Communication Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 147–152
DOI: 10.15439/2023F1961
ISSN 2300-5963 ACSIS, Vol. 37

# Harness Old Media: a cross-disciplinary approach to utilizing television data for media content analysis.

Piotr Jabłoński
0000-0001-9360-3502
Faculty of Mathematics and Computer Science
Adam Mickiewicz University, Poznan, Poland
e-mail: piotr.jablonski@amu.edu.pl

***Abstract*—The phenomenon of disinformation has become a common theme in studies across various fields. Both qualitative and quantitative methodologies are typically used, focusing primarily on content sourced from the internet. This article introduces a method to extend this focus to include content from 'Old Media'[1] specifically from Television which as an unstructured medium, presents a combination of textual and visual layers. Despite this complexity, the integration of these elements allows for the design of algorithms capable of analyzing video streams and extracting individual news from main news programs of nationwide broadcasters. The proposed solution facilitates the extraction of transcriptions generated by the research tool. The aim of this research is to allow access to the content of television to enable its inclusion in research, performed in a manner analogous to Internet content. This research is part of a project that deals with the development of algorithms for combining, classifying and comparing content from different media in order to design an imprecise classifier of disinformation content.**

## I. Introduction

THE spread of the term 'Fake News' worldwide is frequently associated with pivotal political events of 2016, including the US presidential election and the Brexit referendum in the UK, during which the internet and social media were flooded with fabricated content. As outlined in [2], the concept of fake news itself emerged much earlier, dating back to the 19th century in tandem with the rapid development of yellow journalism in the United States. The phenomenon of disinformation, has also reached Poland, causing the emergence of numerous educational campaigns targeting a broad audience in Poland. Among many different initiatives, there was a high rise of activities included the core curriculum of primary and secondary schools, implemented through educational projects. A variety of outreach programs geared toward the general public have also been inaugurated, coordinated by a group of NGO institutions, media consortia and governmental agencies. Post-2016, the phenomenon of Fake News has been a focal point in a multitude of scientific research conducted by different institutions not only in Poland, but also worldwide. These studies span across numerous disciplines, including but not limited to social communication and media studies, linguistics and computer science. It is noteworthy to mention that this phenomenon is also incorporated into the specific objectives of the Infostrateg program, launched by the National Center for Research and Development in 2020 [3].

[1]The terms "Old Media" and "New Media" began to be used by scholars and academics studying the changes in communication caused by the growth of digital technologies in the 1990s. New media includes forms of communication in online form such as electronic books, email, informational web portals, and social media [1].

## II. Current state of research

So far, the research conducted by scientists is mostly focused on the analysis of one medium - print press / television / Internet. In current research, one can notice a tendency to choose the latter medium more often, while as Naturel stated: "Querying and retrieving information from a large television (or video) corpus is still a challenge, for both professional archivers and simple TV users as well." [4]. That's why the justifications for the choice of internet medium points to its universality and great dynamism, unavailable in the old media. Vasoughi also pointed out that on the Internet, news classified as falsehood, was able to reach first 1500 people six times faster than the true news [5].

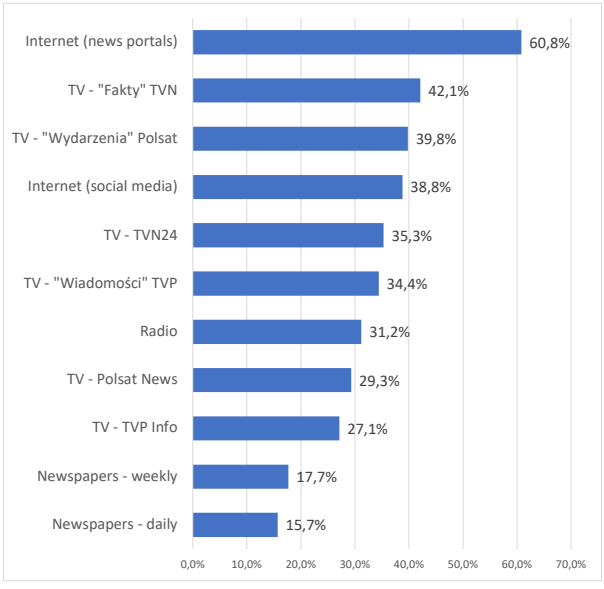


Figure 1. Results of the survey about the preferences of Poles regarding sources of information about Poland and the world. Source: [6]

Undoubtedly, studying the content of a single medium has a major impact on the consistency of research methodology. The



**Thematic track:** International Workshop on AI in Digital Humanities, Computational Social Sciences and Economics Research

Internet as a medium, as mentioned earlier, not only enables a high rate of news dissemination, but is also a frequent source of information for the Polish public. As shown in the "Survey of Poles' preferences for sources of information about Poland and the world" [6] presented in the Figure 1, 60.8 percent of Poles cling to information from online portals. However, this group does not include social media, which is the main choice of information source for 38.8% of respondents. In the same survey, TV programs such as Fakty TVN, Wydarzenia Polsat, and Wiadomości TVP received 42.1%, 39.8%, and 34.4% respectively.

In the same survey, it was also shown that social media is a more frequent choice as a source of information than online portals only in the 18-29 age group, although their advantage is quite small (75.8% to 71.8%). Taking into account the above statistics, as well as a report showing that 93.3% of households in 2022 had access to the Internet [7], television, as a news medium, should be analyzed just as often as internet in media content studies. As stated in [8] "Even in many developed and technologically advanced countries (...), among the people who say they use the Internet daily, a large percentage also say they use television daily for information purposes".

Despite the majority of studies based on material from the Internet, there are studies conducted simultaneously analyzing the content of different media. One of the largest studies conducted to date is the work of Claudia Melladio's international team "Journalistic Role Performance - second wave". This team consists of researchers from 37 countries, studies the content of television, radio, press and Internet portals. The study of Polish team, taking part in this research included 14 outlets. In this group three of them were television programs broadcasted by main nationwide television stations. Uniformity of sampling in this international study consisted of examining two constructed weeks from the entire year 2020. As a result, 541 news cases from Polish news programs broadcast on television were analyzed. This accounted for 8.6% of all news stories analyzed in Poland in the aforementioned study. One of the reasons for conducting research on such a limited research sample is the extremely time-consuming process of coding the video, which involves reviewing and analyzing the content in each message according to a designed codebook.

From the above analysis arose the need to enable access to TV content in a manner equivalent to Internet content - in text form. This form of data makes it possible to use the tools and statistical methods available in natural language processing techniques.

An analysis of the scientific literature has revealed a prevailing shortage in the area of both the discipline of social communication and media sciences and computer science. The research conducted in the field of social sciences on media content analysis using quantitative data analysis methods implemented on a sample of just over half a thousand records is the largest studies conducted in the world to date. The data analysis methods used in the referenced study do not involve any statistical language analysis processes and rely entirely on human input. In contrast, in the discipline of computer science, where such methods would be expected, research is not conducted at all, due to the lack of access to research material. The research carried out by the author, instead, focuses on creating the development of methods for automatic verification and analysis of media content.

## III. Context of the study

The purpose of the project is to develop algorithms to fuse, classify and compare content from different news media for the purpose of designing an imprecise classification of disinformation content. The problems of verifying the authenticity and reliability of published information are a direct result of the oversupply of content and information noise, determined by intertwining elements such as truth, rationality, objectivity, but also rumor, hearsay or conspiracy theories. For this reason, not only the recipients, but also the editors who are the intermediary of the media message have a problem with their verification, and instead of stopping further publications, they amplify the process of spreading unverified information in the media. The oversupply of information is noticeable in social media, where in 2020 Twitter published about 380,000 twitts a day in Polish only, which is more than 260 new messages per minute[2].

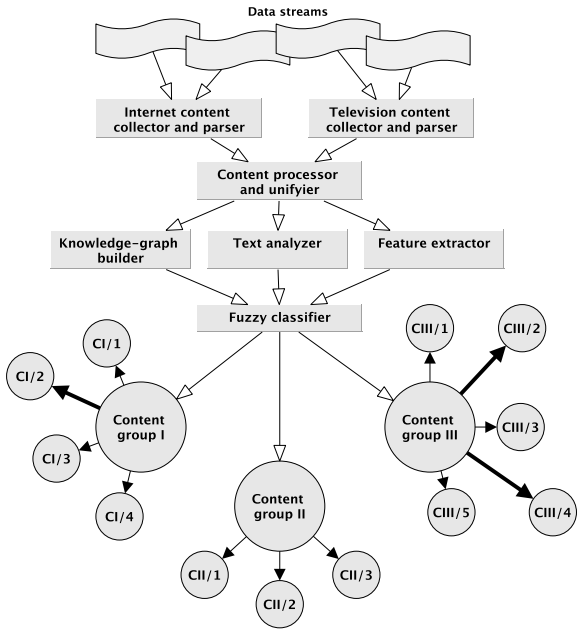


Figure 2. Diagram of the disinformation detection system.

At present, there is a lack of solutions capable of verifying in an automated manner, on the basis of television broadcasts,

[2]The data comes from research conducted by the author in 2021/2022. The research involved an analysis of the volume of information published on web news portals and twitter platform. Their effect was used in the work of the team implementing the study "From urban legend to fake news. A global detector of contemporary falsehood" funded by NCBiR.

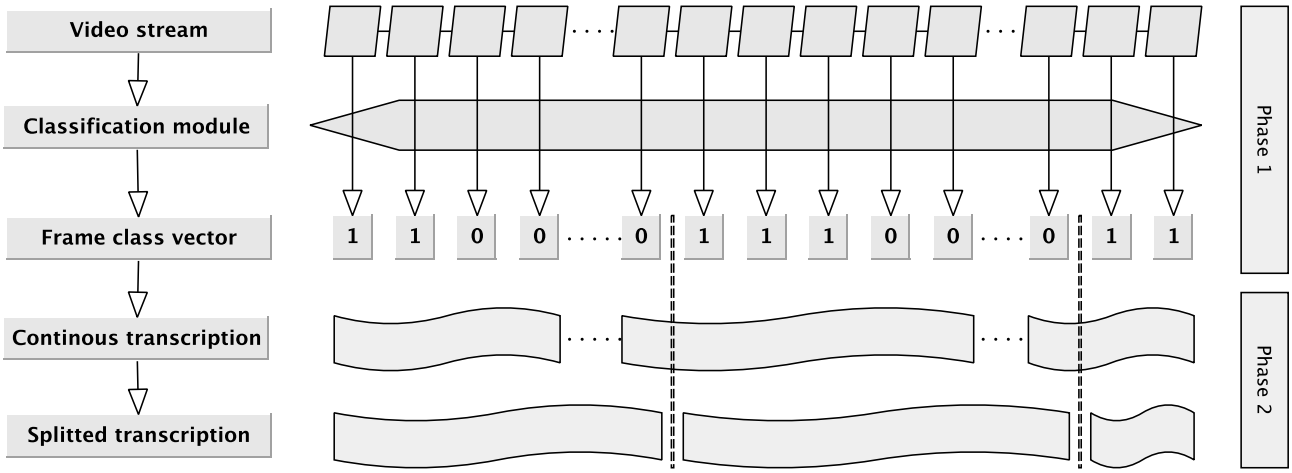


Figure 3. Scheme of the designed system splitting news outlet to separate news.

whether a specific content of a message originating from the Internet media has the hallmarks of disinformation content. Developing a solution to the above problem, will help create a mechanism for detection of disinformation content, which can stop the spread of disinformation (fake news) by professional media broadcasters and Internet users who are unaware of the threat posed by disinformation. The research presented in this article is part of a system presented at Figure 2, which in its next steps uses knowledge-graphs and fuzzy logic to fuse content from different media.

## IV. Developed algorithm

Such a large amount of information delivered in a limited amount of time can affect the creation of information chaos that media audiences have to deal with. In this worldwide confusion, the spread of false information (so-called "fake news") is becoming more frequent, not only on social media, but also by professional news outlets.

In order to collect the research material in the form of transcript text files, an algorithm was designed for news programs broadcasted by major national TV stations (Polsat, TVN, TVP). Developed algorithm detects individual news stories in an outlet and then extracts their transcript, covering detected news which can be later treated as single text documents. The database in the form of video footage and transcription file, for current and archived program outlets, is the CAST (Content Analysis System for Television) system, operating at the Faculty of Political Science and Journalism at Adam Mickiewicz University in Poznan [9] [10]. *The system records 6 channels: TVP 1, TVP 2, Polsat, TVN, TVN24, TVP Info continuously, (24 hours a day), since mid-2014. (...) The broadcasts are stored in a database, described with metadata generated from the EPG (...) Another useful feature present in the system is speech-to-text transcription. Each Polish-language broadcast found in the database contains a text transcription of all the issues spoken in it."* [11].

Thanks to the data collected by the CAST system, it became possible to design a two-phase algorithm. The first one analyzes the video stream, assigning to each frame one of two classes - a frame with a studio image (1), and a frame with a non-studio image (0). Each of broadcasted program have a set of dedicated model and mask in size of frames to determine its class. For this purpose, the OpenCV library is used, with the 'matchTemplate' method additionally using 'masks' that exclude variable parts of the studio image. For each program, minimum of three key frames are defined with a mask applied to it, designed to make the analysis independent of variable elements of the scene. This is an extremely important part of the overall analysis, as current studio arrangements provide not only for multiple presenters, but also for variable camera settings and dynamic backgrounds, often occupying as much as 65,5% percent of the scene area. Figure (4) presents an example set of templates and masks for the TVN Fakty program. The process designed in this phase accepts a video stream as input, which is then analyzed. The video can be of any size but must be encoded with a codec that can be parsed by the OpenCV library. During testing different size of video source[3] was used but the algorithm showed an insensitivity to the resolution of the video analyzed. Currently, it covers a wide range of codecs like MPEG1/2/4, H.264, HEVC, VP8/9, VC1, but also JPEG and uncompressed video. At the beginning of the process, algorithm detects the number of frames per second in the input stream, which is then set as the fixed number of frames to be skipped between successive analyzed key-frames. In the proposed solution, the number of frames to be skipped is equal to the number of frames per second, resulting in the analysis of exactly one frame every one second. The implemented approach optimizes the performance of the image analysis process, which was confirmed empirically

[3]During test following resolutions was checked 1080x1920px, 1280x720px and 640x360px. None of these resolutions have shown greater effectiveness of identifying studio scenes.

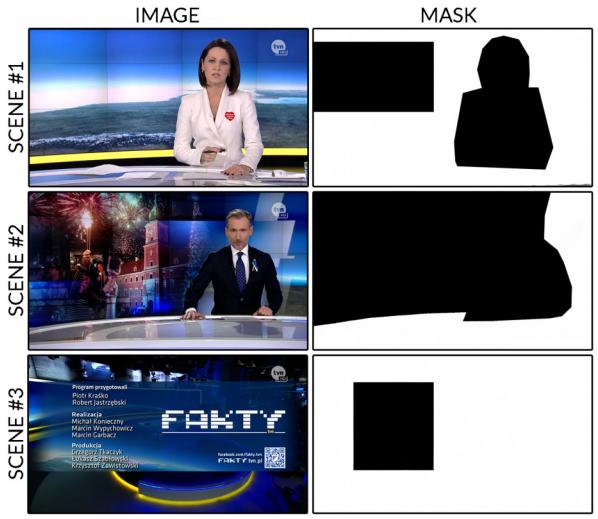


Figure 4. Figure shows examples of different templates and corresponding masks used in algorithm to detect studio frame between news.

during the study. Decreasing the step, did not result in an increase in efficiency in detecting the class of the frame, but increased the number of frame comparison operations. In the CAST system, video is recorded at 50 frames per second. This means that a video stream, lasting 30 minutes, contains 90,000 video frames. As a result of optimization involving frame bypassing, only 1,800 frames are analyzed, which is 2% of all video frames. If a higher step value (equal to two, three or five times the number of frames per second) was adopted, a noticeable delayed detection of the studio's frame was created. In effect the beginning of presenter speech could be cut off from the next material. Then each key-frame is analyzed using the OpenCV library, which determines the class of the frame (studio/non-studio). Comparison of frame with template and mask is made using the function `matchTemplate()` (`cv.matchTemplate(image, templ, method[, result[, mask]])`, where `image` is the analyzed frame, `templ` is the prepared template file and `mask` is the matching mask file. Currently OpenCV uses one of two methods, which support mask usage: `TM_SQDIFF` and `TM_CCORR_NORMED`. In this research was used the last one, which stands for *Correlation Coefficient*. Function this return a matrix of values, which is then search as global maximum with usage of `minMaxLoc()` function. As the effect of first phase of the algorithm, a vector of binary values is generated, containing a values representing the classification of each analyzed key-frame.

In phase two of the algorithm, the previously generated vector is converted into time code values according to the parameters of the analyzed stream. This vector is then used to indicate the locations of studio and news boundaries in the transcription file into parts corresponding to the detected image sequences. Each of isolated transcription part is then supplemented with a metadata metric in the form of parameters

Table I
METADATA AVAILABLE ON ARCHIVED OUTLETS OF NEWS WITH EXAMPLE CONTENT.

| Field | Description | Content |
|---|---|---|
| UUID | Unique object identifier in the CAST system | b61ca13f-b0f2-47f0-ad49-ecedf6107de1 |
| Title | EPG program name | Wiadomości |
| Description | EPG description | News service presenting (...) economy, culture and social life. |
| Channel | Channel | TVP 1 |
| Category | EPG Category | news/current affairs (general) |
| Start date | Scheduled recording start date | 2022-03-30 |
| Start time | Scheduled recording start time | 19:30:00 |
| End date | Scheduled recording end date | 2022-03-30 |
| End time | Scheduled recording end time | 20:05:00 |

presented in Table I for easy transcription file identification. The gathering and saving of metadata on each of the broadcast programs is a key to the ability of determining the publish location and exact time of the video's broadcast. The sole recording of a TV program or analysis of available video footage (especially in the context of archival material, available on the Internet) does not allow for precise temporal placement, which can be crucial for identifying the source of the disinformation. With precise metadata, including the name of the program, it becomes possible not only to develop the route of the spread of content in the media, but also the appearance of actors in a specific time frames.

## V. CONCLUSION

As a result of the research, a set of files was created, representing the content of the main news channels of Polish national TV stations. The result of the program for a period of 26 weeks, is depicted by 6989 text files, representing the extracted news. Each of program outlet consist of average of 12.78 news items per program.

Algorithm, tested on small human annotated test set, consisting of 30 episodes, presents an outstanding performance of scene identification. Table II presents exact results, in which it is shown performance of detecting studio scenes in every news program. The developed algorithm achieves the following results: Precision 97,95%, Recall 97,46% and F1-score: 97,70%.

The data set built with the presented algorithm will contribute positively to the ability to analyze the content of TV

Table II
THE EFFICIENCY OF DETECTING STUDIO SCENES IN THE TEST SET

| Detecting method | Total | TVP | Polsat | TVN |
|---|---|---|---|---|
| Human annotated programs | 30 | 10 | 10 | 10 |
| Human decision | 383 | 124 | 132 | 127 |
| Scene correctly detected by Algorithm (TP) | 366 | 117 | 128 | 121 |
| Scene not detected by Algorithm (FN) | 10 | 4 | 2 | 4 |
| Scene incorrectly detected by Algorithm (FP) | 8 | 2 | 2 | 4 |

programs. Undoubtedly, this is an important step in making the content of this medium available for quantitative research, particularly when comparing it with Internet content.

Each document retains the appropriate structure and detailed metadata, enabling extensive quantitative content analysis. It may be possible to apply methods of automatic text summarization [12] and NLP Tasks with the use of Transformer Models [13]. It can be also used to analyze the affect [14] in news content. The data can also be used to more accurately analyze the content of the messages according to 5W Lasswell's model of communication (who?", "says what?", "in what channel?", "to whom?", "with what effect?") [15] over a broader timeframe. Proposed content distribution can also be another step in the growth of data journalism where big data plays an important role [16].

In the near future, the development of the designed algorithm should also include the possibility of identifying experts and speakers in the broadcast. Currently, the CAST system implements a module for reading the content of information contained in lower third[4], which is in the fine-tuning stage. Thanks to this functionality, another layer of information is being added, which can help in the detection of actors appearing in the media message. All unstructured data collected from television and structured data from the Internet should be organized into a Knowledge Graph, enabling the creation of an efficient connection among all instances of knowledge. According to Zhang[17], it is valuable to employ embedding and clustering algorithms to implement a topic hierarchy for enhancing the Knowledge Graph's performance.

Effects of this research will also positively contribute to the author's project covering methods of imprecise classification of disinformation content. The work is intended to test the possibility of content analysis on the basis of media content from both the Internet and the television broadcasts. Another task is to measure the effectiveness of content clustering and classification using fuzzy logic methods. One of the elements developed in the research will be a system for analyzing various media messages on a selected topic, taking into account similarities between messages. These similarities will be developed, among other things, on the basis of the results of summarizing modules, sentiment analysis and the combination of identified named entities.

[4]Lower third is a graphical overlay, placed in lower part of screen, containing information about current story or appearing person, like name, surname and affiliation.

## REFERENCES


[1] W. J. Dizard, *Old Media New Media: Mass Communications in the Information Age*, Second edition. New York: Longman, 1996, ISBN: 9780801317439.

[2] D. Halagiera, "Fake news jako nowe (stare) wyzwanie dla świata mediów – portal YouTube w walce z nieprawdziwymi informacjami," in *Kryzysy współczesnego świata. Różne ujęcia problemów globalnych i regionalnych*, 2019, pp. 91–105.

[3] Narodowe Centrum Badań i Rozwoju, "Program Strategiczny INFOSTRATEG „Zaawansowane technologie informacyjne, telekomunikacyjne i mechatroniczne"," Narodowe Centrum Badań i Rozwoju, Warszawa, Tech. Rep., Apr. 2020. [Online]. Available: https://archiwum.ncbr.gov.pl/fileadmin/Programy_Strategiczne/Opis_Programu_INFOSTRATEG.pdf.

[4] X. Naturel and P. Gros, "Detecting repeats for video structuring," *Multimedia Tools and Applications*, vol. 38, no. 2, 2008, ISSN: 13807501. DOI: 10.1007/s11042-007-0180-1.

[5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, 2018, ISSN: 10959203. DOI: 10.1126/science.aap9559.

[6] Instytut Badań Internetu i Mediów Społecznościowych and Instytut Badań Rynkowych i Społecznych. "Badanie preferencji Polaków dot. źródeł informacji o Polsce i świecie." (Jan. 2021), [Online]. Available: https://ibims.pl/skad-polacy-czerpia-informacje-o-polsce-i-swiecie-raport-ibims-i-ibris/.

[7] Eurostat, *Digital economy and society statistics*, Dec. 2022. [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/isoc_ci_in_h/default/table?lang=en.

[8] S. J. Shaikh, "Television versus the internet for information seeking: Lessons from global survey research," *International Journal of Communication*, vol. 11, 2017, ISSN: 19328036.

[9] Faculty of Political Science na Journalist. "CAST." (2020), [Online]. Available: https://wnpid.amu.edu.pl/en/home/cast.

[10] J. Wyszyński, *Content Analysis System for Television*, 2017. [Online]. Available: http://cast.info.pl/.

[11] A. Stępińska and J. Wyszyński, "Ilościowa analiza zawartości przekazów w badaniach nad dyskursem populistycznym," in *Badania nad dyskursem populistycznym: wybrane podejścia*, 2020, ch. VII, pp. 107–129.

[12] N. Andhale and L. A. Bewoor, "An overview of text summarization techniques," *Proceedings - 2nd International Conference on Computing, Communication,*

*Control and Automation, ICCUBEA 2016*, 2017. DOI: 10.1109/ICCUBEA.2016.7860024.

[13] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for nlp tasks," *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*, 2020. DOI: 10.15439/2020F20.

[14] P. Subasic and A. Huettner, "Affect analysis of text using fuzzy semantic typing," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 4, pp. 483–496, Aug. 2001, ISSN: 10636706. DOI: 10.1109/91.940962.

[15] H. D. Lasswell, "The structure and function of communication in society," *The Communication of Ideas*, no. 1948, 1948.

[16] A. Veglis and T. A. Maniou, "The mediated data model of communication flow: Big data and data journalism," *KOME*, vol. 6, no. 2, 2018, ISSN: 20637330. DOI: 10.17646/KOME.2018.23.

[17] Y. Zhang, M. Pietrasik, W. Xu, and M. Reformat, "Hierarchical topic modelling for knowledge graphs," pp. 270–286, 2022. DOI: 10.1007/978-3-031-06981-9_16.