Quantitative Analysis of Propagandistic Narratives in Large Text Corpses Using Machine Learning Methods

Illia Dagil[1,*,†], Iryna Vergunova[1,†] and Yaroslav Tereshchenko[1,†]

[1] *Taras Shevchenko National University of Kyiv, Akademika Hlushkova Av. 4d, 03680 Kyiv, Ukraine*

Abstract

This paper presents a novel algorithm for topic modeling, specifically designed to identify and analyze propaganda narratives in large-scale news corpora. The algorithm combines advanced natural language processing techniques, embedding models, and clustering algorithms to assist analysts, communication experts, and government agencies in efficiently processing and identifying propaganda content. A series of experiments was conducted on multiple datasets to thoroughly test the algorithm's performance. In these experiments, five different embedding models were compared along with four clustering algorithms, each tested with various hyperparameters. A significant challenge addressed was determining the appropriate granularity of clusters, balancing between detailed analysis and broader trends. Additionally, narrative extraction was deeply investigated using large language models (LLMs) providing accurate and structured identification of complex narratives. This approach allows not only the identification of propaganda but also the development of counter-narratives, with the potential to be adapted for broader applications such as communication network analysis.

Keywords

Topic Modeling, News Analysis, Natural Language Processing, Embedding Model, Large Language Model, Clustering

1. Introduction

Propaganda and disinformation are among the most significant challenges facing the modern information environment. In a time when people have access to an overwhelming amount of information, the manipulation of facts and the spread of false narratives can have far-reaching effects. These include shaping public opinion, influencing election outcomes, impacting international relations, and even justifying conflicts.

Disinformation is often used as a geopolitical tool, turning the media into a battleground. Although propaganda is not a new phenomenon, modern technologies and social media platforms have enabled it to spread faster and more widely than ever before [1-4].

Analyzing propaganda narratives and disinformation campaigns is essential to defending democratic societies and ensuring information security. Upholding objectivity, information reliability, and source transparency is not only an academic endeavor but also a matter of national security [5-10].

In the current era of information warfare, effectively combating propaganda and disinformation is critical. To achieve this, a comprehensive analysis is needed. One of the key elements of this analysis is identifying propaganda narratives and assessing their prevalence. Natural language processing (NLP) and traditional machine learning techniques [10-13] can be applied to handle large volumes of text efficiently. Also, the study of propaganda has become a highly relevant and timely

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

* Corresponding author.

† These authors contributed equally.

illia.i.dagil@gmail.com (I. Dagil); vergunova@hotmail.com (I. Vergunova); y.ter@gmail.com (Y. Tereshchenko)

0000-0003-3874-6206 (I. Dagil); 0000-0003-3052-9143 (I. Vergunova); 0000-0002-8451-7634 (Y. Tereshchenko)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3933/Short_1.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

topic within the academic community, drawing significant attention due to its critical implications in today's information landscape [1-5]. This approach not only saves time and resources but also provides a more objective and unbiased analysis compared to manual review.

2. Problem Statement

The primary goal of this research is to develop an efficient method for identifying and measuring the prevalence of propaganda narratives within large news corpora. The objective is to accurately detect and quantify narratives, presenting the results as a ranked list based on the frequency of occurrence within the dataset. Furthermore, the potential audience reach for each narrative must be estimated to assess the broader impact of these narratives. To enhance the understanding of how these narratives evolve and spread, an infographic will be created to visually represent their dissemination patterns over time. This visualization will help highlight the influence of key narratives across different regions, channels, and time frames, offering insights into their propagation and reception. The ultimate aim of the research is to provide a tool that can support more informed decision-making by analysts, policymakers, and communication experts, enabling them to counteract disinformation and propaganda more effectively.

3. Algorithm

In this section, we introduce a comprehensive algorithm designed for the analysis and identification of propaganda narratives within large corpora of news texts. The algorithm leverages natural language processing and machine learning techniques to automate the detection of narratives and evaluate their prevalence across different datasets. By combining large language models, embedding models, and clustering algorithms, the method provides a systematic approach to dissecting complex narratives, offering insights into how propaganda themes evolve and spread. The following steps outline the key stages of the algorithm and the models used to achieve this.

3.1. Data Collection

Assume we have access to a corpus of all existing news texts, and the messages we need to analyze are a subset of this corpus. Selecting the appropriate subset is a crucial step in the algorithm. This selection can be made based on various criteria or a combination of them, such as:

- The publication time frame of the news,
- The source of the news (specific social networks, resources, channels, etc.),
- The presence of certain keywords.

3.2. Identifying Propaganda Narratives

Each news item may either contain no propaganda narratives (e.g., "Meteorologists predict rain and strong winds in region N") or include multiple narratives (e.g., "US Navy exercises near Taiwan cross all of China's red lines, as they are an infringement on the territorial integrity of the PRC"). The extraction of these narratives is done using a large language model (LLM). This choice is based on several objective reasons: the most advanced LLMs are capable of following instructions, reformulating, and translating texts into English while maintaining the original meaning. Additionally, LLMs have larger context windows, allowing them to process longer texts more effectively than other neural networks, including transformer-based architectures. They can also provide structured responses (e.g., JSON format), which allows easy parsing. To mitigate the risk of hallucinations, techniques such as prompt engineering and evidence-based model outputs can be applied.

```
{
      "narratives": [
         {
            "narrative_analysis": "The narrative asserts that Taiwan
is an integral part of the People's Republic of China (PRC) and that
US naval exercises near Taiwan are a violation of China's territorial
integrity. This narrative aligns with the Chinese government's
position on Taiwan and seeks to delegitimize US military activities
in the region by framing them as aggressive and disrespectful to
Chinese sovereignty.",
            "based_on_text": "US Navy exercises near Taiwan cross all
of China’s red lines as they infringe on the territorial integrity of
the PRC."
         },
         {
            "narrative_analysis": "The narrative suggests that the US
Navy's military exercises near Taiwan are a deliberate provocation
that crosses China's 'red lines.' This language is intended to
portray the United States as a destabilizing force in the region and
to justify potential retaliatory actions by China as defensive
measures.",
            "based_on_text": "US Navy exercises near Taiwan cross all
of China’s red lines."
         }
      ]
}
```


Figure 1: Example narrative analysis in JSON format.

The above example was generated using the GPT-4 model. The value under the “narratives” key is always a list, allowing for the extraction of any number of narratives (from zero to multiple). The “narrative_analysis” field contains a detailed explanation of the narrative, while the “based_on_text” field provides the exact quote from which the narrative was derived. If this matches the original news text, the analysis can be considered validated to some extent. This structured format also helps to minimize hallucinations.

3.3. Creating Vector Representations of Narrative Analysis

Suppose we have two narrative analyses. To compare their similarity, we need to define a similarity metric. While we could directly compare the words used in the analyses, this approach may reduce the quality of comparison because it would ignore the sequence and context of the words. Two contrasting examples illustrate the limitations of this approach:

- A set of identical narratives expressed with different wording and phrasing.
- A set of narratives that use the same words but describe opposing viewpoints (e.g., "Russia is conducting terrorist acts in Ukraine" vs. "Ukraine is conducting terrorist acts in Russia").

Modern embedding models can solve these issues by representing texts as vectors in latent space, preserving the semantic meaning of the text. As a result, we can create vector representations for each narrative analysis while maintaining a link to the original news item. These embeddings can then be compared using popular distance metrics.

3.4. Clustering the Vector Representations

Our goal is to identify the most popular groups of narratives. Since we now have a measure of distance between objects and an understanding of the data structure, we can apply clustering methods to group similar narratives. Larger clusters will represent more popular narratives.

3.5. Summarization

In the previous step, we obtained clusters, which may contain thousands of news items. Presenting results in this form would not be practical, so we need to identify the overarching narrative within each cluster. One way to do this is by randomly selecting N news items (where N is much smaller than the cluster size) and summarizing them using an LLM. This result can be considered the

"headline" for the cluster. The headline can then be used in further results presentation and the next step, which is cluster validation.

3.6. Validation of Results

We begin the validation with the largest cluster. Using the headline, we can re-annotate the cluster to assess how well each narrative aligns with the main idea identified through the summarization of the randomly selected narratives. The LLM's response at this stage will classify each narrative as:

- “The narrative fully aligns with the cluster's headline”,
- “The narrative aligns with the cluster but indirectly”,
- “The narrative does not align with the cluster's headline”.

Based on these classifications, we can assess the quality of the clustering. We then set a threshold for the acceptable proportion of narratives that do not align with the cluster. If this proportion is low, these narratives can be marked as noise. If the proportion is too high, we must return to step 4 and rerun the clustering with different input hyperparameters or even a different algorithm.

3.7. Presentation of Results

The ultimate goal of this algorithm is to generate an analytical report that provides insights into the popularity of different narrative groups. Assuming we have access to all necessary metadata (publication dates, source names, language, audience size, etc.), we can use data visualization techniques to explore statistical indicators of narrative popularity, identify periods of narrative spikes, and generate word clouds. This gives the reader a deeper understanding of the information campaign and offers insights for further research into the causal links between the publication of the news sample and the overall propaganda narrative.

4. Research results

In this section, we will discuss the research results, covering everything from the data and model descriptions to the experimental outcomes, evaluation metrics, and identified challenges.

4.1. Dataset description

The proposed algorithm has been developed, tested, and refined using three different datasets collected for research purposes:

- The propaganda campaign “Taiwan is an inseparable part of China” in Russian media after February 2022 (2,500 analyzed news articles, from 02.2022 to 08.2023),
- The propaganda campaign “US biological laboratories in Ukraine” (95,800 analyzed news articles, from 02.2022 to 12.2023),
- The information space of the Baltic states during Russia's full-scale invasion of Ukraine (354,700 analyzed news articles from 152 channels, from 02.2022 to 04.2024).

For this research project, a subset of the dataset from the analysis of the Baltic information space during Russia's full-scale invasion of Ukraine was chosen as the demonstration dataset. This subset specifically focuses on Russian-language propaganda channels targeting Lithuania, Latvia, and Estonia. It contains 29,322 news articles published by 14 selected Telegram channels during the period from 02.2022 to 04.2024.

4.2. Machine Learning models

The proposed algorithm employs three types of machine learning models: a large language model, an embedding model, and a clustering algorithm. For the large language models, OpenAI’s GPT-4-

Turbo was chosen for news analysis, and GPT-3.5-Turbo was used for result validation. Several models were compared for the text vectorization task, including:

- OpenAI text-embedding-3-small,
- OpenAI text-embedding-ada-002,
- HuggingFace Alibaba-NLP/gte-Qwen1.5-7B-instruct,
- HuggingFace WhereIsAI/UAE-Large-V1,
- HuggingFace intfloat/multilingual-e5-base.

For clustering the embedding vectors, the following algorithms were applied:

- K-Means++ with the elbow method to determine K,
- Hierarchical Clustering,
- DBSCAN,
- HDBSCAN.

4.3. The problem of determining cluster granularity

The issue of cluster granularity arises from the need to balance between the number of clusters and the level of detail they represent. On one hand, creating a large number of small clusters can capture the unique features of individual texts. On the other hand, merging texts into larger clusters based on common characteristics risks losing important details. In the context of analyzing propaganda narratives, this dilemma becomes especially significant. Too much detail can obscure the broader picture, as propaganda narratives often have complex structures and employ a variety of tactics to achieve their goals. At the same time, excessive generalization may overlook subtle but crucial differences between narratives, which can be critical for understanding the mechanisms of propaganda.

Addressing the problem of cluster granularity requires expert intervention. An expert with deep knowledge of the subject matter can identify which text characteristics are essential for clustering and which can be disregarded. This expertise allows for the creation of clusters that best align with the research objectives.

4.4. Implementation results

Implementation results are presented in tables 1-5 and in Figures 2-3.

Table 2

Clustering metrics and results for **OpenAI text-embedding-ada-002** model

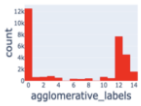
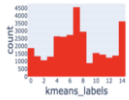
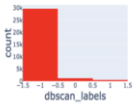
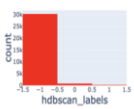
Metric / Algorithm	Number of clusters	Cluster distribution histogram	Silhouette Coefficient	Davies-Bouldin Index	Clainski-Haranasz Index
Hierarchical clustering	15		0.010	5.515	180.944
K-Means	**15**		**0.036**	**3.870**	**490.498**
DBSCAN	3		- 0.087	3.612	215.127
HDBSCAN	3		-0.093	3.486	178.963

Table 3
Clustering metrics and results for **HuggingFace Alibaba-NLP/gte-large-en-v1.5** model

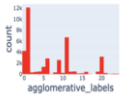
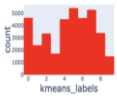
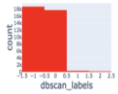
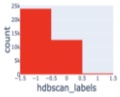
Metric / Algorithm	Number of clusters	Cluster distribution histogram	Silhouette Coefficient	Davies-Bouldin Index	Clainski-Haranasz Index
Hierarchical clustering	25		0.014	4.258	192.817
K-Means	**10**		**0.040**	**3.328**	**723.084**
DBSCAN	4		-0.011	4.493	342.468
HDBSCAN	3		-0.022	5.965	320.535

Table 4
Clustering metrics and results for **HuggingFace Alibaba-NLP/gte-large-en-v1.5** model

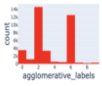
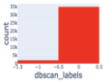
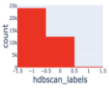
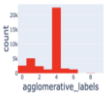
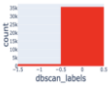
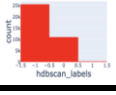
Metric / Algorithm	Number of clusters	Cluster distribution histogram	Silhouette Coefficient	Davies-Bouldin Index	Clainski-Haranasz Index
Hierarchical clustering	10		0.037	4.515	380.300
K-Means	**10**		**0.049**	**3.340**	**788.349**
DBSCAN	2		0.125	4.654	399.316
HDBSCAN	3		-0.015	5.992	327.911

Table 5
Clustering metrics and results for **HuggingFace intfloat/multilingual-e5-base** model

Metric / Algorithm	Number of clusters	Cluster distribution histogram	Silhouette Coefficient	Davies-Bouldin Index	Clainski-Haranasz Index
Hierarchical clustering	10		0.010	5.797	225.086
K-Means	**10**		**0.026**	**4.114**	**512.323**
DBSCAN	2		0.117	5.662	161.783
HDBSCAN	3		-0.037	6.462	249.931

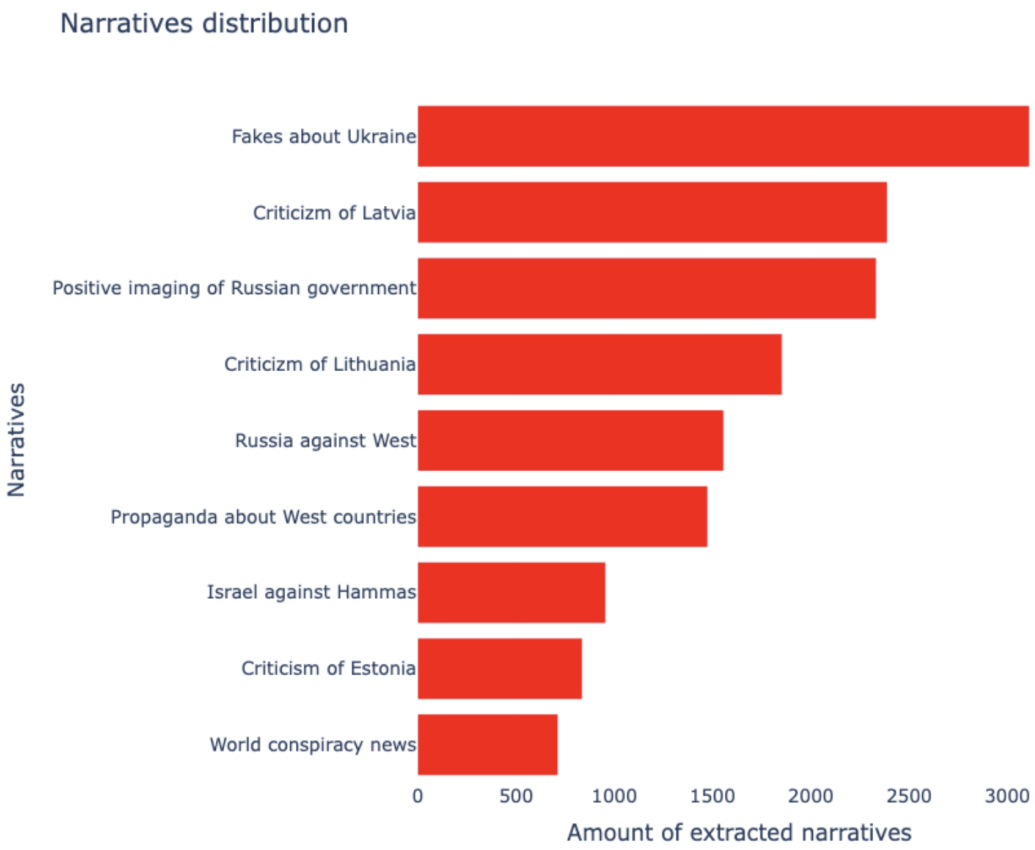


Figure 2: Extracted narratives distribution.

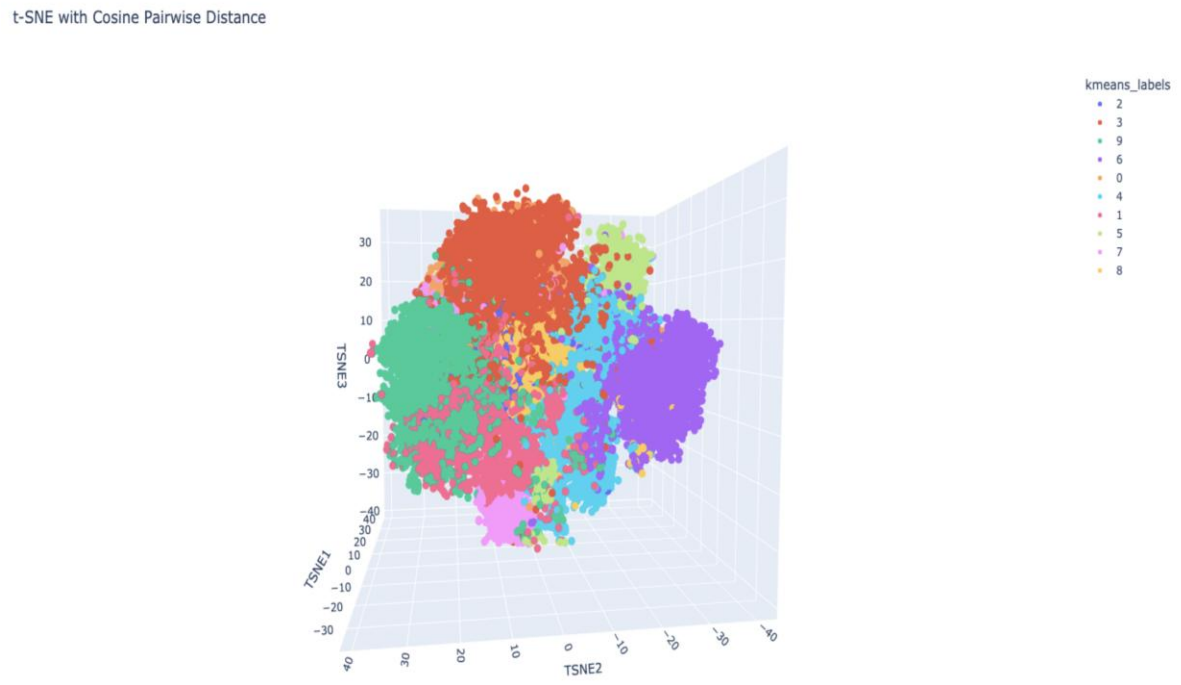


Figure 3: Visual representation of embeddings compressed to 3D space using t-SNE method.

5. Conclusions

This research presents a new method for quantitatively assessing the popularity of propaganda narratives, which enables the systematic and automated analysis of the information space. The applied natural language processing (NLP) and machine learning techniques significantly enhance the efficiency of analyzing large volumes of text data. Furthermore, the objectification of the analysis process is critically important. Human involvement can introduce subjective interpretations, bias, and errors. An algorithmic approach ensures consistency and reproducibility of results, which is essential for any analytical work.

Despite its considerable potential, using AI and machine learning methods for propaganda analysis comes with challenges. First, the availability and quality of data are crucial for the effectiveness of machine learning models. Incomplete or biased data can significantly affect the accuracy of the analysis. Another challenge is that algorithms may struggle to interpret irony, sarcasm, and cultural references, which are often used in propaganda texts. However, with the advancement of modern models, this issue is becoming less of a concern.

During the experiments, a dataset of news articles was collected and annotated, and several hypotheses and models were tested to determine the best approach for analysis. The results of the study include:

- A list of identified narratives from the dataset along with metrics of their popularity,
- Comparative tables of clustering metrics for the results of embedding models,
- Infographics illustrating the relationship between the annotated categories and the semantics of news within clusters,
- An example of an infographic for narrative representation.

Acknowledgements

We would like to extend our sincere gratitude to Mantis Analytics for providing the valuable data and sharing their expertise in propaganda analysis, which were instrumental in the success of this research.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] I. Dagil, Y. Tereshchenko. Assessing the Popularity of Propagandist Narratives Using AI Methods. Proceeding of XXII International scientific and practical conference "Shevchenko Spring - 2024" April 11, 2024, Kyiv, Ukraine. URL: https://probability.knu.ua/shv2024/ShV_2024.pdf.

[2] Piper, A., So, R. J. and Bamman, D. Narrative Theory for Computational Narrative Understanding. In M. F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing (2021). Association for Computational Linguistics. Online and Punta Cana, Dominican Republic. 2021. pp. 298-311. URL: https://doi.org/10.18653/v1/2021.emnlp-main.26.

[3] Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano and Sérgio Nunes. A survey on narrative extraction from textual data. Artificial Intelligence Review. 56 (2023) 8393–8435. doi:10.1007/s10462-022-10338-7.

[4] Burovova Kateryna, Mariana Romanyshyn. Computational Analysis of Dehumanization of Ukrainians on Russian Social Media Proceedings of LaTeCH-CLf, March 22-2024, 2024. pp. 28–39. URL: https://aclanthology.org/2024.latechclfl-1.4.pdf.

[5] Niklas Muennighoff, Nouamane Tazi, Loïc Magne and Nils Reimers, MTEB: Massive Text Embedding Benchmark Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (2023). doi:10.18653/v1/2023.eacl-main.148.

[6] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder and Furu Wei. Improving Text Embeddings with Large Language Models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2024). doi: 10.18653/v1/2024.acl-long.642.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019). doi: 10.18653/v1/N19-1423.

[8] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3982–3992, Hong Kong, China, November 3–7, 2019. URL: https://aclanthology.org/D19-1410.pdf.

[9] Ricardo J. G. B. Campello, Davoud Moulavi and Joerg Sander. Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. Conference paper on Advances in Knowledge Discovery and Data Mining (PAKDD 2013). 2013, vol 7819. doi: 10.1007/978-3-642-37456-2_14.

[10] Frank Nielsen. Introduction to HPC with MPI for data science, Springer International Publishing Switzerland, 2016. URL: https://doi.org/10.1007/978-3-319-21903-5.

[11] Martin Ester, Hans-Peter Kriegel, Jiirg Sander and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of conf. "Knowledge Discovery and Data Mining (KDD-96), 1996. pp. 226-231. https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf.

[12] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data, **10**(1), 2015. 5.1-5.51. URL: https://doi.org/10.1145/2733381.