2020, Vol. 21, No. 1 (80)
ISSN 2451-1617

Identification of Trends in the Polish Media on the Example of the Quarterly *Studia Medioznawcze*

The Use of Big Data Tools

Piotr Pruchnik

Uniwersytet Warszawski

p.pruchnik@uw.edu.pl

ORCID: 0000-0001-6923-1673

ABSTRACT

The media market is developing dynamically; therefore, it is important to forecast trends correctly. **Scientific objective:** The aim of the paper is to identify trends in the Polish media. **Research methods:** The paper presents a unique methodology of information extracting—Big Data for analyzing trends in the media. The source material were the texts published in the quarterly Studia Medioznawcze. The Big Data tools have been used for the analysis. **Results and conclusions:** Author's forecast of media trends based on scientific publications have been formulated, and it was compared with the one prepared by the consulting company PwC. **Cognitive value:** The results show significant discrepancies. The most promising areas in the PwC Report Forecasts VR and OTT have no confirmation in scientific papers.

KEYWORDS

Trend, Big Data, data mining, PwC, media

Bearing in mind the dynamic changes in the image of the Polish media, this study attempts to identify trends for the coming years. Statistical methods of text analysis as the research tool have been used. Source data are publications contained in the quarterly *Studia Medioznawcze.* A comparative analysis of identified trends have also been performed with forecasts published by companies specializing in the media market.

The National Broadcasting Council (KRRiT) attempted at forecasting Polish media and determining development directions in 2015. The document "Strategy for the Development of the Media Market in Poland for 2015-2020" has been prepared (Sztuka, 2015). A group of authorities and experts participated in its creation in cooperation with institutions supporting the media market in Poland: KRRiT, PISF, and ZAIKS. In addition to the analysis of the environment, the document contains forecasts and demands for media development. The document is slowly expiring. For the time being, no summary of this document and / or a similar forecast for the coming years have been published. An attempt to participate in the discussion on the subject of "Strategy" was made at the Department of Media Information Technology of the Faculty of Journalism, Information and Book Studies of the University of Warsaw to identify and evaluate trends in changes in modern media. They are a derivative of interest in the dynamics of changes in the image of Polish media. A unique method of extracting information have been used for the analysis. Modern statistical and analytical tools and the Big Data extracting potential have been used in it (Mayer-Schonberger & Cukier, 2017). The use of statistical methods for text resource exploration is currently a common research method (Surma, 2019). The paper presents key elements of data mining (Silge & Robinson, 2017) for extracting large information resources that can be used to identify trends in the Polish media.

Media Development Forecasts

The media market has a value of many billions. The consulting company PricewaterhouseCoopers (PwC) estimates that the global media and entertainment market will reach USD 2.4 trillion in 2022 and will grow by 4.4% on average annually (PwC, 2018). Accordingly, the market in Poland will reach USD 13.4 billion and will grow 3.5% annually. PwC also indicates the most prospective development directions. In the first place are VR (virtual reality) and OTT (over the top). Media such as newspapers and magazines will see a decline. Television and video will grow slightly.

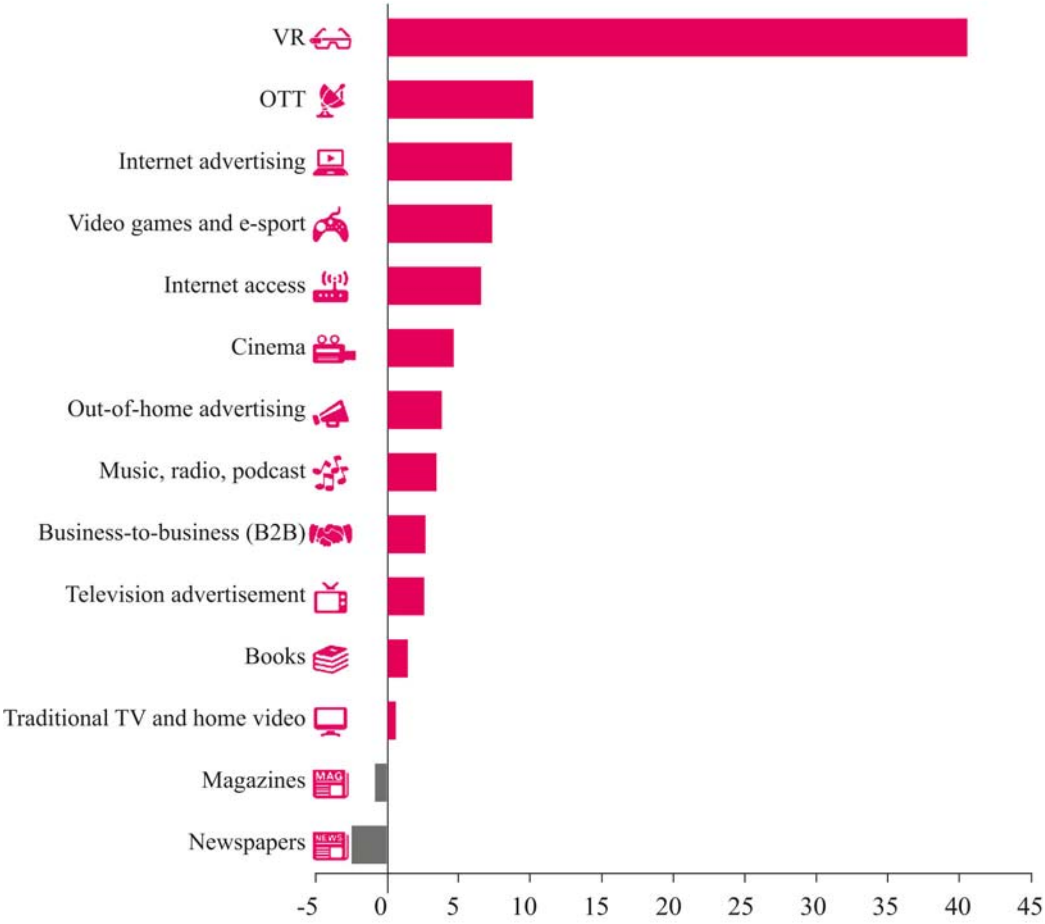

Figure 1. Forecast of media development until 2020 according to PwC

Source: PwC Report "Prospects for the Development of the Entertainment and Media Industry in Poland 2018-2022"

Another company report (PwC, 2017) concerns the forecast of the US media market, which is one of the largest in the world. Growth is projected in the virtual reality segment and E-Sports. Next is Internet advertising and Internet video. Declines are forecast in traditional Newspapers, traditional TV, and home video. One can notice the compliance of the dynamics of media development in the world and in the USA.

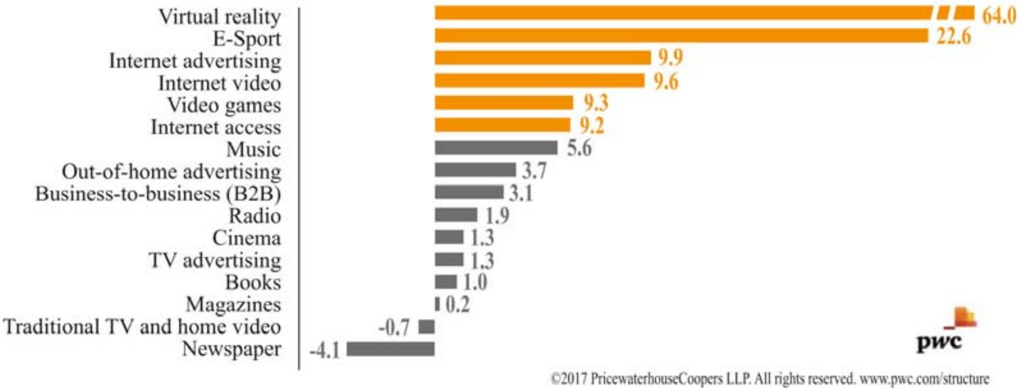

Figure 2. US media development forecast for 2016–2021

Source: PwC Report "Investment Forecast on the Media Market until 2021"

The PwC report did not mention the methods that have been used to prepare the report. It should be guessed that it is based on the knowledge of experts employed in the company, because the document refers to their opinions and statistical data on the media sector (access to them is fragmentary, and due to the lack of transparent classification, they are incomplete).

The British company Kantar Millward Brown, specializing in research, has published a report on the transformation forecast in the media for 2019 (Kantar, 2019). According to it, the biggest changes will be related to artificial intelligence, voice technologies, and advanced analytics. The company does not state what research methods have been used. In the full report, expert opinions are found, so this time it can be assumed that the company relies on their judgments.

Published reports on media development do not include the research methodology. Only the reputation of the company issuing the report is to guarantee reliability of results. An opinion based on such premises may be unreliable, as, for example, the Lehman Brothers Holdings Inc. banking sector has learned painfully Bearing in mind the reliability of opinion about the media, it is reasonable that it be formulated on the basis of systematized, explicit, and clear methods of analysis.

Media Barometer—*Studia Medioznawcze*

It was assumed that the quarterly *Studia Medioznawcze*, given its scientific authority and many years of tradition, is a reliable knowledge base in Poland sufficient to outline the state of media research in Poland. Papers published in the journal contain a wide spectrum of media research areas. Among the journals related to the media, the quarterly was in 30th place (12 points in 2016) in the world rankings (Table below). In the category of international journals it is a distant position, but among the Polish—the highest.

Table 1. Selected media journals, including their scoring in 2016

No.	Journal	Scoring in 2016
1.	New Media & Society	50
2.	Journal of Computer-Mediated Communication	45
3.	Ieee Multimedia	35
4.	Learning Media and Technology	30
5.	Media Culture & Society	30
6.	Journal of Broadcasting & Electronic Media	30
7.	Media Culture & Society	30
8.	Multimedia Tools and Applications	25
9.	Critical Arts-South-North Cultural and Media Studies	25
10.	EMI Educational Media International (Formerly: Educational Media International)	25
11.	Feminist Media Studies	25
12.	Multimedia Systems	25
13.	Networks and Heterogeneous Media	20
14.	Continuum: Journal of Media & Cultural Studies	20
15.	Networks and Heterogeneous Media	20
16.	Television & New Media	20
17.	Convergence: The International Journal of Research into New Media Technologies	20
18.	Crime Media Culture	20
19.	Discourse Context & Media	20
20.	Journal of Mass Media Ethics	20
21.	Journal of Media Economics	20
22.	Media History	20
23.	Television & New Media	20
24.	Waves in Random and Complex Media	20
25.	Critical Studies in Media Communication	15
26.	Journal of African Media Studies	15
27.	Journal of Porous Media	15
28.	New Review of Hypermedia and Multimedia	15
29.	Bioremediation Journal	15
30.	Media International Australia	15
31.	**STUDIA MEDIOZNAWCZE (*Media Studies*)**	**12**
32.	Mediaevistik	10
33.	AS Mediatijdschrift (p)	10
34.	AS Mediatijdschrift (e)	10
35.	Classica et Mediaevalia	10
36.	Intermédialités	10
37.	International Journal of Instructional Media	10
38.	Journal of Educational Multimedia and Hypermedia	10

39.	Mediaevistik	10
40.	Patristica et Mediaevalia	10
41.	Autobiografia. Literatura. Kultura. Media	8
42.	Kultura–Media–Teologia (o)	8
43.	Media–Kultura–Komunikacja Społeczna	8
44.	Media i Społeczeństwo. Medioznawstwo, komunikologia, semiologia, socjologia mediów	7
45.	Biuletyn Edukacji Medialnej	6
46.	CyberEmpathy: Magazine of Visual Communication and New Media in Art Science Humanities Design and Technology	6
47.	Zarządzanie mediami (p)	6
48.	Zarządzanie mediami (e)	6
49.	Dziennikarstwo i Media	5
50.	Nowe Media	3
51.	Nowe Media	3
52.	Media–Kultura–Społeczeństwo	3
53.	International Journal of Multimedia	-
54.	Kognitywistyka i Media w Edukacji	

Explanation of symbols used in Table 1:

(p) - paper version; (e) - electronic version; (o) - online version

Source: Retrieved on August 18, 2018, from https://punktacjaczasopism.pl

Studia Medioznawcze can certainly be considered opinion-forming, highly credible in the Polish scientific environment. The publications published in it are representative of the academic and scientific communities associated with the media. This study focuses on papers published in the journal. It should be noted that the same research methods could also be applied to other source collections of texts, e.g. discussion groups or publications on websites.

The Use of Big Data in the Trend Analysis

The study included papers published in *Studia Medioznawcze* from January 2014 to December 2018 (numbers 56–75). Papers in Polish have been selected. Other languages have been omitted. Some papers have Polish and English versions. In this case, only the Polish version have been selected.

Downloading Data

The publications were downloaded from the archives of the quarterly, from the old version of the website (http://studiamedioznawcze.pl/spis.php). The new version of the journal's website does not contain the full history of issues. It was assumed that the analysis at the level of a single paper is sufficient to draw conclusions. Manual download of individual papers from the *Studia Medioznawcze* website is cumbersome due to the large number of documents. Data mining tools included in the R environment have been used for this purpose (Team, 2019). All documents were downloaded automatically. In total, 354 papers in Polish have been collected.

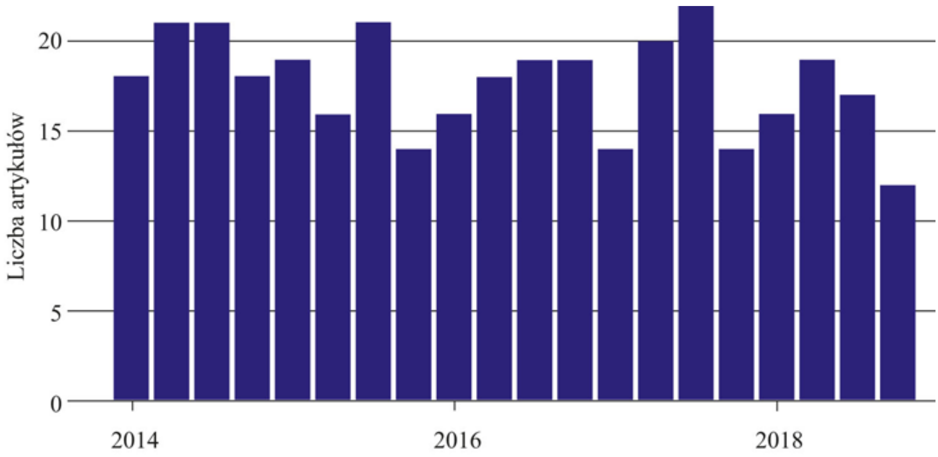

Figure 3. Time series of the number of papers in *Studia Medioznawcze* by years and quarters

Source: Own study

File Conversion

Archive documents on the journal's website are in the Portable Document Format (pdf extension). In order to analyze the data, it was necessary to convert to a text format. The composition of the publication includes two columns. A professional program that recognizes the formatting of the page have been used to convert the files.

Data Structure

When loading the files from the file name, the paper attributes have been extracted:

- The quarter in which the paper was published
- Issue number
- Author of the paper—if there are several authors, only one of them was identified. For this reason, there is no analysis of publications by their authors
- The name of the file in which the paper have been saved on the journal's page.

Tokenization

Text tokenization involves splitting the text into individual words. As a result of this process, each word is extracted from the text and is an independent element (token). Each text is segmented into a set of individual words. In the process of text analysis, statistics were calculated based on the number of tokens. To this end, a table is created that represents the input documents. The rows contain documents and the columns contain words. The table cells contain the multiplicity of words for each publication and word. The table contains many empty cells, i.e. many words appear sporadically in paper. We call this table sparse matrix.

Data Cleansing

An important part of text analysis is the preparation of data for processing by artificial intelligence algorithms. As part of the preparation, text data cleaning is performed (Gogołek & Jaruga, 2016). It consists of five steps:

1. Removing links—links to websites.

2. Removal of words not relevant to the content of the publication (so-called stopwords), conjunctions, pronouns, etc.
3. Delete punctuation.
4. Delete numbers.
5. Convert all words to lowercase.
6. Convert words to basic form. In this way, one avoid different word variations, e.g. the word "Polish" is changed to "Poland." Only basic forms in the quantitative analysis have been used. Words that are unknown to the dictionary, e.g. proper names, remain unchanged. For lemmatization, the dictionary "hunspell" (Ooms, 2018) was widely used in many computer packages, such as LibreOffice, OpenOffice, Mozilla Firefox, Google Chrome, Mac OS-X, InDesign, Opera, and many others.

Quantitative Analysis

The quantitative analysis of the text is to search for the most common words, called attributes. For this purpose, the multiplicity of all words was counted. Sorting results by the highest number, the attribute ranking list was obtained. The list of attributes indicated what words the authors use most often.

Figure 4 contains the most common words. The attribute's font size is proportional to the number of times it occurs. The larger font = word, the more frequent it was. The colors indicate the ranges of the multiplicity of attributes. Color means the same level of word frequency.

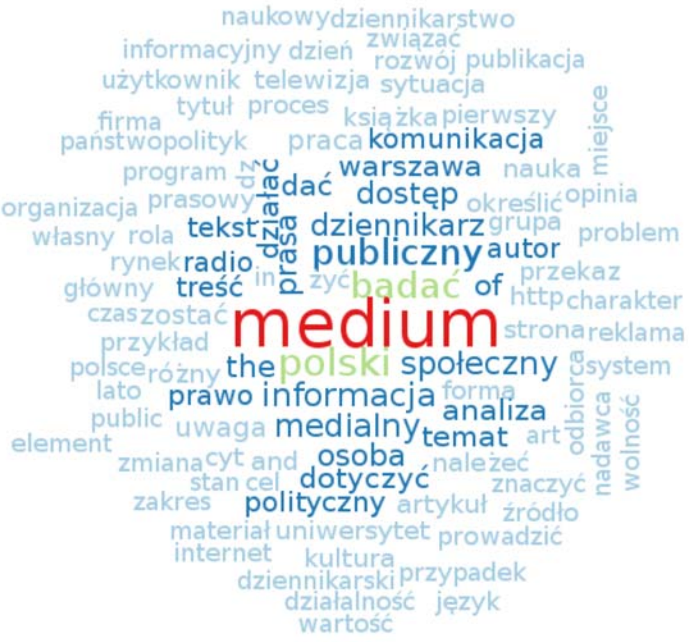

Figure 4. Tag cloud for attributes with the highest multiplicity

Source: Own study

In Figure 4, it can be seen that the authors most often use the word "medium," which is the basic form for the word "media" and all its variations. This is understandable due to the theme of the quarterly. The next group are topics = "research" and "Polish" attributes, which also fall under the profile of *Studia Medioznawcze*, the journal brings together people associated with the academic, science, and research communities. Most of papers are related to local issues, hence the attribute "Polish". In the chart / graph, one can see other attributes indicating the leading topics—"public," "information," "media," "social," "press," "journalist," and "access." They all characterize *Studia Medioznawcze* and point to the journal's profile.

Trends in the Media

With the research tools presented and source materials obtained from *Studia Medioznawcze,* it is possible to examine the changes in the popularity of the most common topics in the media in the last five years. On this basis, trends in Polish media were determined. An increase in the frequency of a topic indicates an increase in popularity, and vice versa—decreasing frequencies means decreasing interest. As before, the topic of publication will be identified with the word-attribute. This section focuses on assessing the intensity of the topic in the publication or its absence, without looking at how often it occurred. It was checked if many authors discussed the topic. In an ideal situation, each author refers in his/her paper at least once to a given issue. The dynamics in time was examined in relation to the issue number. On this basis, the relative incidence of the topic in *Studia Medioznawcze* was calculated. An increase in frequency over time means an increase in interest in the topic, a growing trend. A decrease in frequency means a downward trend. Linear regression and the slope of the regression curve have been to calculate the trend. The regression model concerns the number of words, therefore, instead of a simple linear regression; a generalized linear model with a single independent variable with a logistic binding function have been used. Attributes with the lowest absolute number have been filtered from the data set. In this way, topics that have occurred sporadically do not affect the conclusions.

Upward Trends in the Media

In the first place, the subject of the study were words-attributes, which are characterized by the largest increase in the trend. The frequency of these words increases over time.

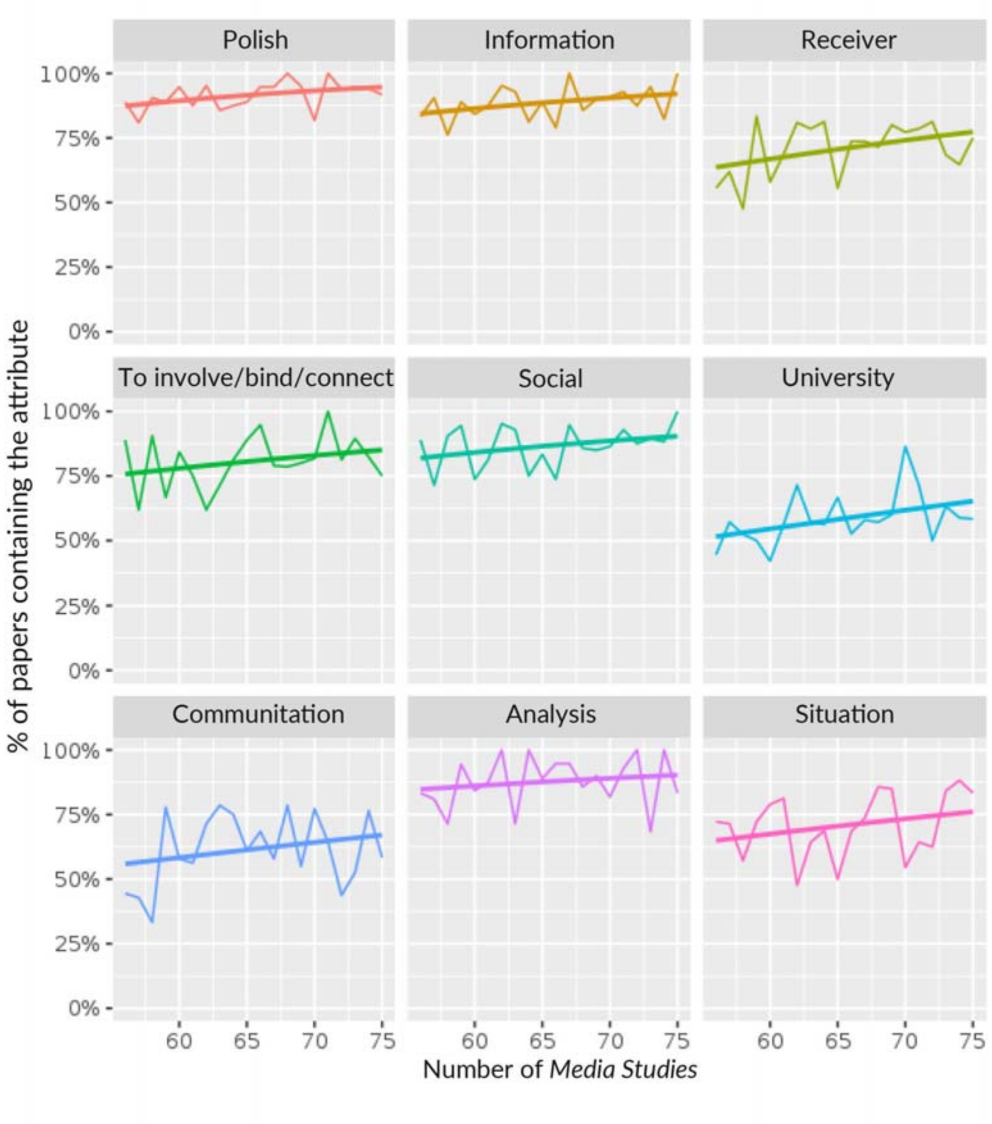

Figure 5. Topics that have the highest increase in citation in 2014–2018

Source: Own study

Figure 5 shows that the largest increase in popularity in the media are topics related to Poland. This shows that interest in patriotic and national values has increased. Next comes information and the related attribute "communication." One can see the increase in the significance attached to the circulation of information in the modern world. The circulation of information. The "receiver" attribute indicates the importance that the receiver plays in the media. Even in scientific and research communities, whose voice is *Studia Medioznawcze*, they grant a significant role to the recipient.

Downward Trends in the Media

The biggest decrease in interest over the last 5 years concerned topics often discussed in the media, but their citation decreased over time.

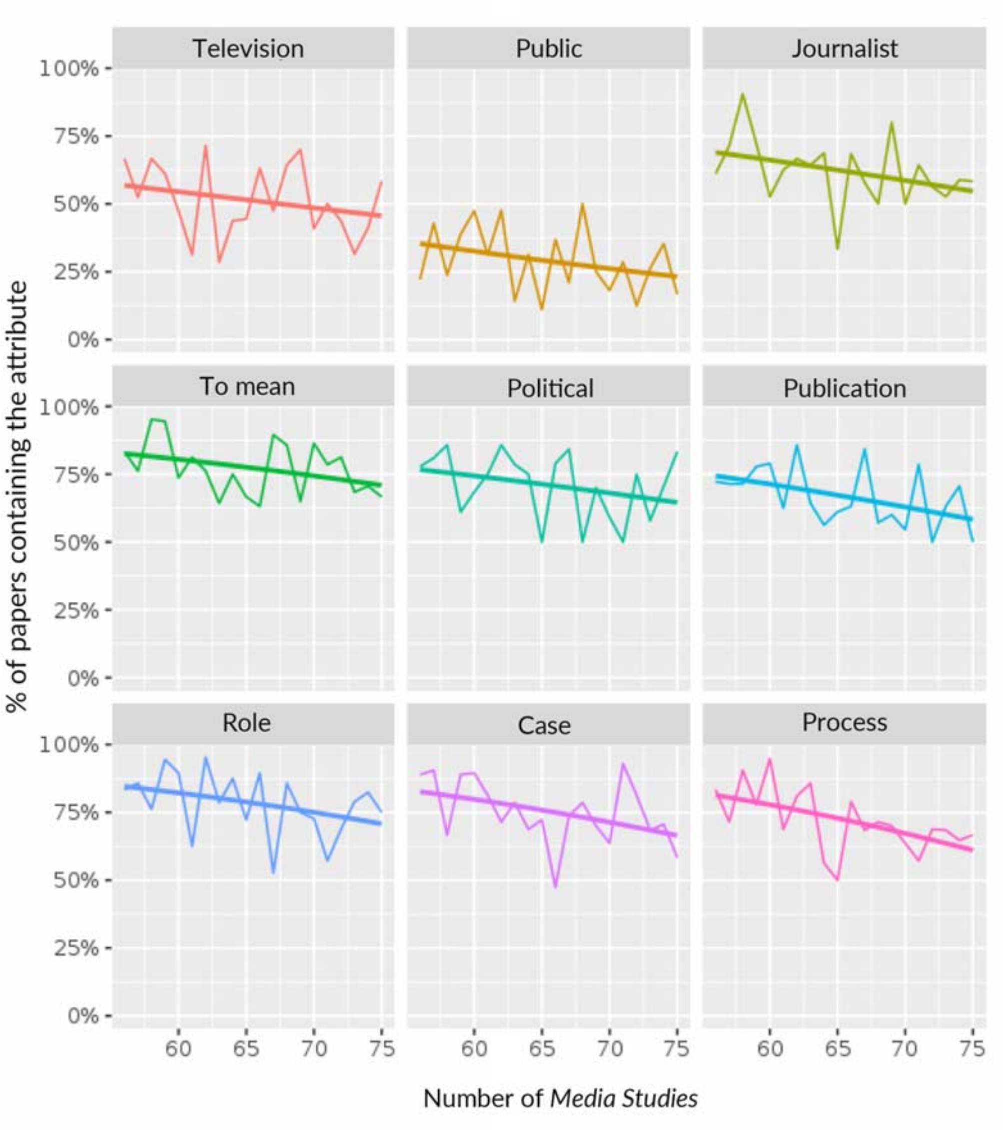

Figure 6. Topics that have the largest decrease in citation in 2014–2018

Source: Own study

The first place in terms of inheritance is “television.” This indicates a progressive decline in the importance of television in modern media. Research reports a gradual decrease in viewership and a decrease in the impact on the opinion of this medium. Another public attribute indicates a decrease in the "public" media so characteristic of the previous era. A "journalist" in the media also plays a smaller role. In modern media, there is ease of speaking through online forums and any member of the Internet community may act as rapporteur or opinion-maker. The next significant topic about the decline in interest is "politics." It seems that the period of profound political changes is behind us and the directions of media interest have moved to other areas.

Substitute in the Media

Analyzing the biggest increases and decreases in trends in the media, it was noticed that some topics popular in the media a few years ago are replaced by other, gaining the interest of the receiver. Below is a list of substitutability of trends in the media. Wherever possible, a corresponding upward topic was assigned to the topic with the largest decrease, which replaces it. Replicability is understood as a change of marking within a given area.

Table 2. Replacement of trends in the media

Downward trend	**Upward trend**
Public	Social
Journalist	Receiver
Publication	Communication
Case	Analysis

Source: Own study

The first direction of substitutability is moving away from public topics and turning to social areas. Addressing content to the public is losing importance. One can see a greater importance attached to directing the media message, selectively to social groups. Another interesting combination is a "journalist" vs. "receiver." A move away from traditional journalism-based media can be in favor of a receiver-oriented message. It is the receiver, not the sender (journalist), who is in the center of attention of the media. A related statement is "publication" vs. "communication." The publication can be equated with journalism in the traditional sense of the word. Communication means active dialogue with the receiver and recognition as an active participant. It seems that the passive role of the receiver in the media (for example, television, printed press) ceases to have such significance, in favor of discussion groups or social networking sites giving the opportunity to speak to all participants of the debate. An interesting combination is "case" and "analysis." This can be interpreted as a change in approach from case reporting to case orientation. The evolution in the media is heading towards analytical methods. The recipient expects not only reliable information, but also answers or reliable forecasts. The answer is the dynamic development of Big Data, which introduces a new quality to our lives by using quantitative data analysis tools and artificial intelligence.

Characteristics of Trends

In the "Quantitative Analysis" part, trends in the media with the highest growth dynamics were identified. Trends represented by words-attributes. A single word-attribute indicates the subject of the field, however, such a generalization can lead to ambiguity and interpretation of the trend may be imprecise. Attribute words that characterize individual trends should be indicated. In this way, each of the selected trends will be described by a set of words that characterize it. An analysis of covariance has been used to select characteristic words. The words that are most closely correlated to trend words have been selected. A binomial analysis of covariance has been performed using a linear Phi correlation coefficient.

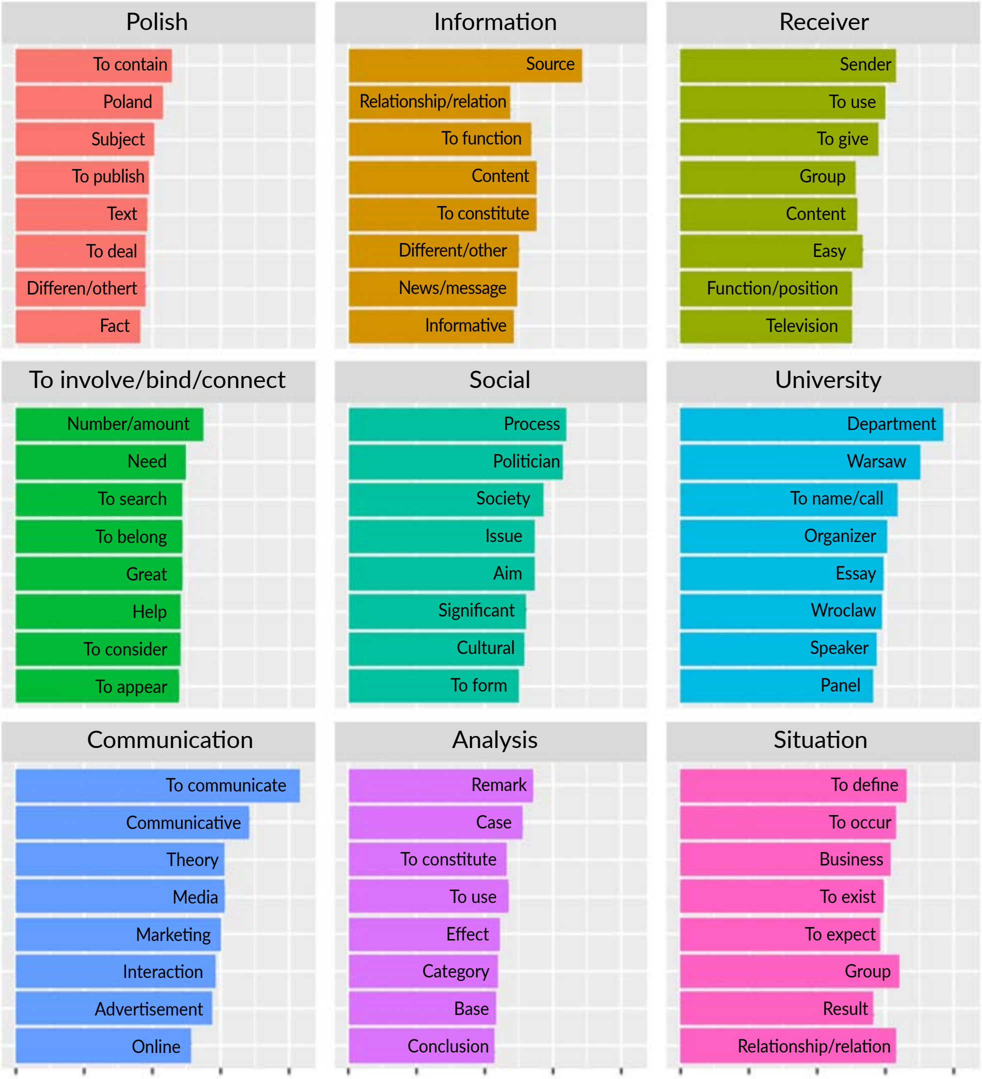

Figure 7. Words correlated with trend words

Source: Own study

Figure 7 contains the words most related to trend words. For example, the "information" trend is described by characteristics that determine the context in which it occurs in *Studia Medioznawcze:*

- source
- relationship/relation

- to function
- content
- to constitute
- different/other
- news/message
- informative.

It can be seen that the most important feature of information is its “source,” which is understandable due to the scientific nature of the quarterly. The subsequent attributes “relationship/relation,” “content,” and “to constitute” confirm the authors’ great care to verify the content published.

The terms “receiver”—“sender,” “group,” “content,” “function” indicate the context in which it is used. It should be concluded that this is about the receiver of the transmitted media content. Attention can be paid to the quality of contact with the receiver—“easy,” “to give.” Trend attributes allow one to place it in the context in which it is used from the source data. A graph has been created to better illustrate the relationship between words-trends and their attributes

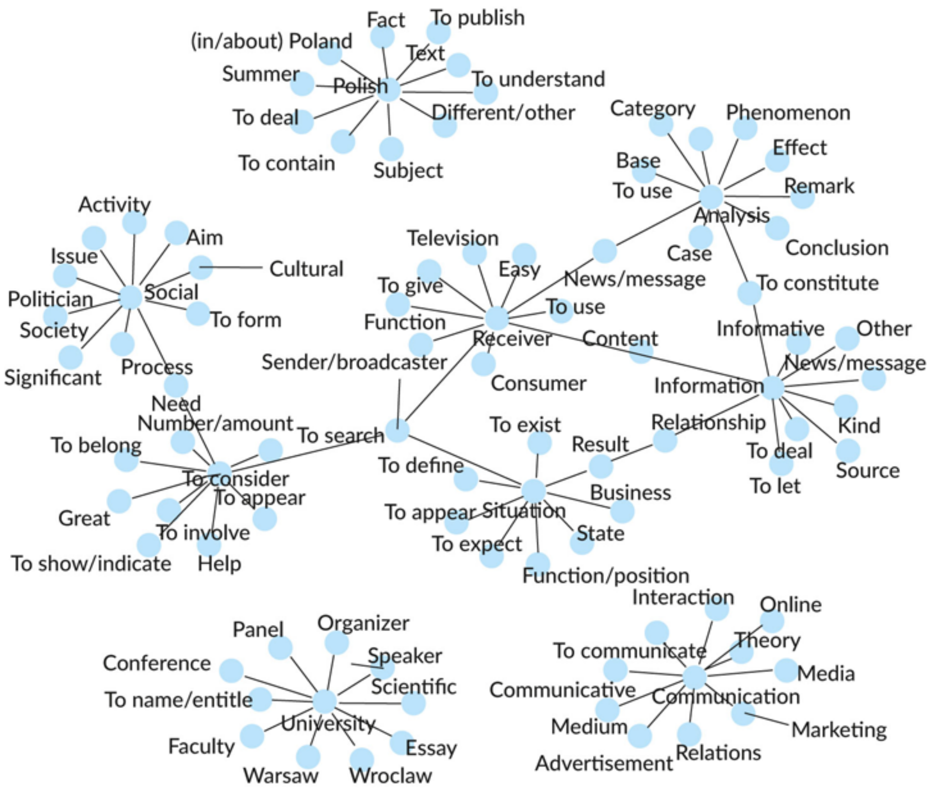

Figure 8. Correlation of words in a graph

Source: Own study

In Figure 8, one can see words that are characteristic of more than one trend. These words are not equated with trends, but due to the universality of occurrence and high correlation with trends, they should be considered an important element of analysis. For example, "television" and "news/message" occur as a characteristic of words-trends for "receiver" and "analysis." In turn, the "group" attribute is associated with three trends—"to involve/bind/connect," "situation," and "receiver." Trend words with at least two connections are a cluster that should be analyzed together. In Figure 7, related words-trends are—"analysis," "receiver," "information," and "situation". "University," "Polish," and "communication" are separate trends.

Comparison

One need to verify the compliance of trends in the media forecasted by consulting companies with trends of scientific and academic communities represented by the quarterly. According to the PwC report, Figures 1 and 2, indicate the most promising directions (columns) for the media:

- VR / virtual reality
- OTT / over the top (delivery of streaming content via the Internet directly to the recipient, bypassing telecommunications and television operators)
- Online advertising
- Video games
- E-sports.

Calculations were made of the frequency of trends forecasted by the consulting company PwC in the quarterly *Studia Medioznawcze.*

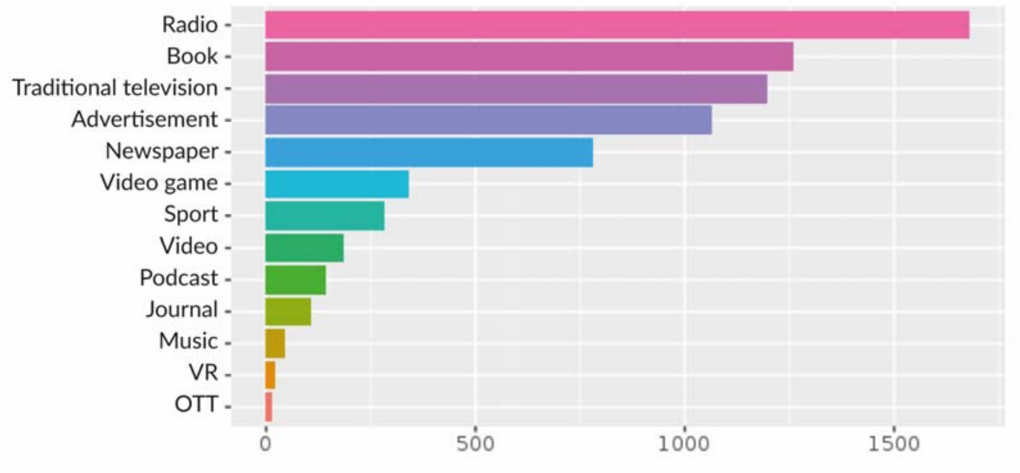

Figure 9. Number of occurrences of PwC trends

Source: Own study

The forecasts for the development of PwC media (Figures 1 and 2) were compared with the number of these trends in *Studia Medioznawcze* (Figure 9). Virtual reality and OTT are the most promising ones according to PwC, and have insignificant representation in *Studia Medioznawcze*. In turn, trends that are characterized by a large representation (above 1000): "radio," "book," "television," and "advertising" according to the PwC report have slight or downward development potential. Therefore, it should be concluded that media forecasts presented by commercial companies, including consulting companies, differ significantly from the trends represented by the scientific and academic community. Unfortunately, here comes the conclusion about the negligible usefulness of the conducted research!

The indicated differences result from the large inertia of the scientific community in relation to technological, organizational, and marketing innovations in the media sector. That is why the willingness to explore already known areas is understandable, even taking into account the innovative processes taking place in them, and avoiding involvement in new emerging media fields (e.g. Internet, games). In fact, a similar phenomenon occurs in the global information market, which is confirmed by data obtained from the site https://books.google.com, which allows a quantitative analysis of publications on specific topics (described using keywords). The obtained data, although their validity is limited to 2008, confirm changes in trends. Since the end of the last century, there has been a clear decline in interest in television, radio, video, and printed magazines, while, although on a small scale (until 2008), interest in games, VR, and sports is increasing.

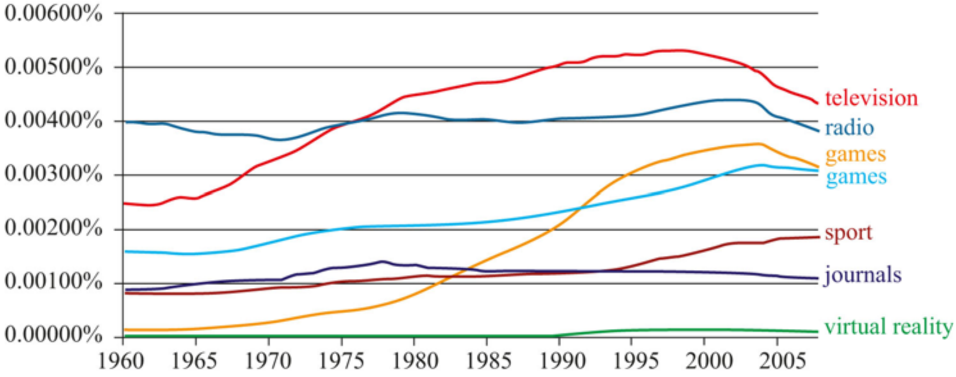

Figure 10. Trends for selected media in 1960–2008

Source: Google trends

As can be seen from the data presented, the scientific and academic communities are balanced and do not follow the latest innovations. Thus, traditional media such as radio, books, and television still occupy a significant place and will not be completely supplanted from the world of media.

Bibliography

Gogołek, W., & Jaruga, D. (2016). Z badań nad systemem rafinacji sieciowej. Identyfikacja sentymentów. *Studia Medioznawcze*, *4*(67), 103–111.

Kantar. (2019). *Kantar Unveils Predictions for 2019 Media Landscape.* Retrieved from https://uk.kantar.com/tech/digital/2018/2019-media-predictions/

Mayer-Schonberger, V., & Cukier, K. (2017). *Big Data, Rewolucja która zmieni nasze myślenie pracę i życie.* Warsaw: MT Biznes.

Ooms, J. (2018). *Hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. Retrieved from https://CRAN.R-project.org/package=hunspell

PwC. (2017). *PwC's Entertainment & Media Outlook Forecasts U.S. Industry Spending to Reach $759 Billion by 2021*. Retrieved from https://www.prnewswire.com/news-releases/pwcs-entertainment--media-outlook-forecasts-us-industry-spending-to-reach-759-billion-by-2021-300469724.html

PwC. (2018). *Perspektywy rozwoju branży rozrywki i mediów w Polsce 2018–2022. Retrieved from https://www.pwc.pl/pl/pdf/publikacje/2018/media-i-rozrywka-2018-raport-pwc.pdf.*

Silge, J., & Robinson, D. (2019). *Text Mining with R A Tidy Approach*. Retrieved from https://www.tidytextmining.com/

Surma, J. (2019). *Cyfryzacja życia w erze Big Data. Warsaw: PWN.*

Sztuka, film, media. (2015). *Strategia rozwoju rynku medialnego w Polsce 2015–2020. Retrieved from http://sztukamediafilm.pl/wp-content/uploads/2014/09/SMF-Strategia-rozwoju-rynku-medialnego-w-Polsce-2015-2020.pdf.*

Team R Core (n.d.). *The R Project for Statistical Computing*. Retrieved from https://www.R-project.org/.