Social Communication
Volume 2 (2019), pp. 32-44
DOI: 10.2478/sc-2019-0007

MASS AUTOMATED INTERNET ANALYSIS AND CYBERSPACE TRANSPARENCY

Marek Robak[1]

Abstract

One of the roles of media research is to explain social phenomena. The Internet became a place where society expresses itself and where society could be influenced or even manipulated. Therefore, online communication analysis becomes a tool that is expected to guarantee the transparency of the social communication process. Unfortunately, the size of the Internet makes analysis difficult, and traditional methods of analysing communication are not always enough or force the researcher to focus on a fragmentary data. The author asks a question which research methods are suitable for Internet research and allow to improve transparency. It focuses on the method group referred to in the article as Mass Automated Internet Analysis. In the final part, the author shows examples of several – existing or being developed – research methods and techniques (including data collection and data analysis field), what research methods can improve the quality of digital communications research.

Key words: The Internet, internet research, online research methodology, Mass Automated Internet Analysis, Webscan, WebStream, transparency of the Internet

The subject of my analysis is the problem of the lack of transparency in research, describing the impact of the Internet on society. Although many Internet analyses are currently being carried out, the weakness of many is the fragmentary nature. As the number of messages and communicators increases, the study becomes complicated or lasts too long. In the article, I wonder if the implementation of large-scale automated Internet research can improve this situation.

STRUCTRURE AND METHODOLOGY

The analysis will be carried out in three steps. Initially, I will focus on fairly general question about whether society really needs objectified media-style research on the Web, referring this question to the context of recent large scale events from last ten years with a high social impact, such as the Facebook – Cambridge Analytica scandal. Based on a historical analysis of those events I will demonstrate, why objective Internet research could play an important social role, comparable to the role of "old media" in pre-Internet times.

In the second part I will characterize a specific category of mass and automated

[1] Marek Robak PhD, Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie, Wydział Teologii, Instytut Edukacji Medialnej I Dziennikarstwa, ORCID: 0000-0001-8331-2891, marek.robak@gmail.com

Internet research in terms of whether they are able to help solve the problems identified before, mostly the lack of transparency in Internet communication This part concentrates on theory of research and its goal is to identify the most significant properties of high-quality internet research, based on experience of classical media analysis, contemporary Data Science techniques and achievements of computer science in the last decade.

The third part is practical – I will discuss several methods of network analysis, showing on selected examples how they can overcome existing limitations. Although this kind of list should never be treated as finished nor completed in such dynamic environment like the Internet, presented techniques could be a good inspiration for researchers and a point for further discussion. Presented list consists of both well-known research techniques and experimental ones, based on author's own research projects from last years.

SOCIAL VALUE OF WEB RESEARCH

Internet research in itself seems interesting. After all, the Internet reaches billions of people, it has a great deal of impact, especially in relation to younger audiences. Network communication is extremely fast and diverse. And most importantly, it appeared relatively recently and has not yet been sufficiently studied. The Internet has also become an important element of the modern economy, and its analysis has practical and commercial value. Younger researchers are interested in the Internet because they like to focus on it and are happy to use this method of communication themselves. Any of these reasons is sufficient to engage in Internet research, even if they were of a niche nature. However, this article is not about this type of Internet research. In addition to many very detailed online studies, the value of which is even based on the fact that they help, for example, to improve the effectiveness of an advertising campaign, there are studies whose results help to better understand the larger processes that take place in society. Today's Internet could be placed rather inside that outside the society, which explains, why a certain group of research on the internet also becomes research on society itself. For many organizations, not only commercial ones, web research tools can help to improve the whole organization, not only it's "digital part" [Robak, 2017b].

Manuel Castells, a famous creator of the theory of network society, introduced the concept of real virtuality. Theories of this sociologist were ahead of their time, so when Castells described real virtuality, the public discourse rather talked about the importance of broadly understood virtual reality and cyberspace as something that comes through the Internet. An important feature of Castells' theory was the rejection of the then popular belief that the internet creates a kind of metaphorically understood alternative world; activity in it seemed to compete with activity in the current world (often called "real"). Meanwhile, Castells pointed out that this duality is not justified. He predicted that these realities would intertwine and social activity would increasingly be implemented online to the extent that it could be said that what does not exist on the Web seems to not exist [Castells. 2000]. Using the term taken from the founder of Microsoft Corporation, Bill Gates, computers act as the digital nervous system of society. Gates anticipated this many years before the spread of the Web [Gates, Hemingway, 2000].

Castells' conclusions at the time of publication seemed revolutionary, today we can take them for granted. However, just over the past decade there have been several significant events that give internet research practical and socially relevant significance. In 2013, an employee of the US National Security Agency (NSA) and the Central Intelligence Agency (CIA) Edward Snowden, decided to provide journalists with several hundred thousand confidential documents, indicating the fact that the US government is constantly tracking electronic communications around the world, including emails, connections telephones, text messages, chats and all kinds of network connections, as well as monitoring of some personal computers. In a recently published autobiography [Snowden, 2019], he explained his motivation. What prompted him to make the decision to disclose the secrets of the NSA's work, was the discovery, that the US agencies were collecting all possible data on a mass scale, gradually extending the period of their storage, and monitoring was carried out in relation to all Internet communication to which they were able reach, and not only in relation to persons suspected of committing a crime, for which the court would agree to their surveillance.

At the same time, in 2018, the world became aware of the use of data by dozens (estimated 30-87) million Facebook users by Cambridge Analytica, without their consent, for targeted political agitation, which could potentially affect Donald Trump's victory in the presidential election [Davies, 2015]. A year later a "hate speech scandal" broke out in Poland. Onet.pl, the leading Polish Internet portal, described how Deputy Minister of Justice Łukasz Piebiak commissioned an online hatred to slander the government of unfavourable judges on the Internet and collect compromising materials about them [Gałczyńska, 2019]. A group of judges who were in favour of the government was also disclosed. The publication resulted in the resignation of the Polish deputy minister of justice, and the parliament even considered the request to dismiss the minister of justice (the application was rejected).

The examples cited may indicate a more serious problem. They confirm that it is possible to monitor society and influence it through new unexplored mechanisms through the Internet. Although propaganda is certainly not a new phenomenon and did not appear with the internet, the complete lack of transparency of this process is worrying. For example, you get the impression that it is easier to document and examine propaganda in history than it is today. E. Snowden also claims that the internet itself, which in the first years of popularity appeared as an oasis of freedom and unrestricted communication, has turned into a network based on "surveillance capitalism" [Snowden, 2019, p. 14].

From this point of view, the scandals mentioned earlier have several features in common. They were not detected from the outside, we know about them only thanks to the information of those involved who, guided by their own conscience, revealed certain facts. Secondly, even after disclosure, the amount of information available is limited. Although they relate to electronic communication, which is theoretically easy to document, in none of the cases mentioned we do not have complete information, for example, all entries that could be independently analysed. Data, if any, is incomplete.

ONLINE RESEARCH TRANSPARENCY

Since many social phenomena occur on the Web, theoretically, Internet research could improve the transparency of social processes. In practice, however, we discover more restrictions: There is too much data to collect. Media experts are not always able to technically analyse large Big Data collections. Many research tools are commercial in nature with limited access. In many ready-made advanced tools, the research method is not explained and there is no access to sources, etc. Therefore, it is worth considering the features of online research that affect high transparency.

Completeness of facts

Each study starts with collecting material for analysis. For articles in magazines, it is only needed to get access to the set of journal years in the library. A collection of articles on a given topic or one author may contain from several dozen to several hundred items – a collection that can be quickly examined by just one person.

In the case of the Internet, the scale immediately becomes much larger. Facts can be: articles, comments, likes and shares, videos, audio, clicks, and many other types of data. The Internet does not have a library, and projects such as Internet Archive do not provide enough data for in-depth research. The average news portal can publish dozens of pieces of information a day, sometimes removing it or changing the place of publication. A popular social profile can trigger reactions in the form of thousands of entries in one go, and if we want to cover the entire ecosystem of information portals, for example in a selected country, the task becomes embarrassingly difficult.

We will quickly conclude that manual collection of facts on the Internet will very quickly have a negative effect on the quality and duration of the study. This is not the only limit. Not every online publisher wants to collaborate with the researcher and provide the necessary data on time. More and more publications, especially on social media and on ad servers, are targeted, so they will appear only in a specific, difficult to grasp context. Some publications, especially on social media, are only available to members of a closed group. Finally, with the growing interest in the protection of privacy and the introduction of GDPR, the possibility of collecting, even for statistical purposes, a lot of data and conducting external analyses and audits of traffic using third party cookies, even when they are pseudonymous [Robak, 2017; Robak, 2018].

Theoretically, we can solve many limitations by involving computers to collect and analyse data. We can thus collect more data and process it faster. However, this is not always the case. Two other trends can be observed. One is to limit the set of facts examined to such a small range that the study can be carried out according to art by traditional methods; the survey is therefore formally correct, but the conclusions are not of great social significance. The second tendency is to replace large-scale surveys with a questionnaire, which was a trend in early years of online social research. So instead of studying phenomena, we ask respondents about them. The survey is the correct form of social research, but in this case, it will reduce the quality of the entire survey, as it will be used to describe general beliefs, not measure the facts. This leads to many problems in social research of the Internet, for example, when there is a widespread belief in the existence of hate speech, but there are no indicators to describe its features, intensity or sources. However, when we measure these phenomena, the results are surprising [Krejtz, 2012].

Sampling

The facts must also be collected in the proper way. General theory of scientific research suggests some good practices [Wimmer, Dominick, 2006]. First of all, you need to collect research material regularly, maintaining the right structure, in the right amount.

Regularity in practice means that data must be collected many times a day, since the web portal can change its content several dozen times a day. If it is possible to extract this data from the archive, the task is simplified, if not, manual methods of analysis are no longer sufficient. Manually conducted research can even disturb them, it is enough to prove that the favourite time of the researcher's work (for example, in the morning, evening, on weekends) influenced the selection of examples.

As in many cases it will not be possible or too costly to record all of the material, sampling can be used. It involves taking a part (sample) from a larger set (population). Statistics rules allow us, based on the sample analysis, to draw conclusions about the entire population, with limited accuracy, provided the sample has been properly constructed. The problem of constructing a sample is widely described [Wimmer, Dominick, 2006; Babbie, 2005]. Although there is no place for deeper analysis, based on my own research it seems that variables such as gender, age, publication time, day of the week and place of publication could significantly affect the final results.

The size and stability of the sample is also important if we are unable to analyse the entire population. In general, we can assume that sets below 30 records are not suitable for classical statistical analysis (based on a central limit theorem), and sets with numbers from several hundred to several thousand records should already give decent stability, sufficient for most analyses. Drawing conclusions based on individual examples, often referred to as "anecdotal evidence", in most cases does not qualify as a scientific method, with some exceptions, such as hypothesis and case study. It seems that the messages on the Internet are so numerous and varied that with the help of anecdotal evidence it is possible to “prove” any theorem.

Comparing

In Internet research, there is often a need to repeat analysis on a larger scale, sometimes after some time. Therefore, the method should be repeatable, returning compatible results. In studies that aspire to diagnose large social processes, this repeatability and ability to compare results is necessary. There are many valuable methods of analysis that work well in the humanities and art, but if they cannot be objectified and extended on a very large scale, in some cases they will not be useful, although they can be used at the stage of preliminary research and hypothesis.

Let us return to the example of hate speech. There is a widespread belief that many statements (mostly posts) on the web are aggressive. However, until we are able to quantify this feature at least on a scale, the following questions cannot be answered: What percentage of comments on this site is aggressive? Which of the websites surveyed has more aggressive readers? Is the slightly ironic statement aggression? According to the popular saying of researchers, apples should be compared with apples and oranges with oranges. Therefore, methods based on the subjective assessment of the researcher may seem tempting to study Internet communication, but they quickly reveal their limitations. This does not mean, however, that we must limit ourselves to primitive collection of numbers. There are many techniques using, for example, text analysis, the method of competent judges, Kansei analysis, fuzzy logic or supervised machine learning, in which in-depth analysis can finally give comparable results. It should not be forgotten that, for example, in psychology, standardized personality tests are used to detect very complex phenomena while maintaining a simple form and comparable results. So, it is feasible even in typical social science.

Ability for quick analysis

The expected feature of Internet research is speed, understood as the ability to provide analysis results in a short time (which, of course, does not exclude longitudinal research). “Short time” is not strictly defined, but I assume that hours or days rather than months and years should elapse from the occurrence of the fact (e.g. publication of the article). If the study lasts for months or years, then it probably won’t be possible to extend its scope at reasonable costs. Delay not exceeding a few days brings many benefits – it allows you to optimize editorial content, improve online campaigns, and also quickly detect a social phenomenon. For example, you can imagine detecting abuse during an election campaign or promoting illegal behaviour – timely detection allows society to react in a timely manner. A great example is the moment when the Cambridge Analytica scandal broke out – although it provoked valuable social discussion, from the point of view of the presidential election in the United States it doesn’t matter anymore, the president was elected and sworn in.

The speed of research and the size of the data analysed also means that in practice we are talking only about partially or fully automated research using computers. In communication research, this conclusion is neither obvious nor simple to implement. There are several methods of analysis recognized in media studies related, for example, to text interpretation, examining the relationship between sets of articles, assessment of the persuasiveness of visual communication, where process automation is not obvious, if at all possible. However, it seems that the achievements of information technologies in the last 15 years significantly expanded these possibilities.

Mass Automated Internet Analysis

I have discussed a certain set of expectations regarding internet research methods.

These methods have some common features, which I will describe as Mass Automated Internet Analysis.

Mass – means that these studies can be carried out on a large scale, which gives statistically stable results and allows characterization of large-scale phenomena that go beyond one information portal or one profile in social media. *Automatic* – means that the tests are partially or completely automated, which allows obtaining fast results from large data sets and easy repetition of the test procedure on a large scale and in subsequent tests, thanks to which the procedure is repeatable, scalable and verifiable.

Although Mass Automated Internet Analysis could trigger a change on the field of media research of internet communication, it seems to have some limitations, which could be summarized in three categories. (1) Automatic processing can cause shallow analysis. There is a tension between the quality and speed of research, which cannot be solved in a simple way. None of the approaches meets fully all requirements. (2) Computer-based analysis is strictly connected with the skills of the researcher. Looking at the agendas of media studies, there is a lack of computer science and Data Science courses. From historical reasons, in comparison to mathematicians, physicians, and life-science specialists, media researchers have smaller experience with automated data processing. On the other hand, there is a risk of "blind" using of software tools with no understanding of technology and context, which is never a good practice for a research. (3) In last years some analytical techniques are seen as a threat to privacy. Even when these threats appear to be overinterpreted [Robak, 2017; Robak, 2018], it is not possible to ignore the social fear of digital research.

EXAMPLES OF ANALYSIS

It is difficult to imagine how more accurately the research procedures would look basing on pure theory,. That is why I have prepared a list of examples of research techniques that can be used in Mass Automatic Internet Analysis. The list is still open, this field is constantly evolving. The examples relate to two areas: data collection and analysis. I tried to make them diverse, which is why both popular techniques used in large commercial research and innovative methods that I am working on as part of my own research were included.

Data collection examples

Email Dataset

The story described here is one of the oldest studies in the history of Data Mining. In 2001, the American energy tycoon, Enron, went bankrupt in connection with disclosure of falsified financial records. Just a year before bankruptcy, the company generated revenues of $ 101 billion and employed 22,000 employees. The bankruptcy also led to the collapse of the prestigious Arthur Andersen consulting company, which was proven to have participated in the forgery of Enron's financial documents.

The story of Enron has been studied in many different ways. In the case of such a large organization and such serious allegations, the basic research problem is to prove how the whole organization behaved, and not just to find a single anecdotal evidence on the basis of which, at most, it is possible to prove the employee's abnormal behaviour.

Investigating the situation of Enron a huge collection of employees' email correspondence was extracted. This collection, called Enron Email Dataset, contains over 600,000 emails from correspondence maintained by 158 employees[2]. The total volume of this collection is 1400 MB, for comparison, the full Bible in English is about 5 MB. Currently, the Enron collection is widely used as training material in text analysis and machine learning. It should be noted, however, that this collection is quite old from today's

[2] https://www.cs.cmu.edu/~enron/

point of view and since its publication it is difficult to find equally rich examples in which the public would gain access to extensive documentation of the case.

Traffic research

By traffic research I understand mass collection of all traffic related events on a website or application. In the most typical case, these are all page views of a given website, sometimes supplemented with additional events such as monitoring of video enabled or adding a product to the basket. Traffic usually brings very large data sets. For example, according to Gemius research, the largest Polish portals reaching over a dozen million people generate tens of millions of page views per day. Thus, the monthly collection of all page views of such a portal gives nearly two billion records for analysis.

The traffic survey consists of two stages: (1) Collecting the full set of events. (2) analysis, for example calculation of basic metrics such as, for example, number of page views, visits, number of unique users, average visit duration, bounce rate, number of entries from external sources or video playback time. A more advanced analysis is also possible, e.g. study of traffic flow paths through the site.

It could seem that traffic research is one of the most popular on the internet, as the tools described above are based on Google Analytics, Gemius Prism and Facebook Stats. Ready tools, however, may lead to superficial analyses, focusing on too simple success rates or on metrics favoured in a given tool. In the early years of Facebook's popularity, a known misunderstanding was using the number of profileslike as an indicator of success and estimating the number of people you reach on that basis. However, this indicator was used cumulatively, adding up likes from the entire period, which almost always increased; at the same time, the number of unique users of web services was given on a monthly basis, which was a much better estimation of the number of people interested in a given content. Another type of mistake can be made by focusing too much on the number of pageviews. It may indicate popularity, but also too much fragmentation of the page structure, poorly designed navigation through which users enter a larger number of pages or the existence of a small group of fans who make a lot of clicks but do not increase the reach. Therefore, as Kaushik (2009) rightly points out, every research on the Internet must be preceded by understanding its own context and goals. Thus, the use of ready-made popular tools at the first stage speeds up the work of the researcher, but can also limit him. That is why it is worth considering conducting more independent traffic research on raw data. The growth of Data Science in last years resulted in the increase in popularity of tools for self-analysis of data, such as R language, Python Pandas, SAS, Hadoop, Spark.

Web & social media scrapping

The scrapping type test involves the automatic downloading of website content, including social media. This way you can analyse various elements – articles, comments, photos, multimedia, likes, etc. Research consists of three steps. (1) Creating a list of resources to download. For this purpose, a special program passing through the given website (web spider), news feed (e.g. RSS) or publication download mechanism (API) made available by the platform owner is used. (2) Elements of interest to the researcher (e.g. text, graphics) are extracted from the internet and pre-processed. (3) The collected set is subject to further analysis, it can be content analysis or links between its elements (social network analysis), [Wasserman, Faust, 1994].

The operating diagram described here is, for example, the basis of the Google search engine, which indexes a significant part of the content available on the Internet. Then creates a map of links between pages, assesses their content and importance using an algorithm whose first version was known as PageRank [Brin, Page, 1998]. In the scientific community, a good example of attempts to apply this approach is the web content refining method developed at the University of Warsaw [Gogołek, Jaruga, 2016], in which the content of a set of web pages is downloaded and, after cleaning, subjected to text analysis.

Some may fear that scrapping as a technique used by Google is advanced and too extensive. This is indeed the case with the global internet survey. This technique, in a much simpler version, can be and is used on a smaller scale: to extract the current prices of products in stores, obtaining spammers' e-mail addresses, and above all to study the content of communication: texts, posts or comments. For communication research, this primarily means being able to quickly download a set of content that would be difficult to gather manually.

OSINT and WebScan

On the Internet, the term OSINT (Open Source Intelligence), also known as white intelligence, refers to obtaining information based on publicly available data sets. Over the last dozen or so years of researching the internet, it has been discovered that the simultaneous use of many different independent sources can give information about a value much greater than the sum of these sources separately. The idea of grouping data from various sources was an inspiration to create in 2016-2018 the author's assumptions of the WebScan method [Robak, 2018b]. The idea is very simple: there are many technical tools for testing websites and extracting information from public records. These tools are not used every day for media studies and seemingly do not give results interesting to the world of media. However, when I conducted tests on real collections of websites, it turned out that I could get results that are interesting from the point of view of communication research. It is important because this data can be obtained automatically in a very short time, without the need to contact the publishers of the site individually or obtain additional consent.

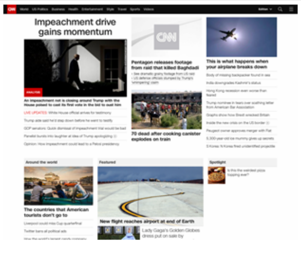
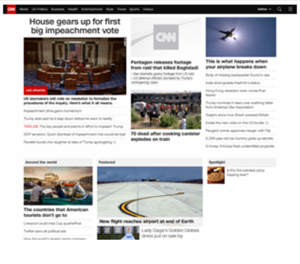
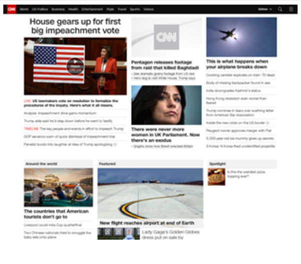
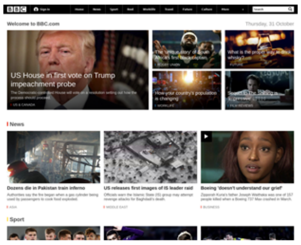
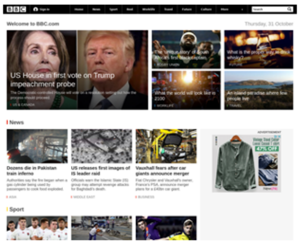
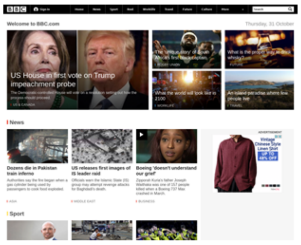
One of the collections, on which I conducted tests, was a collection of all official websites of Roman Catholic dioceses in Poland, on which 19 different tests were performed. The collection seemed interesting because of the similarities arising from religious affiliation, but also the differences resulting from the regional nature of the diocese and their high autonomy. It turned out that with the help of automated WebScan procedures it is possible to extract a lot of information, here are examples of conclusions drawn from automatic and semi-automatic analysis:

> (1) Without contacting the publishers, it was possible to determine the time of domain creation. The average age of the domain was 10.8 years (study 5.14). (2) 31% of websites used WordPress content management system to edit the site, 26.2% used Joomla and 2.4% from Drupal. These results differ from the internet average. (3) Traffic measurement tools were found on most sites (61.9%), mainly Google Analytics. (4) The parties differed significantly in technical quality. 31% charged very quickly (<100 milliseconds) and 21% unacceptably slow (> 1000 milliseconds).

WebStream

The WebStream research method was proposed for the first time in 2019 to solve the previously described problem of sampling website content. Content on the Internet changes many times throughout the day, making it difficult for researchers to precisely analyse content and how it changes. Therefore, the needs of the study have been developed a special platform that archives selected websites several times a day. The collection developed in this way not only has historical value. It allows you to accurately analyse how pages change. It also allows you to compare several different websites registered at the same time.

Table1: Time-series research of CNN and BBC websites using WebStream method

2019-10-31		
13:00 GMT	14:00 GMT	15:00 GMT
CNN		
		
BBC		
		

Data analysis examples

Text mining

In communication research, the most important subject of analysis is still the text itself. There are, however, many excellent text interpretation techniques taken from literary studies or philosophy. Although they provide in-depth analysis, it is very costly, if possible, and difficult to compare on a large scale.

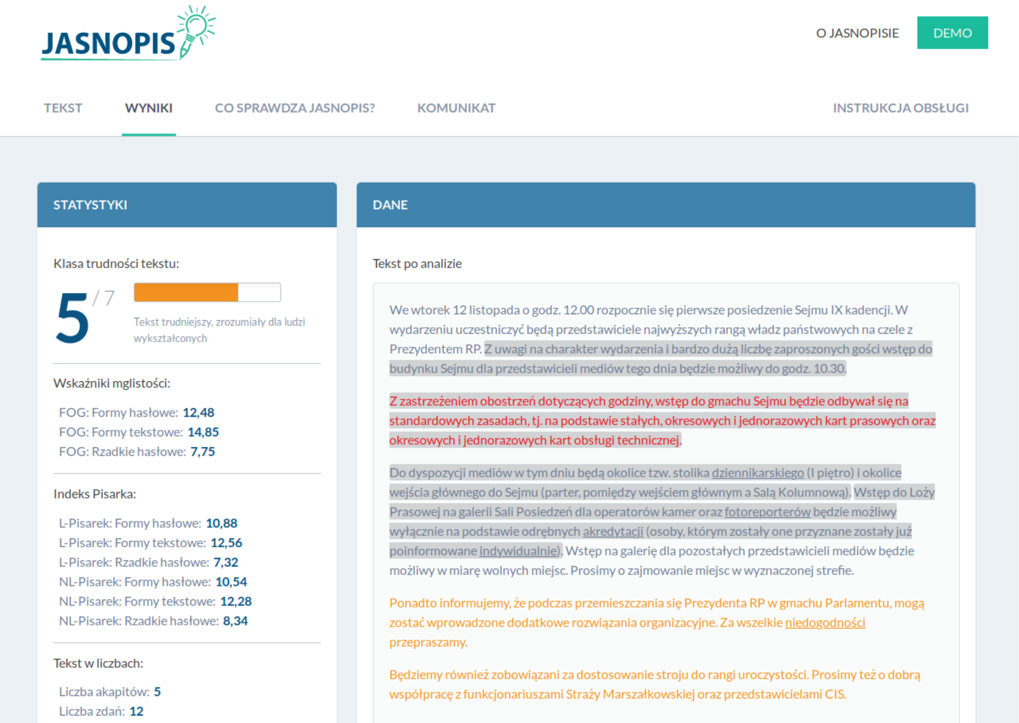
The second group of analysis is based on the method of critical judges. They read each text and categorize according to strict criteria. This approach can be used to analyse the average size of sets and gives results that are easy to analyse and present. It is also an approach well grounded in early media research [Berelson, 1952].

The third group of analysis is based on fully automatic text analysis. This makes the procedure very efficient and enables the processing of millions of texts. It also allows in-depth analysis of hypertext (links, hashtags). On the other hand, this approach can be accused of insufficient quality resulting from the limited possibilities of algorithms to understand the text. The quality assessment of these algorithms goes beyond the topic of the article, so I will limit myself to indicating the most important methods of analysis in this group.

(1) Basic text statistics. It comes down to measuring the length of elements of the text, such as title, lead or content of the statement. The method seems rather primitive, but it allows to describe a set of texts and group them into categories quickly and easily. It is also worth noting that the length of the text significantly affects its intelligibility and may also characterize the way the page is edited. For example, one of the studies [Robak, 2018b] showed that websites in the same category can have very different text lengths, in the studied case for Polish language the average was 51 characters, the average text lengths ranged from 23 to 100 characters per service, with a standard deviation of 15.6 characters. In this case, the simplest statistical measurement revealed large differences in article editing strategies, which was confirmed in further studies.

(2) Foggy text index. An extension of the usual statistical analysis of the text is the method developed by R. Gunning (1952), also adapted to the needs of Polish lan-

guage [Gruszczyński, Ogrodniczuk, 2016]. Based on the analysis of the length of words and sentences, the index allows you to determine what level of education the reader must have to understand the text. This allows you to compare texts and services with each other. Here is an example of an article in Polish taken from the Polish parliament. Analysis by the Jasnopis tool shows that it is "a more difficult text, understandable for educated people".



Klasa trudności tekstu:

5 / 7

Tekst trudniejszy, zrozumiały dla ludzi wykształconych

Wskaźniki mglistości:

FOG: Formy hasłowe: **12,48**
FOG: Formy tekstowe: **14,85**
FOG: Rzadkie hasłowe: **7,75**

Indeks Pisarka:

L-Pisarek: Formy hasłowe: **10,88**
L-Pisarek: Formy tekstowe: **12,56**
L-Pisarek: Rzadkie hasłowe: **7,32**
NL-Pisarek: Formy hasłowe: **10,54**
NL-Pisarek: Formy tekstowe: **12,28**
NL-Pisarek: Rzadkie hasłowe: **8,34**

Tekst w liczbach:

Liczba akapitów: **5**
Liczba zdań: **12**

Tekst po analizie

We wtorek 12 listopada o godz. 12.00 rozpocznie się pierwsze posiedzenie Sejmu IX kadencji. W wydarzeniu uczestniczyć będą przedstawiciele najwyższych rangą władz państwowych na czele z Prezydentem RP. Z uwagi na charakter wydarzenia i bardzo dużą liczbę zaproszonych gości wstęp do budynku Sejmu dla przedstawicieli mediów tego dnia będzie możliwy do godz. 10.30.

Z zastrzeżeniem obostrzeń dotyczących godziny, wstęp do gmachu Sejmu będzie odbywał się na standardowych zasadach, tj. na podstawie stałych, okresowych i jednorazowych kart prasowych oraz okresowych i jednorazowych kart obsługi technicznej.

Do dyspozycji mediów w tym dniu będą okolice tzw. stolika dziennikarskiego (I piętro) i okolice wejścia głównego do Sejmu (parter, pomiędzy wejściem głównym a Salą Kolumnową). Wstęp do Loży Prasowej na galerii Sali Posiedzeń dla operatorów kamer oraz fotoreporterów będzie możliwy wyłącznie na podstawie odrębnych akredytacji (osoby, którym zostały one przyznane zostały już poinformowane indywidualnie). Wstęp na galerię dla pozostałych przedstawicieli mediów będzie możliwy w miarę wolnych miejsc. Prosimy o zajmowanie miejsc w wyznaczonej strefie.

Ponadto informujemy, że podczas przemieszczania się Prezydenta RP w gmachu Parlamentu, mogą zostać wprowadzone dodatkowe rozwiązania organizacyjne. Za wszelkie niedogodności przepraszamy.

Będziemy również zobowiązani za dostosowanie stroju do rangi uroczystości. Prosimy też o dobrą współpracę z funkcjonariuszami Straży Marszałkowskiej oraz przedstawicielami CIS.

Figure 1: Text from the parliament website examined by Jasnopis tool.

(3) Frequencies. The widely used technique is based on calculating the frequency of words in text collections and comparing them with each other [Płaneta, 2018]. On this basis, you can determine the subject of the text, identify the characteristics of the indicated sets of texts (e.g. one author or from one subject), highlight keywords, examine the similarity of the text to the reference group, determine semantic field, detect language or dialect automatically. The following example shows the most important words in media in Poland on September 1 (anniversary of the outbreak of World War II):

Table 2: Word of the day, Polish language, September 1, 2019. Source: Clarin Network, http://slowadnia.clarin-pl.eu/#/default/2024/true

	Original keyword (PL)	Translation (EN)
1	osiemdziesiąty	eightieth
2	wojna	war
3	światowy	worldwide
4	wybuch	explosion
5	drugi	second
6	rocznica	anniversary
7	obchody	celebration
8	Westerplatte	Westerplatte
9	Wieluń	Wielun

	Original keyword (PL)	Translation (EN)
10	Frank-Walter Steinmeier	Frank-Walter Steinmeier
11	uroczystość	celebration
12	huragan	hurricane

(4) Sentiment analysis and Natural Language Processing. Language analysis software is used to determine quickly whether a statement is neutral, negative or positive. The effectiveness and validity of this method can be discussed. However, it is used not so much to precisely define the overtones of individual texts, but rather to massively evaluate large sets of utterances, for example comments, and to compare these sets with each other. There is also a wider field of language research – Natural Language Processing. The purpose of this research is to create algorithms that allow for full grammatical language analysis, automatic creation of correct sentences, semantic web analysis [Dobosz, 2012], as well as machine translations. These studies are included in the area of artificial intelligence and although it is impossible to summarize their capabilities in one sentence, in recent years there has been significant progress in this field. It must also be added that the effectiveness of NLP algorithms varies depending on the specific language.

Colour mining

The colour used in publications is certainly relevant to their overall perception [Kasperski, Boguska-Torbicz, 2008] and is of interest in the areas of psychology, brand marketing, media studies, User Experience, art, but also medicine. An interesting research question, however, is the possibility of analysing it for larger collections of publications. Searching for the answer to this question, I used the data collected for the purposes of the WebStream study also to automatically analyse dominant colours. Here is an example of comparing the dominant colours in a set of websites from the same category.

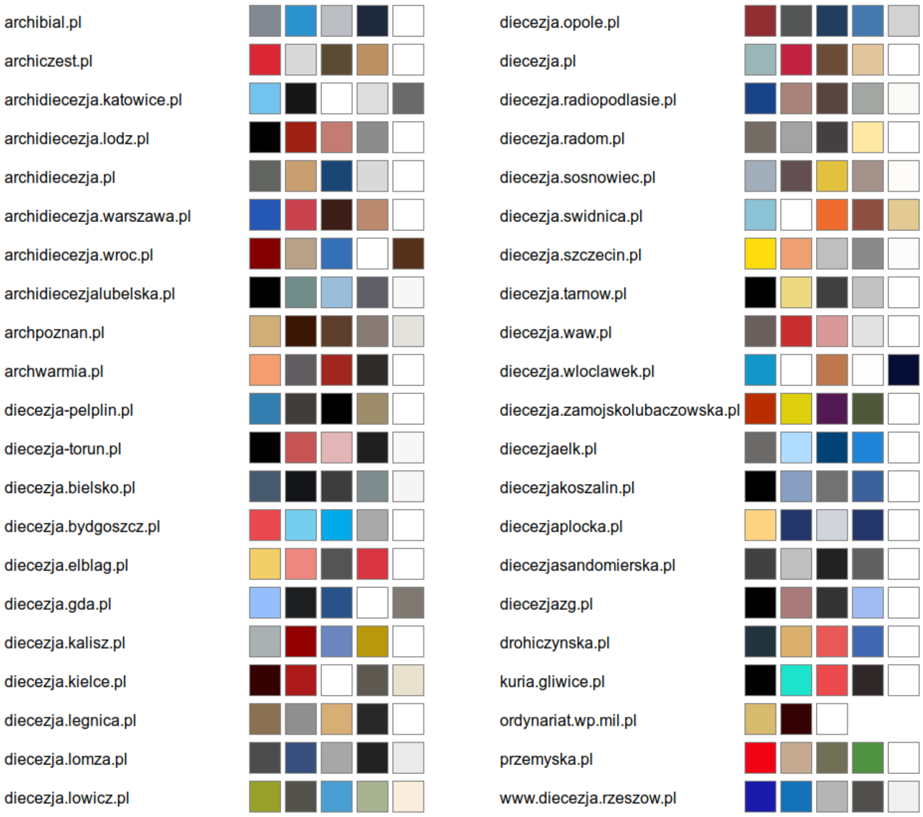


Figure 2: Analysis of dominant colors of websites. Source: Robak 2018b.

SUMMARY: THE FUTURE OF COMMUNICATION RESEARCH

In my analysis, I have focused on the social role and method of conducting Internet research, which aims to understand the medium and large-scale communication process. Although Edward Snowden, who was cited at the beginning, shows the dark possibilities of certain data analysis techniques, the remedy for these threats seems to be increasing number of large-scale research, thanks to which we will understand better how various, more complex elements of the Internet communication affect society. This will allow not only to understand the social process itself, but also to catch situations in which there are attempts to covertly influence public opinion, which we know little about today. That is why I described the theoretical and practical possibilities of a specific Internet research group, defined as Mass Automated Internet Analysis.

The need for a new type of research is also strongly associated with changes in society itself, leading to much greater fragmentation. This was noticed by many researchers, describing, for example, the new "long tail" model – selling a large number of very diverse goods instead of mass selling typical products [Anderson, 2006]. Although Anderson mainly described sales models, the same disproportion appears in Internet communication, forcing to reach for many sources instead of a dominant one.

For this reason, one must ask about the challenges faced by media and social communication researchers. The appearance of the Internet has caused a rapid increase in the volume of information available, and more and more examples confirm that the ability to analyse large data sets is a key development factor today [Mayer-Schönberger, Sugar, 2014]. Large-scale methods of data analysis are the daily bread of modern computer scientists and statisticians, but this is due to the simple fact that these issues appear already at the stage of study. In study programs in the field of media and social communication, subjects related to computer science, statistics and automatic data analysis appear relatively rarely, if at all. In addition, in many media studies, mass research results are presented from external tools, but they are rarely the result of independent data collection and processing.

Does this mean that research on social communication ceases to be right? It seems the opposite – there is a huge social demand for high-quality non-commercial research explaining contemporary phenomena related to Internet communication. As I mentioned at the beginning, we are not able to explain fully many cases of online hate, fake news mechanism, attempts to influence public opinion. The reason for these difficulties is the lack of data, insufficient number or too slow analysis. That is why in research in the field of communication it is so important to adapt to the scale of currently processed data in order to ensure appropriate quality results, and at the same time relieve researchers from performing mechanical activities. As noted by M. Szpunar [2018] there is a need to harden the social science, including, among others, bringing media research closer to computer science.

Finally, it should be noted that many of the research methods described here, although they may seem innovative, fall within the framework described by Berelson [1952] over half a century ago. He pointed out that the content testing process should be repeatable, that a scheme for dividing content into units of analysis and a scheme for classifying these units should be developed. A new element in current research is the use of computers for mass data collection and processing, which, however, in many cases does not so much change the formal way of conducting the research as it reduces the cost, increases the scale and reduces the time needed for analysis. The use of software tools is not due to fashion, but to necessity. Almost every modern data set is large, communication on the internet is fragmented, and research results will have social significance and improve transparency if delivered really quickly.

References

ANDERSON CH. (2006), The Long Tail: Why the Future of Business Is Selling Less of More, Hyperion.

BABBIE E. (2005), The Basics of Social Research, Wadsworth.

BERELSON B. (1952), Content Analysis in Communication Research, Free Press.

BRIN S., PAGE L. (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine, [in:] Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

CASTELLS, M. (2000), The Rise of the Network Society, Cambridge.

CROUCH C. (2004), Post-Democracy, Polity Press.

DAVIES H. (2015), Ted Cruz campaign using firm that harvested data on millions of unwitting Facebook users, „The Guardian", 2015-12-11, https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

DOBOSZ K. (2012), Przeszukiwanie zasobów Internetu, Warszawa, PJATK.

GAŁCZYŃSKA M. (2019), Śledztwo Onetu. Farma trolli w Ministerstwie Sprawiedliwości, czyli „za czynienie dobra nie wsadzamy", https://wiadomosci.onet.pl/tylko-w-onecie/sledztwo-onetu-farma-trolli-w-ministerstwie-sprawiedliwosci-czyli-za-czynienie-dobra/j6hwp7f

GATES B., HEMINGWAY C. (2000), Business at The Speed of Thought: Succeeding in the Digital Economy, Penguin.

GOGOŁEK W., JARUGA D. (2016), Z badań nad systemem rafinacji sieciowej Identyfikacja sentymentów, „Media Studies" 4 (67) 2016, http://studiamedioznawcze.pl/article.php?date=2016_4_67&content=gogolek&lang=pl.

GRUSZCZYŃSKI W., OGRODNICZUK M. (2016), JASNOPIS czyli mierzenie zrozumiałości polskich tekstów użytkowych, Warszawa.

GUNNING R. (1952), The Technique of Clear Writing. McGraw-Hill.

KASPERSKI M., BOGUSKA-TORBICZ A. (2008), Projektowanie stron WWW. Użyteczność w praktyce, Gliwice, Helion.

KAUSHIK A. (2009), Web Analytics 2.0: The Art of Online Accountability and Science of Customer Centricity, Sybex.

KREJTZ K. (2012), W jaki sposób badać kulturę wypowiedzi w internecie, [in:] Krejtz K. (ed.), Internetowa kultura obrażania?, Warszawa, SWPS.

MAYER-SCHÖNBERGER V., CUKIER K. (2014), Big Data: A Revolution That Will Transform How We Live, Work, and Think.

PŁANETA P. (2018), Komputerowa analiza tekstu w dyskursach medialnych, [in:] Szymańska A., Lisowska-Magdziarz M., Hess A. (ed.), Metody badań medioznawczych i ich zastosowanie, Kraków.

ROBAK M. (2017), Lęk przed ciasteczkami a realna ochrona prywatności, [in:]Szetela M., Kaleta M., Piech K., PIECH M. (ed.), Zaplątani w sieci. Społeczeństwo wobec wyzwań nowych mediów, Toruń.

ROBAK M. (2017b), Analityka internetowa i jej potencjał w trzecim sektorze, „Kultura – Media – Teologia", 2017 (30) vol. 7, pp. 52-69.

ROBAK M. (2018), Ograniczenia badań aplikacji wobec batalii o prywatność, [in:] Gackowski T., Brylska K., Patera M. (ed.), Komunikowanie w świecie aplikacji, Warszawa.

ROBAK M. (2018b), Analiza porównawcza stron diecezji w Polsce metodą WebScan, [in:] Olędzki J. (ed.), Teologia środków społecznego przekazu w naukach o mediach, Warszawa.

SNOWDEN E. (2019), Pamięć nieulotna, Kraków.

SZPUNAR M. (2018), Nowe media – nowe metody badawcze?, [in:] Szymańska A., Lisowska-Magdziarz M., Hess A. (ed.), Metody badań medioznawczych i ich zastosowanie, Kraków.

WASSERMAN S., FAUST K. (1994), Social Network Analysis. Methods and Applications, Cambridge.

WIMMER R., DOMINICK J. (2006). Mass Media Research. An Introduction, Wadsworth.