Włodzimierz Gogołek, Paweł Kuczma

Refining Network Information on the Example of Parliamentary Elections. Part 1. Blogs, Forums, Sentimental Analysis

KEY WORDS
Internet, Big Data, social networking, sentiments, blogs, forums, refining network information, new sources of journalistic information, parliamentary elections 2011.

ABSTRACT
Over 90 per cent of information generated in 2012 was registered in the form of a digital recording. Sources of this scale are described as Big Data. An analysis of this data creates new sources of valuable information. The process of their retrieval – mainly from social network services – was termed refining Big Data. A confirmation of the usefulness of its application is research results on the support for certain political parties participating in parliamentary elections in Poland in 2011. The accepted methodology and results from quantitative research prove that the refining of records obtained from social networks can be a reliable source of information on the state and changes of political sympathies in the period before elections.

The use of Internet resources, especially social networks and traditional forms of online distribution of media information, is becoming an important source of information for social research, in particular journalism. This potential is derived from the communication power of the Internet and the strength of both information and service resources. Already in 2010, for the first time the total amount of digital information produced in the world in one year exceeded one zeta byte (10^{21}). Resources on this scale, known as Big Data – i.e. huge unstructured data warehouses – exceeded the critical size. They have created a new dimension to the value and attractiveness of information resources for all types of research, including those related to social research. The critical size means a faint usefulness of conventional tools for the analysis of such large databases. This justifies the commencement of work on the exploration/ expert analysis of Big Data. The results obtained from the analysis of Big Data form previously unavailable data sources, the creation of which can be seen as a new phase in the development of IT applications (tools and digital networks for exchange of information)[1].

[1] *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation [access from: 22.04.2013].

Skilful analysis of Big Data enables more precise, and in due time, provision of the necessary, critical, or even credible-looking information[2]. This information will allow for improvement and development of new generations of products and services used by the media.

A significant part of Big Data is formed by the Internet resources, including social networks. Data of this type are created by and about individual users of social networking (blogs, posts, portals, e-mails or stream of queries to the Internet), professional publications, and other rich information resources.[3] The interesting part of Big Data are resources hidden in the network– Deep Web (Dark Net) and pNet, built on P2P – called F2F (friend to friend), such as: Freenet. These resources are a thousand times greater than those available in traditional, indexed by the search engines on the web[4].

It is assumed that the resources collected in Big Data constitute the primary source of information, and the result of their analysis secondary information. The process of this analysis is referred to as the refining of network information (refining).

Refining

One of the well-established pillars of refining is culturomics, which is 'a form of computational lexicology examining human behaviours and cultural trends through quantitative analysis of digitized texts. Scientists explore huge *data mining*[5] in order to explore cultural phenomenon through their impact on language and the way words are used'[6]. Using the *culturomics* tools efficiently identified important changes, including: culture, science and history. Refining enables the perception of original information in hidden resources– i.e. secondary information. It is like a microscope that enables a fuller view and measurement of things – at the level of both individuals as well as social groups. It is a kind

[2] D. Copeland, *Harvard Researcher Uses Social Media To Predict Stock Market Volume*, http://readwrite.com/2012/02/08/harvard_researcher_uses_social_media_to_predict_st [access from: 20.04.2013].

[3] S. Stephens-Davidowitz, *Google's Crystal Ball*, http://campaignstops.blogs.nytimes.com/2012/10/20/googles-crystal-ball/ [access from: 20.04.2013].

[4] W. Boswell, *Five Search Engines You Can Use to Search the Deep Web*, http://websearch.about.com/od/invisibleweb/tp/deep-web-search-engines.htm [dostęp: 31.03.2012].

[5] *Data mining* – „'exploration of data (it can be also determined as knowledge acquisition, data extraction) – one of the stages in the process of discovery of knowledge in databases (English: 'Knowledge Discovery in Databases, KDD). The idea of *data mining* means the use of the speed of your computer in order to find the accuracy in the information stored in data warehouses (just because of the limited time), hidden to man'. http://pl.wikipedia.org/wiki/Eksploracja_danych [access from: 20.04.2013].

[6] Term: *culturomics* was used firstly by the end of 2010 by researchers from the Harvard University, namely: Jean-Baptiste Michel and Erez Lieberman Aiden in the article titled: *Quantitative Analysis Of Culture Using Millions Of Digitized Books*, „Science" Vol. 331 (2011), no. 6014, p. 176–182, www.sciencemag.org/content/331/6014/176 [access from: 1.06.2011].

of revolution in measurements. The data, obtained through such measurements, form a picture of the needs and behaviours of individual users, and also the community as a whole.

The following tools may be directly used for refining of network resources: Attentio, Radian6, Sysomos, NetBase, Collective Intellect, Alterian, and Google Alerts. Network refining can be effectively carried out using Attentio Brand Dashboard[7]. This is demonstrated by the results of dynamics of information image changes of candidates in the presidential election of 2010.[8] Another professional refining tool is Summary of World Broadcasts (SWB) – a network service that monitors information services. It allows the monitoring of full texts and summaries of newspaper articles, conference proceedings, and television and radio materials as well as other non-classified technical reports (grey literature) in 130 languages[9].

Purpose and scope of the study

Taking into account the potential of Big Data, repeated demand for current information related to the election on a national scale, the Institute of Journalism at the University of Warsaw, as part of research work, assumed the object and purpose of research illustrating the potential of network refining to be the identification and verification of information processing tools, to assess current electoral preferences before the parliamentary elections in Poland in 2011.

A parallel objective of the study was to outline the methodology which is the key element of network refining. This methodology was used to look for these ratings of electoral preferences on the basis of information obtained from the network.

Achieving the goals has helped identify a way of creation of a meaningful data source, supporting the diagnosis of the condition and dynamics of changes of information image of activities of the electoral committees (political parties) taking part in the elections. This knowledge can be a valuable source of information about the election campaign for the media, interested individuals and social groups.

---

[7] *Attentio Brand Dashboard – monitoring mediów społecznościowych,*[Social Media Monitoring] www.blog.mediafun.pl/attentio-brand-dashboard-monitoring-mediow-spolecznosciowych/ [access from: 20.04.2013], see also a websity of Attentio – http://attentio.com/.

[8] P. Kuczma, W. Gogołek, *Informacyjny potencjał sieci – na przykładzie wyborów prezydenckich 2010* [The Informational Potential of the Internet – Illustrated with the Example of Presidential eEection of 2010], „Studia Medioznawcze" 2010, no. 4, p. 35–49.

[9] K.H. Leetaru, *Culturomics 2.0: Forecasting large-scale human behavior using global news media in time and space*, „First Monday" Vol. 16 (2011), no. 9, www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040 [access from: 20.04.2013].

A similar study was conducted in 2010 on the occasion of the presidential election[10]. The results fully confirmed the continuing validity of the research path, based on network information refining[11].

The following hypothesis was adopted: refining of the network allows current, reliable monitoring of variables, describing the voting preferences of Poles in the run-up to parliamentary elections in 2011.

This hypothesis is equivalent to saying that the content of the network, especially in social media, is a reflection of the real attitudes of users and may presage their real actions, such as: voting for the candidate, the party, the choice of a specific response in a referendum. Thus, there is a statistical relationship between quantitative measures of the content created in the network and political preferences the result of which is the choice of a particular political option.

The study classified the campaign committees associated with parties/ political circles, whose members sat in the RP Parliament on 1 January 2011 (including the newly established political structures present in the Parliament, associated with Mr Janusz Palikot and Ms Joanna Kluzik-Rostkowska), namely: Platforma Obywatelska (PO), Polskie Stronnictwo Ludowe (PSL), Prawo i Sprawiedliwość (PiS), Sojusz Lewicy Demokratycznej (SLD), Polska Jest Najważniejsza (PJN)[12] and Ruch Palikota (RP)[13].

In order to quantitatively assess the number of times party names appeared in texts published in the Internet, the relevant contexts were distinguished[14]. These included the keys (phrases/ words) associated with the government and its functions as well as powers of the various ministries, accepted as factual contexts: 1) education, 2) finance, 3) economy, 4) infrastructure, 5) culture, 6) science and higher education, 7) defence, 8) labour and social policy, 9) agriculture, 10) regional development, 11) Treasury, 12) sport and tourism, 13) justice, 14) internal affairs and administration, 15) foreign affairs, 16) environment and 17) health. Words describing the expertise of each of the ministries are based on the competences of ministries written in their statutes[15].

---

[10] P. Kuczma, W. Gogołek, *Informacyjny potencjał...* [The Informational Potential...]

[11] Analysis and conclusions included in the text were developed by W. Gogołek on the basis of the source data collected, verified and properly processed by P. Kuczma.

[12] Polska Jest Najważniejsza was registered as a political party on 17 March 2011.

[13] Previously, Ruch Poparcia, and as a political party called Ruch Palikota was registered on 1 June 2011.

[14] On the basis of the structure of the Council of Ministers for: Order of the President of the Republic of Poland from 16 November 2007, no. 1131-50-07 on the appointment of the Council of Ministers, M.P. 2007, No. 87, item. 947.

[15] Table of statutes, see: www.id.uw.edu.pl/zasoby/profile/59/Aneks_nr_2-Wykaz_statutow_ministerstw.pdf [access from: 23.04.2013].

The second group of contexts – i.e. media contexts – are those related to current events published in the media. They are chosen through a formal analysis of the content of press releases (using the QDA Miner v3.2 and the WordStat 6.0.1. program)[16] out of the largest Polish opinion-forming heavies (in the electronic version) [17] with a different political profile: 'Gazeta Wyborcza' and 'Rzeczpospolita'. For this analysis, the electronic version of heavies, available via the Factiva search was used[18]. Articles came from the period of 1-28 February 2011, i.e. the month prior to the commencement of the relevant studies. All articles, including titles, were analysed quantitatively. This resulted in a list of 39153 words. One thousand (1000), most frequently recurring and statistically significant words – at least 32 times in all the analysed articles were chosen[19]. Since some of the analysed words recurred (e.g. in different cases or synonymous words appeared in the analysed set), eight groups/ sets of words were distinguished, the so-called groups of contexts. As a result of this analysis, the following groups of media contexts were selected: 1) EU (European Union) – including words such as: union, EU, European, Presidency, Europe; 2) disaster (Smolensk) – disaster, in Smolensk, Russia, MAK, tragedy, Tupolev, attack; 3) power – government, management, parliament, leader, president;
4) the media – media, newspaper, TVP, TVN, television; 5) money – money, finance, budget, NBP; 6) reforms – reforms; 7) church - church; 8) law – prosecutor's office, legislation, court, tribunal, etc.

The media contexts obtained this way have been used to analyse the substantive nature of the election campaign in 2011. Its goal was to answer the question of whether in the content available online, more intensive in terms of quantity, there are represented factual contexts or media contexts[20].

In the course of the study, the basic content published in social media (forums, blogs, Facebook, Twitter) was analysed, where the content is created by the users, and in

[16] Programs available at: www.provalisresearch.com/Download/download.html [access from: 31.05.2010]. One used the test version.
[17] *„Gazeta Wyborcza" i „Fakt" to najchętniej czytane dzienniki* [„Gazeta Wyborcza" and „Fakt" are the most popular dailies], www.wirtualnemedia.pl/artykul/gazeta-wyborcza-i-fakt-to-najchetniej-czytane-dzienniki# [access from: 24.05.2010].
[18] https://han.buw.uw.edu.pl/han/ISIEM/site.securities.com/search/pub_search.html?pc=PL&sv=EMIS [access from: maj 2010].
[19] Because the word at the 1000 place had the frequency of occurrence of 32, the analysis included all the words with a frequency of at least 32. There were a total of 1016.
[20] The results of these studies will be published in the second part of the research. *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych. Konteksty medialne i merytoryczne*. [The Rafination of the Internet Information Based on the Example of Parliamentary Elections. Media and Content-related Contexts]

information services, created by professional editors. It is assumed that a single entry, record, part of a blog, downloaded from the Internet for further analysis will be referred to as 'entry'.

Refining of data from network resources was carried out on collections published from March 1 to October 17, 2011. Monitoring, archiving and preliminary analysis of contextual content published in the Internet were performed using the Attentio Brand Dashboard tool. The data has been highlighted by key words (in this case – contexts) describing the examined political parties[21].

The study was subjected to the following indicators derived from the analysis of online content: 1) the amount of content on a given party; 2) trends/ dynamics of the change of the content amount[22], 3) qualitative assessment, that is: results of the context analysis, in which content on the party appear and the colouring of the content, i.e. sentiment – a quantitative distribution of content on the positive and negative.

Index of amount of content is based on the total number of entries/information in the files collected by the Attention Brands Dashboard, taken from online sources concerning the party and contexts. Entries included content obtained from forums, blogs, Facebook, tweets and newspaper articles.

Assessment of the dynamics of changes and trends in the content/ search was performed on the basis of a number of entries concerning the party, depending on the context and sentiments in the analysed period.

Qualitative assesment was carried out in two categories. The first was a preliminary contextual analysis involving the grouping of the content into contexts, obtained by monitoring, on the basis of substantial and media list of contexts discussed above. The second was – carried out in parallel with the analysis of context – a preliminary analysis of sentiments understood as highlighting of entries that contained any party name and a word recognized as 'sentiment'. Due to the importance of the sentiment analysis, it is a separate part of the whole study. Highlighting the words deemed 'sentiments', due to the absence of an authoritative list of such Polish words, was made on the basis of a list of expressions emotionally saturated, the so-called ANEW 2012[23]. Out of the 1031 words of this collection– extremely positive and extremely negative ones were selected, including those that appeared

[21] The definitione of the monitored words is included in the Annex published on the website: www.id.uw.edu.pl/zasoby/profile/59/Aneks_nr_1-definicja_wyszukiwania.pdf.
[22] This term is understood as the relationship between the amount of content on each candidate during the period.
[23] M.M. Bradley, P.J. Lang, *Affective Norms for English Words (ANEW)*. List of all words is not publicly available. It was obtained at the special request of the authors from its creators from The Center for the Study of Emotion and Attention from the University of Florida, http://csea.phhp.ufl.edu/media/anewmessage.html [access from: 2.04.2013].

most frequently. These words were then translated into Polish[24], extending their meaning - if necessary, for example, while translating the word 'love' both forms were used, i.e. – 'love' – noun, and 'to love' – verb, etc. Words used in the analysis are shown in Table 1.

Table 1: List of positive and negative words with the frequency of their occurrence in the entries.

Words	Frequency
Positive word	
sure	930
win, prize	869
target	840
car	769
enthusiast, enthusiasm	573
surprise, amaze	500
profit, gain	407
Success	397
opportunity	341
party, celebration	312
fun, toy	311
love, to love	306
to like	299
rich, wealth	297
Negative word	
blame, guilty, guilt	616
war	516
with difficulty, difficult	396
death	345
to destroy, damages	344
disaster	308
worst, to worse	279
accident	275
bankrupt, bankruptcy, collapse, insolvency, fail, fail	247

Filtration of entries associated with a particular party

The first stage of filtration

Due to the fact that the monitoring of acquired online content brought above-average results, which show the advantage of Platforma Obywatelska in the number of entries awarded, an additional analysis consisting of entries obtained was conducted, in order to increase the

[24] The translation of the ANEW set was made using the Google Translate (http://translate.google.com/).

accuracy of assigning entries to the appropriate party. For this purpose, a software was created[25], which consists of two modules. The first converts the data gathered from the monitoring of online sources using the Attentio application. Conversion is used to prepare a file for the verification phase in the second module, and does not affect the content of the file to be converted. The second module verifies whether the contents and title of an entry include references to the party, to which an entry has been assigned. For this purpose, appropriate standards of variations/ varieties (resulting, among others, from Polish grammar) of the names of the party were applied.

The second stage of filtration

Filtration of the first stage did not provide satisfactory results, especially with regards to the entries for Platforma Obywatelska (PO) – a large amount of entries attributed to the party did not apply to it, because of the repeated occurrence of the preposition 'po'. This brought about the need to launch the second stage of content verification.

To provide a precise automatic selection of entries concerning the distinguished parties – e.g. the abbreviation of the PO party and distinguishing it from the preposition 'po', an appropriate program was developed that analyses the input data for the common occurrence of designated words and a set of words defining them [26]. The occurrences of these words were counted, together with the words designated in the text boxes, containing relevant data from the analytical point of view.

To analyse the occurrence of needed words, the software uses regular expressions. The input data are subjected to filtration prior to analysis. Filtration involves the removal of formal character sequences that were included in the data set, obtained after the initial context analysis (html tags, html entities, and 'script' sections), replacement of multiple whitespace characters with a single space, replacement of multiple punctuation marks with single and removal of extra spaces before punctuation marks.

The program counts the common occurrence of each standard of the party name with each standard determining sentiment/ emotion (contexts were defined at the stage of obtaining the content from the Internet). Additionally, some character limits within which the party standard and sentiment must belong, for the pair to be counted was specified. The adopted number of 30 characters is half the average number of characters of a sentence in Polish. Surveys have shown that this value is optimal and can thus be considered as highlighted

[25] The author of this program is Mr Marcin Łączyński.
[26] The author of this program is Mr Piotr Celiński, pgc@post.pl.

entries associated with the party and adopted model (context, sentiment). Standards of contexts and sentiments are counted for both the left-hand and right-hand presentations relative to the standard of the party name.

Standards of party names used in social media were determined on the basis of the most common (including colloquial and pejorative) definitions of parties appearing in social media. For this purpose, a file with the standards for party names was created.

The definition of a party consists of a name, which will be included in the output file, as well as a standard of a regular expression, describing the possible names (and their variants and variations) occurring in social media.

The definition of a party includes its acronym, i.e.: PiS, PJN, PO, PSL, RP, SLD, with uppercase and lowercase letters[27] as well as lots of variety names of a party, i.e.: for Polskie Stronnictwo Ludowe the definition includes the following keywords: PSL, psl, PsL, Psl, pSl, psL, Polskie Stronnictwo Ludowe and declension according to the Polish language, i.e. polskie stronnictwo ludowe, polskiego stronnictwa ludowego, polskiemu stronnictwu ludowemu, polskim stronnictwem ludowym, polskim stronnictwie ludowym, and these present in all cases of the plural form of words which determine the members of the party in the media, i.e.: ludowcy, ludowców, ludowcom, ludowcach, ludowcami (also in upper or lowercase letters). Analogous definitions have been created for the other parties.

Input data

The input data to the first stage of refining included all those available and related to political parties during the period between March and September 2011, information resources of forums, blogs, Facebook, Twitter and network information portals[28].

Using the Attentio tool, in the first stage from the social portals 1 418 267 entries were acquired, of which 565 868 came from forums, 754 850 from blogs, 23 883 from Facebook and 73 667 from Twitter. In the next stage, after the first stage of filtration of the content there was 339 403 entries, of which 175 356 came from forums, 162 527 from blogs, 197 from Facebook and 1323 from Twitter.

[27] The exception is the acronym 'PO' coinciding with the preposition 'po', as already explained.
[28] Number of entries obtained from some online newspapers in the first stage of filtration exceeded 550.000, and in the next one – 100.000. Detailed data obtained from online newspapers and the results of their analysis will be presented in the second part of the study in a separate article.

A B

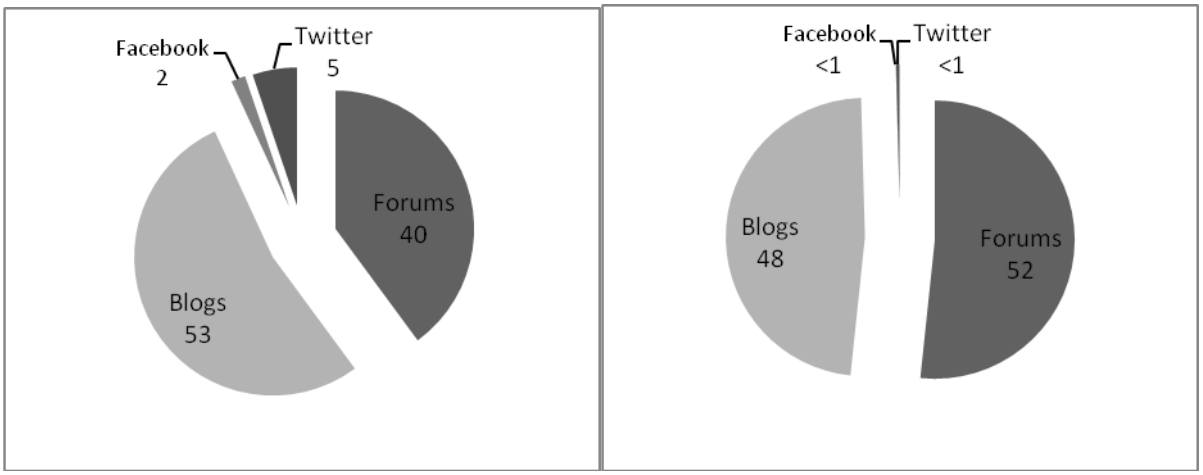

Figure 1: Distribution of the number of entries of the social networking sites; A – before filtration and B – after filtration (per cent).

Source: Own calculations.

Due to the relatively small numbers of entries on Facebook and Twitter, the data obtained from these sources are not subject to further analysis.

Due to the relatively high volumes of source data acquired for the calculation of awarded entries – in the first stage of refining – it was necessary to use a computer of high computing capability. For the analysis of the entries, the IBM Boreasz - IBM Power 775, available under the Calculation Program of the Great Challenges of Science and Technology (POWIEW) and servers of the Institute of Journalism at the University of Warsaw among others were used[29].

A significant element of the procedure for the use of source data for further research was to identify the independent variables. They provided a reference point for assessing the reliability of the results of refining. It was assumed that these variables consist of:

– costs incurred by the parties for the election campaign (Table 2),

– number of votes cast for each party (Table 3),

– results of the CBOS opinion polls (Table 4).

[29] Calculations were performed at the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) at the University of Warsaw under Grant No. G51-14.

Table 2: Campaign costs (in PLN)

	PO	PiS	RP	SLD	PSL	PJN
Creating and maintaining a website	60 709,57	924	2 076,98	108 100,52	41 572,90	20
Using the media and poster carriers	15 881 840,10	16 224 914	572 172,52	14 334 58,14	5 018 560,83	647 485,71
Advertising on the Internet (cost of the service)	2 206 623,42	1 170 403	52 438,76	1 144 647,58	321 025,43	31 284,25
Share of expenditure on the Internet in total/ media spending (per cent)	13,89	7,21	9,16	7,99	6,4	4,83
Implementing the election materials, including: conceptual, design works and manufacture	6 195 905,01	7 477 332	854 201,60	6 694 452,58	4 926 782,56	710 929,75
Advertising on the Internet	531 144,33	344 725	29 892,90	47 859,51	28 918,65	27 727,19
Share of expenditures on the Internet/ in total on the performances (per cent)	8,57	4,61	3,5	0,71	0,59	3,9
In total	22 138 455	23 703 170	1 428 451	21 137 211	9 986 916	1 358 435
Total spending on the Internet	2 798 477	1 516 052	84 409	1 300 608	391 517	59 031
Share of expenditures on the Internet/ total costs (per cent)	12,64	6,4	5,91	6,15	3,92	4,35

Source: National Electoral Commission Communication of 13 February 2012 on the financial statements of the electoral committees participating in the elections to the Polish Sejm and Senate of the Republic of Poland, held on 9 October 2011 (submitted for publication in Monitor Polski): http://pkw.gov.pl/wybory-do-sejmu-rp-i-do-senatu-rp-2011/komunikat-panstwowej-komisji-wyborczej-z-dnia-13-lutego-2012-r.html [access from: 12.11.2012].

Table 3: Total votes in the parliamentary elections of 2011 – election results

	PO	PiS	RP	SLD	PSL	PJN
Number of votes	5 629 773	4 295 016	1 439 490	1 184 303	1 201 628	315 393

Number of votes in per cent	39,18	29,89	10,02	8,24	8,36	2,19

Source: Announcement of the National Electoral Commission of 11 October 2011 on the results of the elections to the Polish Sejm, held on 9 October 2001 (Journal of Laws 2011, No. 218, item. 1294): http://pkw.gov.pl/g2/2011_10/0a409df0d614d3d42f854615f3ab6286.pdf [access from: 12.11.2012].

Table 4: The results of pre-election polls of 2011 conducted by CBOS (per cent)

Month	**PO**	**PiS**	**RP**	**SLD**	**PSL**	**PJN**
March	35	18	1	16	4	3
April	31	23	1	12	5	2
May	37	21	1	12	4	1
June	34	22	1	11	5	3
July	38	17	1	9	4	0
August	36	20	1	8	4	0
September	37	20	2	7	6	1
October	34	20	7	9	6	1

Source: *Parties' preferences before the election*, www.cbos.pl/SPISKOM.POL/2011/K_124_11.PDF [access from: 12.11.2012].

Quantitative analysis

Setting about the quantitative analysis of the results, the Pearson correlation – r of the number of votes obtained by all parties in the CBOS survey results were calculated. It amounts to 0.96 ($p<0.001$). This constitutes the accepted form for assessing the credibility of opinion polls conducted by CBOS and justifies the assumption that the CBOS results (in particular months) shall be a reference point for further research.

Due to dominant numbers of positive and negative entries (sentiments) concerning PO and PIS together (forums – 62 per cent; blogs – 67 per cent), further analysis of the power of information refining was performed on the example of these two parties (see Figure 2).

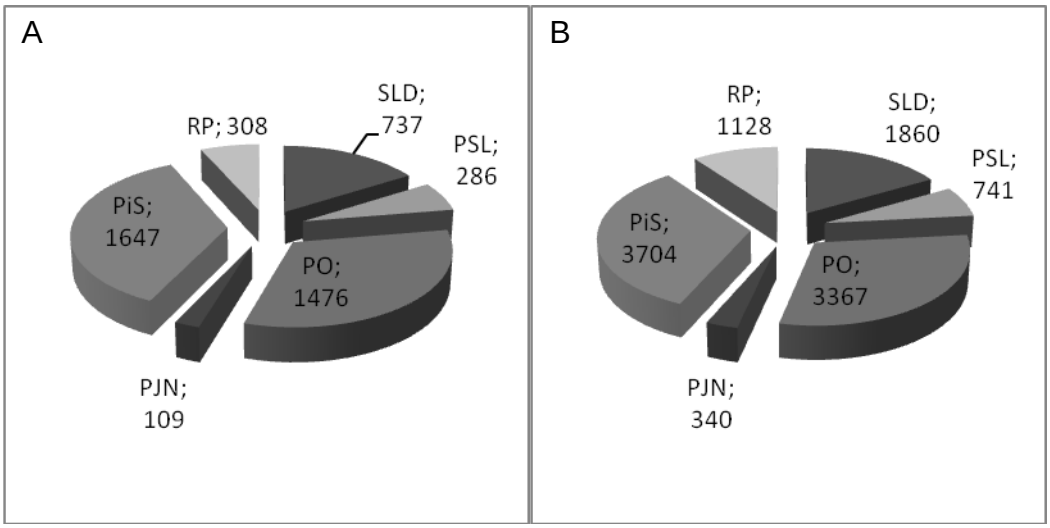

Figure 2: Distribution of sentiments – positive and negative entries – (March – October) on forums (A) and blogs (B) for all parties (in per cent)

Source: Own calculations.

The correlation of r; the number of votes obtained by the parties with a number of positive entries on the blogs amounts to 0.93 (p>0,001) – see Figure 5. This proves an almost certain dependence of the results obtained from refining on the real results of voting. At the same time, it shows the justification for in-depth analysis of the indicated regularity – a significant convergence of refining results with the official results of voting.

Forums

In the third stage of refining, a preliminary quantitative analysis of the results obtained was conducted – mainly on the basis of their visualization. Besides the direct results obtained from refining, an additional form of visualization are curves, illustrating the trend of changes in the number of sentiment – positive and negative entries – for the two leading parties in the election, i.e.: PIS and PO.

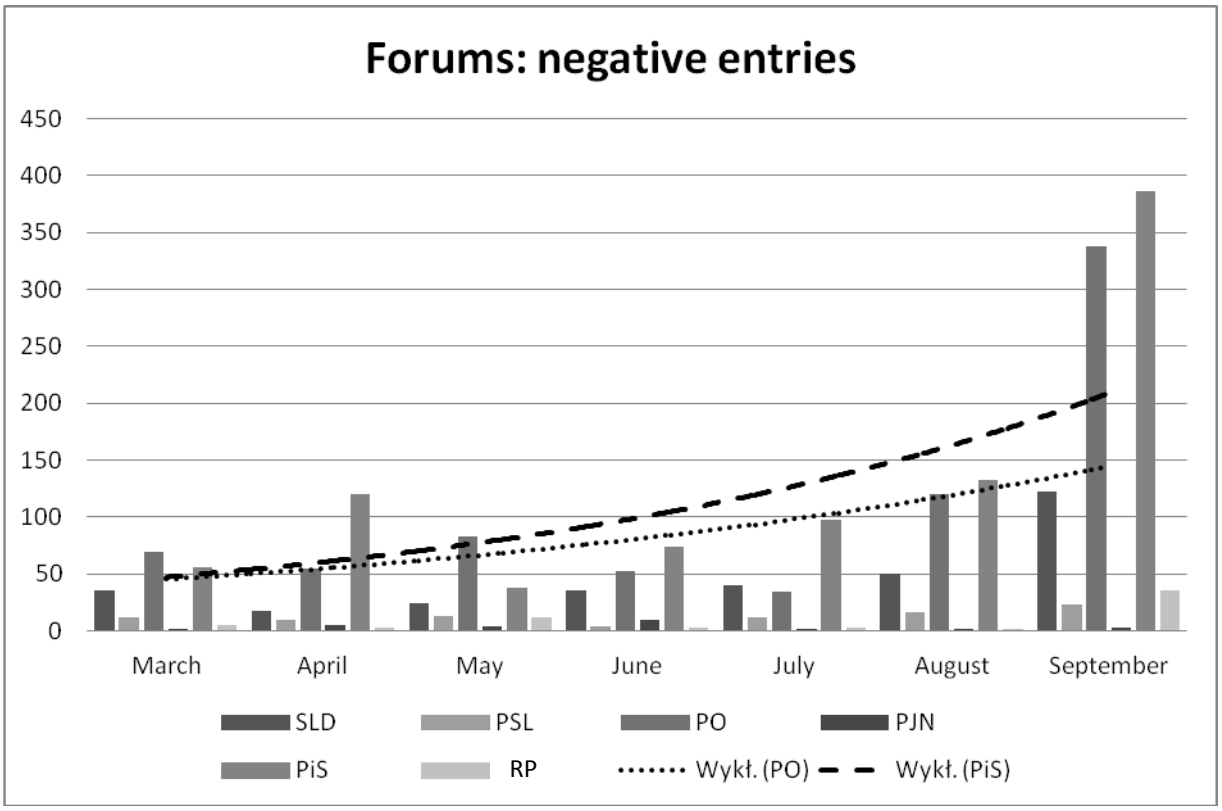

Figure 3: Distribution of the number of negative entries (sentiments) on the forums and exponential trend curves of changes in these numbers for PO and PIS

Source: Own calculations.

Thanks to this, apparent regularity/ dependence of variables obtained from refining of entries were distinguished. Number of sentiments – positive and negative entries – on forums, pointed out the desirability of calculating the difference between the numbers of these entries. A clear predictive power of the final results of the elections on the basis of the refining results is shown in Figure 4.

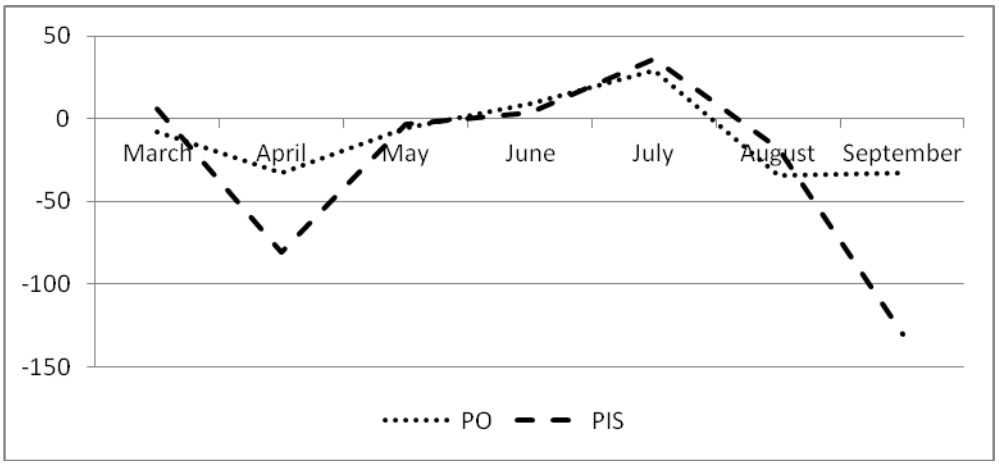

Figure 4: Distribution of the differences between the number of positive entries and the number of negative entries (sentiments) on the forums for PO and PIS

Source: Own calculations. The figure shows the balance between negative and positive entries, positive values reflect the advantage of positive entries.

The importance of information obtained, through refining, is proved by the quantitative measures of dependence between research variables: independent variables (source data) with the results of refining (dependent variables).

The first stage included the calculation of the correlation of quantitative results of the official (CBOS) polls with the refining results. An example of this is a correlation between the number of positive and negative entries (sentiments) on the forums with the CBOS poll results for all electoral committees in the subsequent months – from March to October 2011. The values of these coefficients are statistically significant ($p<0.001$), with the exception of October (due to the date of the election, it did not include data from the entire month). This shows a statistically significant convergence of the poll results with the refining results.

Table 5: Correlation of r - the relationship of the CBOS poll results with the number of positive and negative entries (sentiments) on the forums for all the parties in the subsequent months

Month	**r for positive entries**	**r for negative entries**
March	0,70	0,95
April	0,81	0,76
May	0,83	0,98
June	0,89	0,84
July	0,55	0,46
August	0,88	0,93
September	0,90	0,87

Source: Own calculations.

It also showed a significant association between the numbers of positive entries (sentiments) on the forums with the number of votes obtained by the parties (Figure 5).

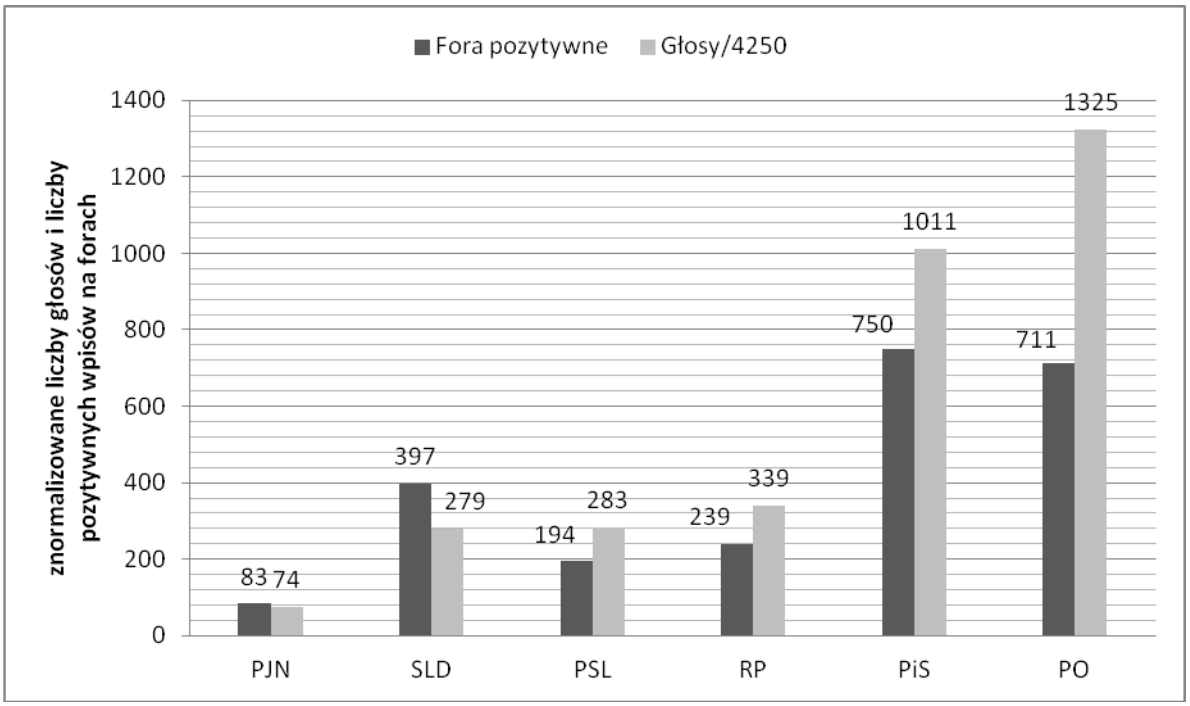

Figure 5: Illustration of similarity ratio (un-certain size) of the positive entries (sentiments) on the forums with the number of votes obtained

Source: Own calculations. In order to improve the effect of data visualization, input data was normalized (the number of votes were reduced 4250 times).

The correlation coefficient of positive entries on the forums with the number of votes is 0.93 ($p<0.001$).

Table 6: The number of positive and negative entries (sentiments) on the forums

Month	SLD		PSL		PO		PJN		PiS		RP	
	positive	negative	positive	negative	positive	negative	positive	negative	positive	negative	positive	negative
March	32	35	52	12	61	69	4	1	62	56	1	5
April	8	18	3	9	22	55	2	5	39	120	5	3
May	68	24	4	13	77	83	2	4	35	38	5	12
June	23	35	12	4	61	52	7	9	78	74	3	3
July	36	40	23	12	63	34	10	2	134	98	16	3
August	51	57	49	16	97	131	12	2	107	125	19	1
September	146	122	47	23	305	338	44	3	256	386	133	36

October	33	9	4	3	25	3	2	0	39	0	57	6
In total	397	340	194	92	711	765	83	26	750	897	239	69

Source: Own calculations.

Indicated relationships: positive entries from polls and votes prove the expected reliability of the information obtained from refining and confirm the assumed hypothesis that refining of network allows continuous monitoring of changing political preferences.

Blogs

Blogs had a slightly higher value of information, in terms of the number of votes obtained by the parties from the forums.

Table 7: The number of positive and negative entries (sentiments) on blogs

Month	**SLD**		**PSL**		**PO**		**PJN**		**PiS**		**RP**	
	positive	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**	**positive**	**negative**
March	72	74	45	44	157	173	41	23	129	230	19	16
April	121	61	71	20	202	143	26	23	144	248	14	3
May	152	53	37	4	295	171	3	32	152	218	12	5
June	152	210	13	30	304	362	78	35	255	365	31	4
July	58	71	52	17	226	78	2	12	347	277	12	2
August	90	107	96	33	183	168	5	43	215	238	13	40
September	90	57	86	1	219	87	4	12	215	267	194	201
October	222	270	87	105	254	345	1	0	214	190	393	169
In total	957	903	487	254	1840	1527	160	180	1671	2033	688	440

Source: Own calculations.

The value of information mentioned shows Figure 6 - ratios between positive entries on blogs and election results. The r correlation coefficient between the two variables was 0.95 ($p<0.001$). This proves an almost certain reliability of the information obtained from the refining of entries on blogs. It can be assumed that blogs are an important source of information on the number of votes obtained in the election.

Table: The correlation of r - the relationship of the CBOS poll results with the number of positive and negative entries (sentiments) on blogs for all parties

Month	r for positive entries	r for negative entries
March	0,95	0,78
April	0,97	0,84
May	0,98	0,84
June	0,95	0,93
July	0,72	0,42
August	0,84	0,79
September	0,67	0,24
October	0,38	0,81

Source: Own calculations.

One interesting issue is the identical value of r = 0.99 ($p<0.001$) the dependence of the number of positive entries on forums and blogs and the number of negative entries on forums and blogs. This proves the significant similarities of monthly distributions of numbers of positive and negative entries, which are the result of refining.

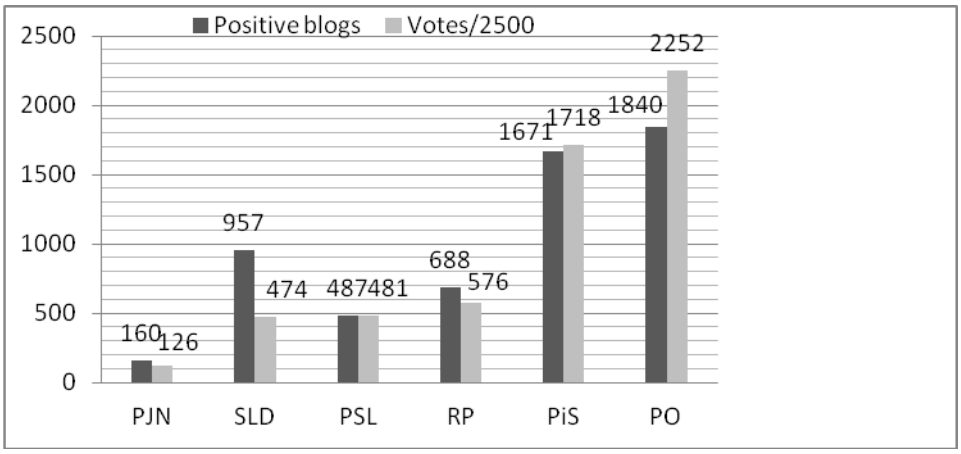

Figure 6: Similarities/ differences in ratio (un-certain size) of the positive entries on the blogs to the number of votes obtained

Source: Own calculations. In order to improve the effect of data visualization, the normalization of the input data was performed: the number of votes was reduced 2500 times.

The credibility of the statistics calculated

The value of the information of coefficients calculated is confirmed by their statistical significance. For this purpose, a study of the dependence between the CBOS poll results and the number of negative and positive entries (sentiments) on blogs for all parties was conducted. It was assumed, that these results can be generalized and they also account for the reliability of the results obtained from refining of entries on the forums.

Table 10: The value of Pearson's and Spearman's correlation of the dependence of the CBOS poll results with the numbers of negative and positive entries on the blogs for all the parties, together with the values of significance.

Entries	**Months**	**Pearson**		**Spearman**	
		correlation	**significance**	**correlation**	**significance**
Positive	March	0,948	1,000	1,000	1,000
	April	0,969	1,000	1,000	1,000
	May	0,976	1,000	0,971	1,000
	June	0,954	0,996	0,829	0,983
	July	0,720	0,926	0,943	0,999
	August	0,844	0,981	0,886	0,992
	September	0,672	0,954	0,829	0,983
	October	0,382	0,776	0,600	0,912
Negative	March	0,779	0,947	0,943	0,999
	April	0,836	0,965	0,886	0,992
	May	0,843	0,961	0,754	0,961
	June	0,935	0,996	0,886	0,992
	July	0,42	0,832	0,886	0,992
	August	0,785	0,951	0,714	0,949
	September	0,237	0,683	0,429	0,822
	October	0,806	0,986	0,943	0,999

Source: Calculations were performer by PhD. Piotr Pokarowski from the Institute of Applied Mathematics and Mechanics at the University of Warsaw.

Taking into account the relatively small number of testing/ data (Table 8), in order to verify the significance of the findings of the statistical analysis, the correlation value for all combinations of the month - the party (720 combinations) were calculated. It was expected that such incidental correlations will be significantly different from the essential results shown in Table 10. Column 3 of this table, contains the values of linear correlation (Pearson), while column 5 – the ranking correlation (Spearman correlation, i.e. the correlation calculated on the ranks, i.e. positions in the ranking of values). In contrast to the linear correlation, it is a

measure of resistance to scaling – i.e. it does not change after transition from the original data, for example to their logarithms. In the columns next to the correlation values, the probability (statistical significances) was entered, which shows that the observed correlation is greater than or equal to the correlation between all randomized (random) data (720 combinations), i.e. vectors: poll and blog, one of which has converted values. In other words, one can with a negligible statistical error speak of an almost full statistical dependence of both positive and negative content of entries (sentiments) on blogs with polls.

This regularity includes other results on the correlation value of entries with other variables (polls, number of votes, financial outlay). It can be considered as an indication of the desirability of further analysis of these types of relationships for a statistically reasonable prediction of e.g. parliamentary elections.

Expenditures for the campaign

Quantitative analysis of costs incurred by individual parties in the election period indicates a complete lack of relationship (r =0,06) between the cost of creation and maintenance of the website of the committee and the number of votes obtained. This proves an extremely small (meaningful) effectiveness of expenditure declared by the committees on the websites. A much larger dependence on the number of votes was found with other inputs.

Table 11: The correlation r costs of the campaign with the number of votes obtained

Costs of campaign	Pearson correlation coefficient
Costs of creation and maintenance of a committee website	0,06
Implementation of election materials, including conceptual works, design works and production	0,61
Sum of all expenditures	0,72
Use of the media and poster carriers	0,75
Share of expenditure on the Internet in total/ media spending	0,76
Advertisement on the Internet (service cost)	0,86
Expenditures on the Internet, in total	0,89
Internet Advertising	0,98

Source: Own calculations.

The data included in Table 8 show the extraordinary effectiveness of expenditures on the Internet advertising, much higher than the investment in the media, and poster carriers.

Summary

Refining of the Big Data resources enables a quantitative analysis – refining – of a wide range of original, unstructured information (in the discussed research results – more than 2 million entries). It has been proved that refining creates a new space of valuable sources of information and opens new avenues for research on the quest for information.

Such studies require changes in the methodology and the tools as well as significant computing power, used for exploration. This includes a collection of thematic information, their precise filtration (in this case about 90 per cent of entries were rejected) and a fundamental analysis of the information on sentiments and their statistical processing.

Thanks to data obtained through refining, the research results show the similarities and an almost identical support for specific political parties participating in the parliamentary elections of 2011, with the results of public opinion polls and the official results announced by NEC.

An in-depth analysis of the distribution of sentiments creates the possibility of predicting future changes in the estimated value of the data (see Figure 4). Further towards this goal – the creation of the function of prediction – requires the involvement of statistical tools and at the same time taking into account many parameters, such as: multiple regression function. These parameters include, among others, a reliable indication in the article of sentiments derived from entries on forums and blogs.

No less important than the substantive value of the results is the much lower cost of information obtained through refining, compared to the costs of acquisition of the same information in the traditional way – through surveys.

The results of this study show that Big Data cease to be a *terra incognito* for the social sciences. An important challenge is to improve the methodology for refining and developing appropriate tools for refining network information, user-friendly form of delivering results, and most importantly – the belief in the usefulness of this new source of information.