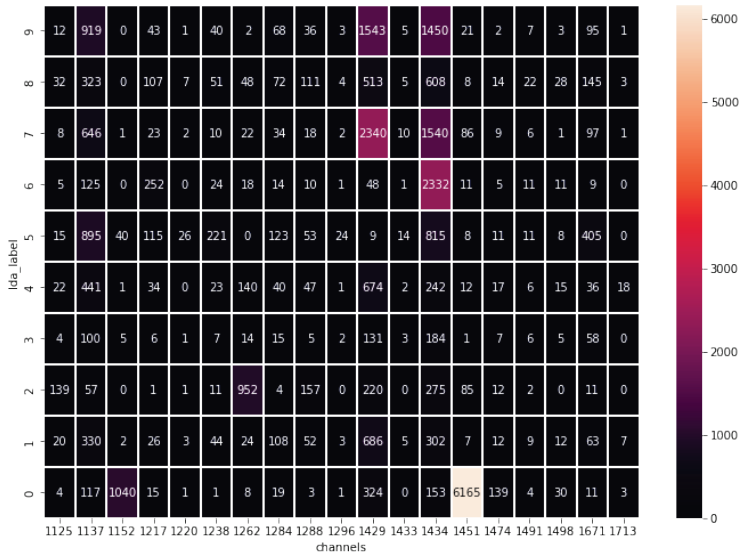


Figure 7: Pivot table LDA-Channels

Tag clouds can be built to discover new topics, but the topic "politics" is broad, so only the topic "machine learning" is visualized. The best cluster extraction result is achieved by applying the DBSCAN algorithm for LDA probabilities projections into t-SNE space (Figure 8). The K-means algorithm was unable to fully isolate the topic of "machine learning" in any of the three spaces.

Figure 8: Tag cloud for "machine learning" cluster

5. Conclusion

With the increasing amount of information and easier access to it, there is a need to track and analyze emerging data. Since different types of information sources have different opportunities for publishing and distributing content, it seems reasonable to perform analysis for each type of source separately.

For the Telegram messaging platform, the basic unit is an individual message that can be described by Telegram tools not only with content (text), but also with additional information about the date/time of publication, the original source, etc. In this case, the substantive part is the text message, which can be analyzed using NLP methods.

This paper presents a processing algorithm that divides publications into clusters and visualizes them. This algorithm uses K-means, DBSCAN, HDBSCAN, or any other clustering algorithm. DBSCAN and HDBSCAN are more adapted to the definition of large topics, independently determining the clusters and their number. The K-means algorithm can be used if the number of clusters is known in advance and they are approximately equal in cardinality.

The authors found that dimensionality reduction by any of the proposed algorithms does not allow proper interpretation of the results. However, the clustering for the original probability space and for the probability projections on the plane work approximately the same, even despite the visual differences. This confirms that dimensionality reduction algorithms retain most of the information about an object.

6. References

[1] Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. In Public Opinion Quarterly (Vol. 80, Issue S1, pp. 298–320). Oxford University Press (OUP). doi:10.1093/poq/nfw006.

[2] Liang, Y., Guo, N., Xing, C., Zhang, Y., & Li, C. (2015). Multilingual Information Retrieval and Smart News Feed Based on Big Data. In 2015 12th Web Information System and Application Conference (WISA). 2015 12th Web Information System and Application Conference (WISA). IEEE. doi:10.1109/wisa.2015.44.

[3] Ulizko, M. S., Artamonov, A. A., Tukumbetova, R. R., Antonov, E. V., & Vasilev, M. I. (2022). Critical Paths of Information Dissemination in Networks. In Scientific Visualization (Vol. 14, Issue 2). National Research Nuclear University MEPhI (Moscow Engineering Physics Institute). doi:10.26583/sv.14.2.09.

[4] Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks. In ACM SIGMOD Record (Vol. 42, Issue 2, pp. 17–28). Association for Computing Machinery (ACM). doi:10.1145/2503792.2503797.

[5] Gomez-Rodriguez, M., Leskovec, J., & Krause, A. (2012). Inferring Networks of Diffusion and Influence. In ACM Transactions on Knowledge Discovery from Data (Vol. 5, Issue 4, pp. 1–37). Association for Computing Machinery (ACM). doi:10.1145/2086737.2086741

[6] Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. In IEEE Computational Intelligence Magazine (Vol. 13, Issue 3, pp. 55–75). Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/mci.2018.2840738.

[7] Diego Lopez Yse, Your Guide to Natural Language Processing (NLP), 2019. URL: https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1.

[8] Tretyakov, E., Savić, D., Korpusenko, A., & Ionkina, K. (2022). Sentiment Analysis of Social Networks Messages. In Studies in Computational Intelligence (pp. 552–560). Springer International Publishing. https://doi.org/10.1007/978-3-030-96993-6_61.

[9] Kulik, S. D., Belov, A. N., & Matveeva, K. I. (2018). Development of generation special short articles for the given topic. In International Journal of Engineering & Technology (Vol. 7, Issue 2.23, p. 171). Science Publishing Corporation. https://doi.org/10.14419/ijet.v7i2.23.11909.

[10] Grin, D., Grigorieva, M., & Artamonov, A. (2021). Visual Analysis Application for the Error Messages Clustering Framework. In Procedia Computer Science (Vol. 190, pp. 274–283). Elsevier BV. https://doi.org/10.1016/j.procs.2021.06.037.

[11] Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. In Information Processing & Management (Vol. 57, Issue 2, p. 102034). Elsevier BV. doi:10.1016/j.ipm.2019.04.002.

[12] Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T., & Bian, J. (2021). Evaluation of clustering and topic modeling methods over health-related tweets and emails. In Artificial Intelligence in Medicine (Vol. 117, p. 102096). Elsevier BV. doi:10.1016/j.artmed.2021.102096.

[13] Korogodina, O., Karpik, O., & Klyshinsky, E. (2020). Evaluation of Vector Transformations for Russian Word2Vec and FastText Embeddings. In Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2 (pp. paper18-1-paper18-12). MONOMAX Limited Liability Company. doi:10.51130/graphicon-2020-2-3-18.

[14] Spathis, D., Passalis, N., & Tefas, A. (2019). Interactive dimensionality reduction using similarity projections. In Knowledge-Based Systems (Vol. 165, pp. 77–91). Elsevier BV. doi:10.1016/j.knosys.2018.11.015.
[15] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(4-5), 993-1022.
[16] Susan Li, Topic Modeling and Latent Dirichlet Allocation (LDA) in Python, 2018. URL: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24.
[17] Medium, Topic modelling using LDA, 2021. URL: https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13.
[18] George Seif, The 5 Clustering Algorithms Data Scientists Need to Know, 2018. URL: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68.
[19] R. J. G. B. Campello, D. Moulavi, A. Zimek and J. Sander, "Hierarchical density estimates for data clustering visualization and outlier detection", ACM Trans. Knowl. Discov. Data, vol. 10, no. 1, pp. 5:1-5:51, Jul. 2015.
[20] Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). In Computer Science Review (Vol. 40, p. 100378). Elsevier BV. doi:10.1016/j.cosrev.2021.100378.