Political sentiment analysis of press freedom

Krzysztof Rybiński

Machine learning applications are nowadays found in almost every aspect of our life[1], and politics and political research are no exceptions. Machine learning algorithms used by Oxford Analytica could have affected two recent, probably globally most significant political events: the election of Donald Trump for US president and the British referendum that decided about Brexit[2]. Algorithms (computer programs) are increasingly used to help political parties and researchers understand and predict political events. They are also used to optimize PR and marketing activities of political parties or inform and guide the political discourse. As illustrated by the literature review below, the use of machine learning in politics will continue to rise, as new, more efficient tools suitable for such analysis are continuously developed, more data becomes available and computing power rises.

This article applies machine learning to political news stories sourced from influential news portals in Kazakhstan and Poland. It is the first ever attempt in the political science literature to assess the degree of press freedom using political sentiment analysis.

Literature review: application of political sentiment models and methodology of freedom rankings

Wikipedia provides the following definition of sentiment analysis: "Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to the voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event". [accessed on 4 September 2017]

This paper applies the sentiment analysis defined above to political texts published by major news portals in order to detect what opinion about key politicians is presented in the media. It is postulated that higher press freedom should be associated with a more open and

[1] Alpaydin (2016), McAfee, Brynjolfsson (2017).

[2] "Did Cambridge Analytica influence the Brexit vote and the US election?", The Guardian, 4 March 2017, https://www.theguardian.com/politics/2017/mar/04/nigel-oakes-cambridge-analytica-what-role-brexit-trump [accessed: 15.12.2017].

critical assessment of the government, while a lack of press freedom would cause media to publish mostly positive opinions about the government.

Review of political sentiment literature

To date, there has been no research applying political sentiment analysis to the issue of press freedom. Therefore, this section briefly presents existing applications of political sentiment analysis models and discusses the methodology of the most significant freedom rankings. Earliest examples of sentiment analysis in general can be found in papers by Carbonell (1979) and Wilks and Bien (1984). Pang and Lee (2008) document that sentiment analysis and opinion mining took off around 2001, with fourteen important articles published on this topic around that date. In the following few years, there were hundreds of articles on sentiment analysis published. The list of references by Pang and Lee (2008) has 332 positions. In the past fifteen years, the applications of automated sentiment analysis truly exploded, from single papers in the 2001–2004 period to 100–120 in 2014–2015 (Piryani et al., 2017). The authors of this largest to date scientometric study on sentiment analysis, covering 488 research papers, show that nearly 40 percent of all reviewed articles used product reviews as research data sets, 32.5 percent used Twitter, blogs, messaging services and web sites, and only 9.4 percent used news articles. Hence, the contribution of this paper is in the relatively less researched area of political news stories.

Early development of the sentiment analysis was related to commercial and intelligence purposes. It was later fueled by advances in natural language processing and information retrieval tools that utilized machine learning, by rising the availability of data sets, and by an exponential rise of the amount of data that could be processed, including Internet-based customer reviews, blogs, and more recently – tweets. Also, a massive increase in computing power and the development of cloud-based computing services (such as Amazon Web Services) played a significant role. These trends are captured in the survey of sentiment analysis tasks, approaches and applications (Ravi and Ravi, 2015), that provides a detailed analysis on 251 papers written between 2002 and 2015. The most recent review of the literature on applications of text mining in the field of politics was conducted by Ngai and Lee (2016). The authors classified 55 articles published in English that applied text mining in politics according to the stage of the policy cycle: agenda setting (18 articles written between 2000 and 2013), policy formulation and decision making (10, 2007-2014), policy implementation (25, 2002–2014) and policy evaluation (1, 2010).

Political sentiment literature can be grouped by the purpose of the analysis. Many papers try to infer or predict political agenda and opinion about politicians or parties from news, tweets and blog posts: (Sindhwani and Melville, 2008), (Melville et al., 2009), (Taddy, 2013). There is a very large number of literature on predicting election results or outcomes of other important political events in many – mostly developed – countries using political sentiment analysis: (e.g. Rill et al., 2014), (Franch, 2013), (Ceron et al., 2013), (Fortuny et al., 2012).

Some papers make an attempt to infer media political bias from published news texts: (Gonzales-Bailon at al., 2014), (Niculae et al., 2015), or apply political sentiment analysis to conduct lexical analysis of news, tweets or posts: (Bosco et al., 2013), (Bakken et al., 2016).

Political sentiment analysis of news has never been conducted for Kazakhstan, and the literature about Poland is also very limited. The studies documented the polarization of the Polish political scene (Sobkowicz et. al., 2012), analyzed the emotional dimensions of Polish political forums (Sobkowicz and Sobkowicz, 2012), or calculated the correlation between political sentiment polarity and election results

(Gogolek et al., 2015). One study analyzed the potential of Twitter in analyzing individual Polish politicians' position on various issues (Ecker, 2017). There is also one example of linguistic analysis of Polish political tweets (Ogrodniczuk and Kopec, 2017).

Review of press freedom rankings methodology

Among any press freedom rankings, four stand out in terms of their publication history, country coverage, influence and presence in the media: the Word Press Freedom Index published by Reporters Without Borders; the Freedom in the World report published by the Freedom House; Freedom of Press published by the Freedom House; and the Human Freedom Index published by the CATO Institute as well as cooperating think tanks that contain a freedom of expression subcategory[3]. They use a mix of factual analysis and expert opinion to derive the relative degree of freedom in various countries, including press freedom. Their data and opinion sources are as follows:

Press freedom expert surveys and opinions sourced from, among others, think tanks, NGOs, media professionals, sociologists and scholars (RWB, Freedom House) Record of abuses and violence against journalists and media outlets (RWB, CATO) Aggregation of press freedom measures provided by other organizations (CATO).

A major source of the data used to compile a freedom of press (expression) assessment for a given country is an expert survey. Both surveys used by Freedom House and Reporters Without Borders collect expert opinion, not factual data. For example, question B6 in the Reporters Without Borders survey asks, "How easy it is for authorities to foster a firing of a journalist?". And the expert answer is in the range from 1 (authorities are powerless) to 10 (authorities can fire at will). So, the survey does not collect the data about how many journalists were actually fired by authorities, but expert opinions about this very important issue.

Similarly, the Freedom House expert survey asks, for example, "To what degree are journalists subject to editorial direction or pressure from the authorities or from private owners?", and does not attempt to collect the data in such cases. Press freedom rankings are compiled by experts, using expert knowledge and opinion, and hard data has a very small direct influence on how rankings are created. Of course, factual data has an indirect influence on the country press freedom assessment, as expert opinion is based partly on facts, but partly on expert beliefs, values, and personal experience. It is surprising, however, that in the era of rapidly advancing machine learning models, including text mining and natural language processing, they have not been used in the press freedom assessment.

This article aims to address several gaps identified in the literature review. Firstly, it presents the first ever comprehensive political sentiment analysis of news in Kazakhstan and Poland. Secondly, it demonstrates that a real political power structure can be inferred from the text mining analysis. Thirdly, it suggests that the existing freedom rankings can be improved by applying text mining models to the news published in local languages. Finally and most importantly, it creates an unbiased, automated tool to inform both the public and the government about cabinet members' perception in the media.

[3] https://rsf.org/en/ranking, https://freedomhouse.org/report/freedom-world/freedom-world-2017, https://freedomhouse.org/report/freedom-press/freedom-press-2017, https://www.cato.org/economic-freedom-world [accessed: 15.12.2017].

Data and methodology description

This article applies key steps of the political sentiment analysis research methodology. A positive opinion about a given politician is formed if, in the vicinity of his name mentioned in the media, there are more words with positive than with negative sentiment polarity. A negative opinion is formed when negative words dominate, and a neutral opinion if the number of positive and negative words is the same.

Next steps are: robustness analysis and human validation of automated political sentiment calculations[4]. The results are robust with respect to the selection of the size of the analyzed text, and human assessment of a sample of texts positively validates computer-generated outcomes. Additionally, for Kazakhstan, the results were also positively validated by applying the same methodology to a corpus of almost 50,000 texts from the portal informburo.kz. A detailed research methodology discussion follows, see figure 1.

Between 22 June 2017 and 20 September 2017, an internet spider written in R language was scraping the main websites of major and most influential news portals in Kazakhstan (in Russian) and in Poland (in Polish). In Kazakhstan, they were: informburo.kz, nur.kz, tenrginews.kz and zakon.kz. Informburo.kz is owned by Verny Capital, a private equity firm that manages over 3 bn dollars in assets, owned by the second richest Kazakhstani businessman, Bulat Utemuratov, also a member of president Nazarbayev's closest circle. Informburo has a wide coverage and publishes more financial news than other portals. Nur.kz and tengrinews.kz are affiliated with the former prime minister and the current head of the Secret Service, Karim Massimov. Tengri is a large media holding that also includes a radio station and a news agency. Zakon.kz is considered by local media specialists as a very reliable source of information, respecting good standards of journalism and presenting well-balanced views.

In Poland, scraped portals were: gazeta.pl, rp.pl and wpolityce.pl. All three institutions in Poland also publish printed daily or weekly papers and are among the most influential written media in Poland. Gazeta.pl is leaning towards liberal and center-left wing, supporting the former Polish coalition government formed by Platforma Obywatelska and PSL[5] and sometimes sympathizing with former communists SLD. Wpolityce.pl is associated with the ruling conservative, right wing party PiS, while rp.pl has a more neutral association.

In total, between 22 June and 20 September 2017, 5707 unique texts were downloaded and analyzed in Kazakhstan and 8407 in Poland, by a computer program written in R language[6]. When scraping the above portals, only articles from political, economic, international affairs, legal affairs and society sections were considered, while local news or gossip sections were ignored.

Sometimes, one text stayed on the portal's main page for longer than one day, so duplicates were removed. Then, standard text mining preprocessing steps from the bag of words[7] method

[4] Human validation methodology and results are presented after the discussion of the main results.

[5] The "Platforma Obywatelska" name in English is "Civic Platform", "PSL" translates into "Peasants Party and "PiS" into Law and Order. This article uses the parties' names in Polish.

[6] The paper uses the quanteda library in R language to conduct the text mining analysis. See (Benoit 2017).

[7] For a description of "bag of words" method in text mining with quanteda library used in this paper, see, for example, https://blog.paperspace.com/intro-to-datascience/ [accessed: 15.12.2017]. The method name was first mentioned by (Harris, 1954). The basic idea is to replace the various forms of a given word (declination, conjugation) with its lemma, so that the computer algorithm can count how many times a given word (lemma) appeared in the text. Sentiment dictionaries are developed for lemmas; hence words should be changed into lemmas so that

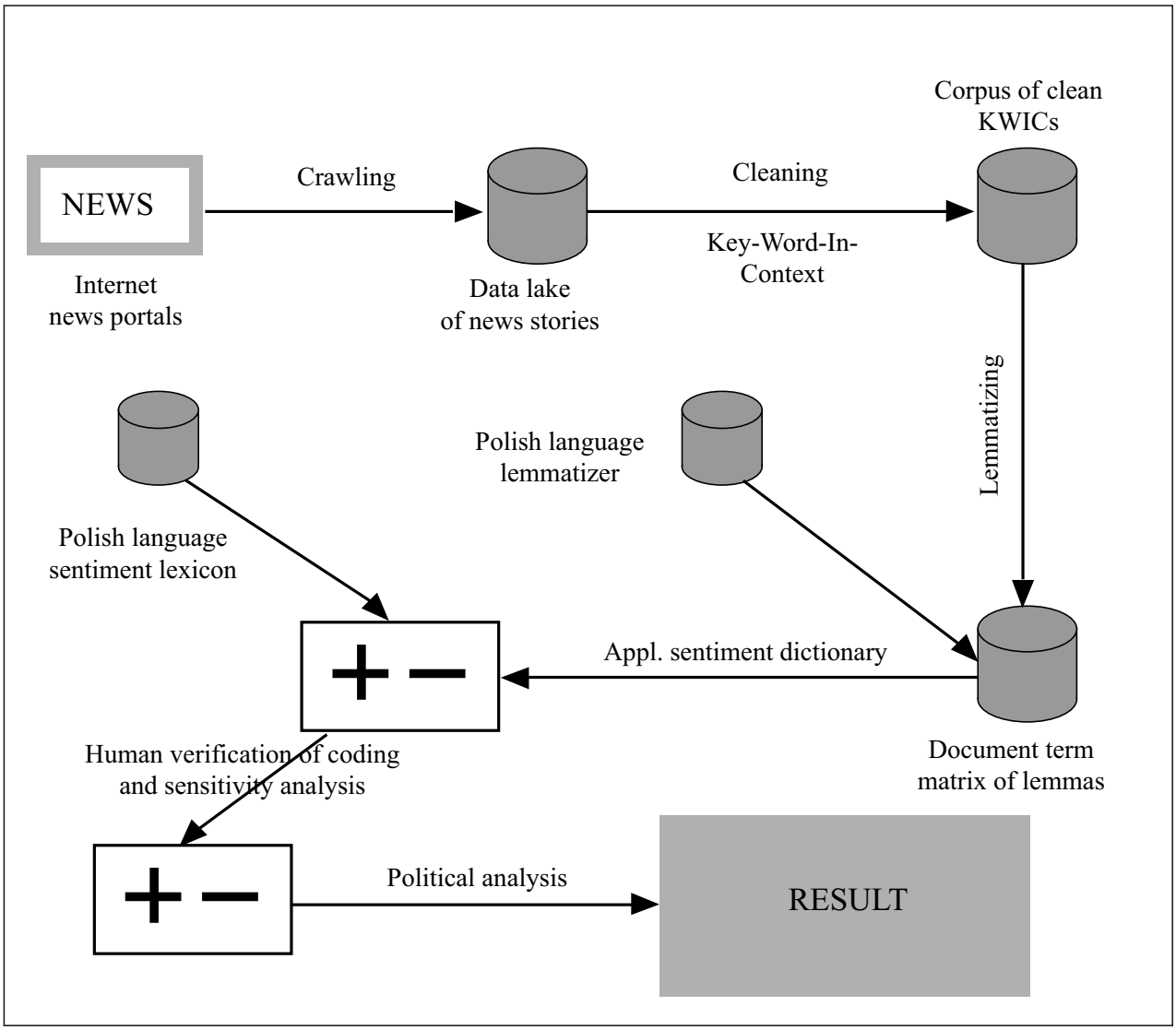

Figure 1. Research methodology (the case of Polish news portals)

Note: The research method for Kazakhstani news portals was analogous, instead of the Polish language lemmatizer and sentiment lexicon, Russian language versions were used

Source: the author

were applied. The texts were cleaned by removing punctuation, numbers, stop-words[8] and rare words[9] and capital letters were converted to lower case. After the cleaning stage, the Key--Word-In-Context[10] (KWIC) algorithm of the R quanteda library was applied to the corpus

computer program can calculate their sentiment polarity (positive, neutral or negative). An alternative method is stemming, not used in this paper.

[8] For stopwords removal, in the Polish paper the "lsa" (latent semantic analysis) library is used by R. See Wild (2015).

[9] All words that appeared less than three times were removed, as they introduce noise into the analysis.

[10] For a description of Key-Word-In-Context in the quanteda R library, see: http://quanteda.io/reference/kwic.html [accessed: 15.12.2017].

of texts. The program looked for the appearance of selected politicians' names in the corpus of texts using regular expressions, which allowed to find all declination forms of names. Then, 15 words preceding each name appearance in the text, and 15 words following the name, were extracted and formed a KWIC[11]. For each politician, a corpus of all KWICs containing her or his name was created. After converting the KWIC corpus into a document-term-matrix[12], all KWICs were lemmatized using large, freely available lemmatizers, containing 1.5 million word-lemma pairs in Russian and 3.3 million pairs in Polish[13]. Then, sentiment analysis was applied to the lemmatized KWICs using large sentiment polarity dictionaries containing more than 10,000 unique words in Russian and in Polish. The dictionaries were manually annotated with sentiment polarity. In Russian (Loukachevitch Levchik, 2016), three sentiment polarity levels were available: positive, neutral and negative[14], while in Polish, there were five levels: strongly negative, moderately negative, neutral, moderately positive and strongly positive (Zaśko-Zielińska et al., 2015). Words with a strongly negative sentiment polarity were assigned a "-2" score, negative/moderately negative - a "-1" score, neutral – a zero score, positive/moderately positive – a "+1" score, and strongly positive - a "+2" score. The sentiment algorithm calculated how many words with identified sentiment polarity appeared in each KWIC, applied the assigned scores and calculated the total sentiment score for each KWIC. The higher (positive) the total sentiment score, the better opinion (sentiment) about the politician was expressed in a given KWIC. Conversely, the lower (more negative) the KWIC score, the worse the opinion about a given politician was expressed in that particular KWIC. The following statistics were calculated for the KWIC corpora for each analyzed politician:

- Number of KWICs (mentions of politician's name in the media).
- Average sentiment score for all KWICs for each politician.
- Percent of positive KWICs.
- Percent of negative KWICs.
- Net difference (percent positive minus percent negative).

The paper only reports the results for a 15-word KWIC size, but a sensitivity analysis for KWIC size between 5 and 25 words was conducted and the results were very similar, i.e. average political sentiment score for every analyzed politician remains almost the same irrespectively of the KWIC size.

Research hypothesis

Political sentiment analysis was applied to the name of president Nazarbayev and all government members in Kazakhstan. No opposition member was selected as in the Kazakhstan opposition, as a real political power does not exist. In Poland, the analysis covered president Duda, all cabinet members and Jaroslaw Kaczynski, the ruling party leader who does not hold any formal position in the government.

As shown in table 1, in all influential freedom rankings, Poland is rated as "free" or "partly" country, while Kazakhstan is judged as "not free", and the freedom gap between both countries is very large. Consequently, it was expected that the number (percentage) of

[11] We repeated the analysis for KWIC size between 5 and 25. The results were similar, i.e. ranking of opinions about politicians remained unchanged, as documented in the sensitivity analysis part of this article.

[12] Document-term-matrix (dtm) is used for bag-of-words algorithms. Each row represents one document in the corpus, and each column one word. Dtm[i,j] describes how many times jth word appears in ith document.

[13] Polish language lemmatizer used for this research is available at www.lexiconista.com/datasets/lemmatization/

[14] See: www.labinform.ru/pub/rusentilex [accessed: 15.12.2017].

Table 1. Kazakhstan and Poland in the most influential freedom rankings

Ranking name	Kazakhstan	Poland	Authors, source
Word press freedom index 2017	Ranked as 157th among 180 countries	Ranked as 54th among 180 counties	Ranking prepared by Reporters Without Borders, https://rsf.org/en/ranking_table
Freedom in the world report 2017	Not free (political rights – 7, civil rights – 5)*	Free (political rights – 1, civil rights – 2)*	Freedom house, https://freedomhouse.org/report/freedom-world/freedom-world-2017
Freedom of the press 2017	Not free, score 85 (100)**	Partly free, score 34 (100)**	Freedom house, https://freedomhouse.org/report/freedom-press/freedom-press-2017
The human freedom index 2016	Ranked 96th in the overall ranking, a score of 5.17 in the Expression and Information category***	Ranked 21st in the overall ranking, a score of 8.9 in the Expression and information category***	CATO Institute et al., https://www.cato.org/human-freedom-index

*1 – least free, 7 – most free; **0 – most free, 100 – least free; ***0 – least free, 10 – most free

Source: freedom rankings and reports

KWICs with a negative sentiment polarity will be much higher in Polish media than in Kazakhstani ones. A lack of political and press freedom in Kazakhstan, as described in the freedom rankings, should have prevented the publication of critical opinions about the president and government in Kazakhstan, while critical views should be more common in Poland.

It was also expected that political sentiment analysis would reveal the real power structure, with most influential politicians appearing in the media very frequently and possibly enjoying a more positive opinion.

Key results: revealing the real political power structure and verifying the freedom rankings

The first basic measure of each politician's activity and popularity is the number of mentions of her or his name in the media. If a politician is not mentioned often, it means that she or he is not active and/or is not associated with important reforms. Of course, there may be cases of "celebrity-politicians" that are covered by the media not because of their influence or professional activity, but because of their "celebrity status". But among the analyzed names in both countries, there were no such cases.

Figure 2 shows that president Nazarbayev totally dominates political news in Kazakhstan, which reflects his real influence. Despite recent reforms that transfer some powers from the Kazakhstani president to the parliament and government, both his formal and informal dominance in Kazakhstani politics is undisputed. He sets the political, economic and social agenda, appoints ministers and regional leaders at will, and even appoints and dismisses the central bank governor as he wishes. The political status of key members of his administration is much higher than that of the Kazakhstani government members. In Poland, two politicians enjoy very high and similar coverage in the media: president Duda and the ruling party leader Kaczynski. It is somewhat surprising, as Jaroslaw Kaczynski is the main architect of the

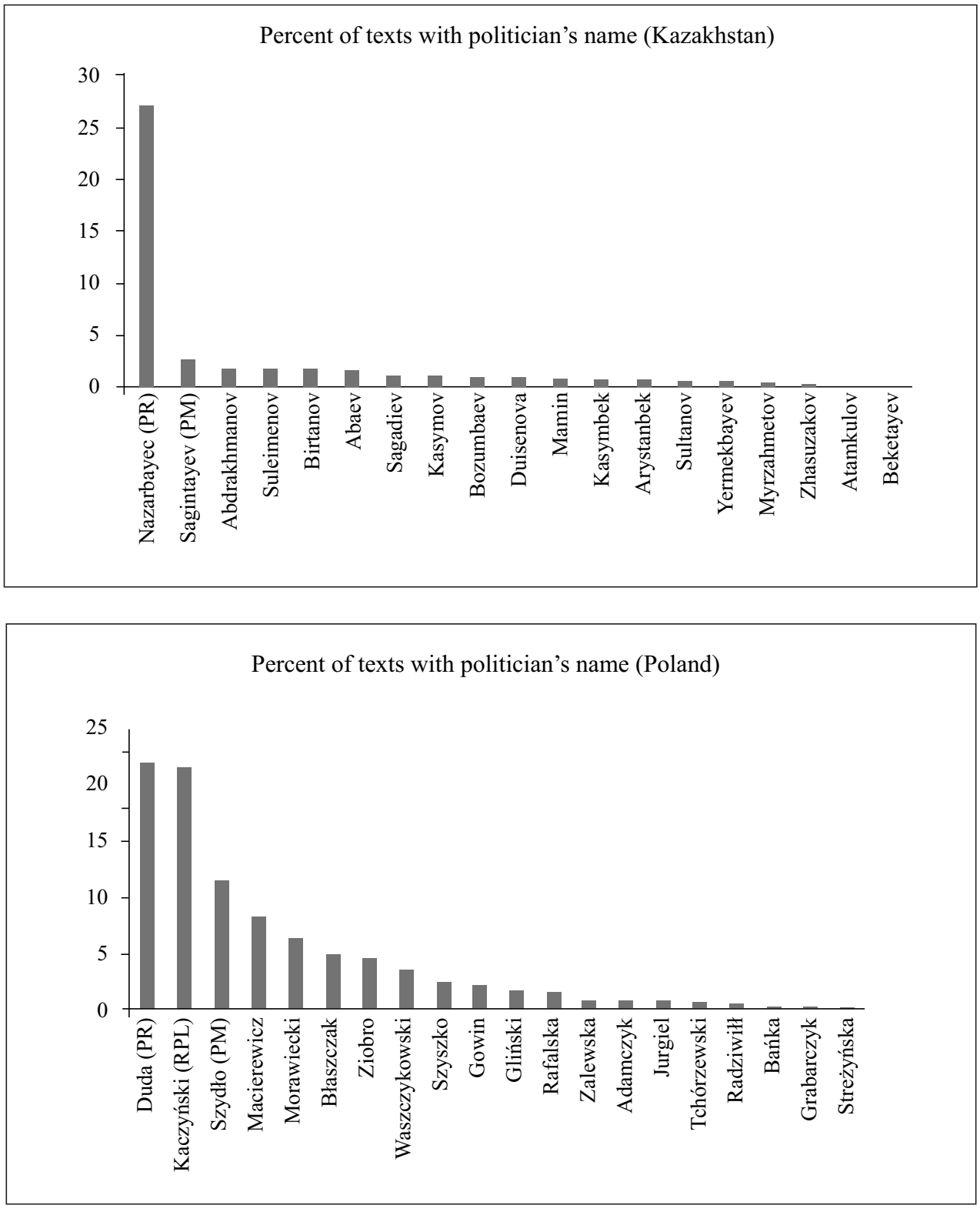

Figure 2. Percentage of mentions of politicians' names in the media (%)

Note: The percentages on the chart are calculated as the number of appearances of a given politician's name divided by the total number of articles in the corpus for a given country

Source: analysis of news portals in Kazakhstan and Poland

PiS ruling party agenda and has a decisive voice in all personal appointments, including the selection of Andrzej Duda as the PiS candidate in the 2015 presidential elections. It confirms that president Duda, after winning in the 2015 elections, successfully avoids the role of a "PiS puppet", forms an independent political center, and even makes decisions that are highly unpopular among PiS elites, such as vetoing two out of three laws transforming the Polish judiciary system.

While in the case of the Kazakhstani president Nazarbayev's dominance in the media is incredibly high, exceeding prime minister Sagintayev's appearance in the media by a factor of ten, the Polish political scene exhibits much more pluralism, with prime minister Szydlo and several key ministers also appearing in the media frequently.

The above results show that machine learning techniques can be used to automatically retrieve the real power structure of a country's political elites from news portals and to identify the most influential politicians. Interestingly, it can be done automatically, in many languages, for many countries without any prior knowledge about the local political scene, providing that local news portals allow internet robots access to their sites.

Opinion about politicians in the media

Tables 2 and 3 below present the average sentiment score for each analyzed politician in Kazakhstan and Poland, percent of KWICs with positive and negative polarity as well as the net result (positive minus negative). Names are

Table 2. Political sentiment analysis for politicians in Kazakhstan

Politician	Average sentiment score	Positive (%)	Negative (%)	Net (pp.)	Number of mentions
Zhasuzakov	1.0	60.0	0.0	60.0	20
Nazarbayev (PR)	0.9	53.1	10.1	43.0	1542
Arystanbek	0.8	62.2	11.1	51.1	45
Beketayev	0.8	50.0	0.0	50.0	4
Mamin	0.7	52.0	8.0	44.0	50
Abaev	0.4	41.8	12.1	29.7	91
Sagadiev	0.4	34.8	7.2	27.5	69
Myrzahmetov	0.4	29.2	8.3	20.8	24
Sagintayev (PM)	0.4	38.0	12.7	25.3	150
Abdrakhmanov	0.4	34.9	15.1	19.8	106
Suleimenov	0.2	36.2	18.1	18.1	105
Sultanov	0.2	36.6	22.0	14.6	41
Yermekbayev	0.2	47.1	41.2	5.9	34
Atamkulov	0.1	44.4	11.1	33.3	9
Bozumbaev	0.0	21.8	21.8	0.0	55
Birtanov	0.0	28.0	26.0	2.0	100
Duisenova	-0.1	19.2	23.1	-3.8	52
Kasymov	-0.5	18.0	42.6	-24.6	61
Kasymbek	-0.6	17.0	38.3	-21.3	47

Source: four news portals, 22 June–20 September 2017, author's calculations

Table 3. Political sentiment analysis for politicians in Poland

Politician	Average sentiment score	Positive (%)	Negative (%)	Net (pp.)	Number of mentions
Banka	2.4	61.1	16.7	44.4	18
Tchorzewski	1.6	59.3	14.8	44.4	54
Glinski	1.5	52.4	19.3	33.1	145
Morawiecki	1.4	54.2	19.2	35.0	537
Strezynska	1.2	40.0	60.0	-20.0	5
Duda (PR)	1.1	50.6	20.1	30.5	1854
Rafalska	1.0	50.8	22.7	28.0	132
Gowin	0.9	46.4	24.9	21.5	181
Macierewicz	0.7	48.6	25.1	23.4	704
Szydlo (PM)	0.7	44.5	26.7	17.8	969
Waszczykowski	0.7	45.2	23.9	21.3	301
Adamczyk	0.4	44.3	19.7	24.6	61
Ziobro	0.3	44.9	31.6	13.3	376
Zalewska	0.3	45.5	33.3	12.1	66
Kaczynski (RPL)	0.2	39.8	32.9	7.0	1823
Szyszko	0.2	40.3	32.0	8.3	206
Jurgiel	0.0	29.3	29.3	0.0	58
Radziwill	0.0	33.3	31.1	2.2	45
Grobarczyk	0.0	42.9	35.7	7.1	14
Blaszczak	-0.4	33.8	42.6	-8.8	420

Source: three news portals, 22 June–20 September 2017, author's calculations

sorted by the value of the average sentiment score, from most positive to most negative. PR denotes president, PM – prime minister, RPL – ruling party leader in Poland.

The main research hypothesis postulated that critical views about ruling party politicians and the president will be rare in Kazakhstan, ranked as a "not free" country, and much more frequent in Poland, defined in freedom rankings as a "free" or "mostly free" country. The political sentiment analysis showed that, in reality, critical views are much more frequent in Kazakhstan than in Poland. Four cabinet members received a negative average sentiment score in Kazakhstan, and only one in Poland. The simple average opinion for all politicians in Kazakhstan is 0.3 and in Poland 0.7. Also, the president score in Kazakhstan is 0.9, and in Poland – 1.1. So, it seems that media in Kazakhstan are much more critical about domestic politicians than the media in Poland.

These results for Kazakhstan were surprising, so an additional attempt was made to verify them. Additional data was obtained from the informburo.kz portal; it provided 48512 articles published on the portal's main page between April 2015 (portal launch date) and July 2017. Analogous political sentiment analysis was applied to this large corpus and the results are reported in table 4 below.

Over a period of more than two years, the portal informburo.kz was very critical of the cabinet members and more than half of them have a negative average sentiment score. At the same time, president Nazarbayev emerges as a clear leader, both in terms of the frequency of

Table 4. Political sentiment analysis for politicians in Kazakhstan (additional data)

Politician	Average sentiment score	Positive (%)	Negative (%)	Net (pp.)	Number of mentions
Nazarbayev	0.39	41.8	17.9	23.9	9735
Mamin	0.26	28.8	13.7	15.0	153
Abaev	0.16	32.4	23.3	9.1	519
Sagadiev	0.09	30.0	21.8	8.2	463
Yermekbayev	0.07	34.7	34.7	0.0	118
Sagintayev	0.03	27.5	23.7	3.7	914
Zhasuzakov	0.02	25.7	26.2	-0.5	183
Kasymbek	0.01	28.0	23.3	4.7	296
Atamkulov	0.00	20.3	21.6	-1.4	148
Sultanov	-0.02	26.5	26.7	-0.2	663
Abdrakhmanov	-0.10	26.6	31.2	-4.6	523
Beketayev	-0.10	28.4	31.9	-3.4	116
Arystanbek	-0.14	26.3	33.3	-7.0	676
Birtanov	-0.14	26.7	35.5	-8.8	251
Myrzahmetov	-0.18	21.1	29.0	-7.8	383
Duisenova	-0.25	21.3	33.3	-12.1	447
Suleimenov	-0.33	22.7	37.8	-15.1	669
Kasymov	-0.44	20.6	43.4	-22.8	913
Bozumbaev	-0.57	15.9	46.5	-30.5	691

Source: portal informburo.kz, 48502 articles, April 2015–July 2017, author's calculations

media presence and as the politician seen in the most positive light. However, his average sentiment score in significantly below levels scored by top Polish politicians. Hence, earlier results were confirmed and reinforced.

There are three ways one may challenge the results of this comparative analysis of Kazakhstan and Poland. Firstly, the government in Poland may be better than that found in Kazakhstan, so it enjoys better opinions in the media. However, the quality of government is a very complex issue, which depends on many factors and is very subjective. Moreover, we cannot verify this claim by looking at the election results, as there is no real political opposition in Kazakhstan. So, it is not possible to debate the "good/bad government" argument here. Secondly, automated text analysis may be misleading, i.e. it may classify negative texts as positive or vice versa. In what follows, we present the human verification of automated text classification, which positively validates the results. Thirdly, one may argue that sentiment lexicons in Russian and Polish languages are biased, i.e. the share of positive sentiment polarity words in Polish is higher than their share in Russian, and for negative sentiment polarity, the relation is opposite.

The shares of positive/negative sentiment polarity words are respectively: 30/70 percent in Polish and 34/66 percent in Russian. This small difference is unlikely to bias the results towards more positive opinions in Kazakhstan. It is interesting that there are more words with a negative sentiment polarity in both languages, but in political texts, positive polarity words are used more often than negative ones. This is true for both countries, but less so in the case of Kazakhstan.

Finally, one could also argue that Kazakhstan set up a system where you can criticize every politician, with the exception of the president. And because the real power is concentrated in the hands of one man and his family, the allowed media criticism of cabinet ministers tells very little of the real media freedom in Kazakhstan. However, we found that ten percent of texts in the smaller corpus and 18 percent in the larger corpus mentioning Nazarbayev have a negative sentiment polarity.

The author has published many articles in newspapers, internet portals and social networks in both Poland and Kazakhstan criticizing low quality government and local administration decisions and policies. In Poland, this criticism was related to many aspects of the government's activity (excessive bureaucracy, failed government investments, ineffective structural policies, misguided monetary policy, lack of accountability or strategic vision, excessive fiscal deficit). In Kazakhstan, the author's articles were focused on bad regulation in the education sector, low quality of monetary policy and lack of independence of the central bank). In Poland, the author received personal threats from politicians, one think-tank was told to stop cooperating with the author, and the author's spouse, working at the state bank at the time, was told that she could lose her job unless her husband stops criticizing the government. In Kazakhstan, there has not been a single case of such pressure from the government or its affiliates. However, the situation changes when it comes to criticizing the president. When the results were shown during the class on quantitative text analysis at the Narxoz University in Almaty, the first reaction of the participants was that the minister that came ahead of president in the ranking may be fired, and that the authors of critical articles may be targeted by authorities.

The authors' own experience and presented research results show that the large gap between press freedom assessment in official rankings for Kazakhstan and Poland is not justified and that these rankings would benefit from incorporating the political sentiment models into their methodology.

Validation of results by comparing human coding to computer generated results

The key question in text mining is whether the results make sense at all. The bag of words methodology used in this paper does not take into account negation[15], irony or sarcasm. There are two general approaches to tackling this problem. One can apply different text mining methods to verify results or one can apply human coding to a sample of texts and check whether computer and human judgment match. This paper takes the second approach.

The average sentiment score for a given politician depends strongly on the correct classification of documents with the highest (positive) and lowest (negative) sentiment score. Therefore, three Kazakhstani and two[16] Polish ruling party politicians were selected to verify the results. For Kazakhstan, the validation was conducted for president Nazarbayev, who enjoys a high sentiment score, prime minister Sagintayev who is in the middle of the sentiment polarity table, and minister of internal affairs Kasymov, who is presented in the media in

[15] For example, any adjective with positive polarity will retain its polarity even when it is preceded by "not" or "lack of", or any other negation phrase.

[16] For another research project, Poland's results verification was conducted for three politicians, two from the ruling party and one from the opposition. The results for the opposition match the ones for the ruling party. As this paper covers only ruling party politicians, only their results are reported here.

Table 5. Verification of computer text classification by human coders (Kazakhstan)

Kazakhstani politicians	Nazarbayev	Sagintayev	Kasymov
Average judgment	2.93	3.16	2.72
Average judgment positive sentiment	3.87	3.45	3.34
Average judgment negative sentiment	1.99	2.85	2.09
Correlation: human coding vs. computer assessment	0.93	0.51	0.78
Number of coders with proper sentiment ordering	14 (17)	16 (17)	16 (16)

Source: human coding experiment, 17 coders

Table 6. Verification of computer text classification by human coders (Poland)

Polish politicians	Morawiecki	Ziobro
Average judgment	2.95	2.83
Average judgment positive sentiment	3.43	3.24
Average judgment negative sentiment	2.47	2.42
Correlation: human coding vs. computer assessment	0.44	0.48
Number of coders with proper sentiment ordering	6 (7)	8 (8)

Source: human coding experiment, 8 coders

a negative light. For Poland, the validation was performed for government deputy prime minister Morawiecki, who supervises finance and economic development ministries, is often mentioned in the media and enjoys good opinions. The second choice was justice minister Ziobro, who is in the lower part of the sentiment ranking and has been heavily criticized in the international media for his judiciary reforms.

As this research was conducted without any funding, the availability of human coders was limited, and each coder was only willing to read a limited number of texts. Therefore, the following approach was adopted. The texts that were most positive and most negative on the sentiment scale were selected. These texts contain many positive or negative sentiment polarity words and as such should be easier to code by humans, but they also have a relatively large impact on the politician's average sentiment score. So, if there were no agreement between the humans and the computer for such a sample of texts, the research method would be rejected. However, the obtained positive validation results of the experiment may still leave room for criticism that a much larger, random sample of texts should be selected. That is why additional validation methods were used (analysis of a large corpus of near 50,000 texts in Kazakhstan, reported in this paper; and application of the same methodology and lexicons to another sample of political texts in Polish news portals; in Polish, not reported here, but available upon request). In each case, the obtained results confirmed the ones reported in this paper. A detailed description of the human validation procedure follows.

Ten most positive and ten most negative KWICs for each politician were printed and human coders were asked to mark each KWIC with the sentiment score: 1 for strongly negative sentiment, 2 for moderately negative sentiment, 3 for neutral sentiment, 4 for moderately positive sentiment and 5 for strongly positive sentiment. There were 17 human coders for Kazakhstan and eight coders for Poland, all were nationals of the respective country and fluent in their home language. The results are reported in tables 5 and 6 below.

Table 7. KWICs with most positive and most negative sentiment polarity determined by computer for each politician covered by the validation procedure

Most Positive / Negative Score	Name, country (original KWIC)	Name, country (KWIC English translation by Google Translate* with minor edits)
	Nazarbayev, Kazakhstan	
8	„Главу государства с днем рождения и назвал все успехи независимого Казахстана наглядным результатом проводимой Нурсултаном дальновидной политики. Ваш богатый профессиональный и жизненный путь является прочной основой Вашей успешной”	„The head of state has a happy birthday and he described all the successes of independent Kazakhstan as a visual result of the far-sighted policy pursued by Nursultan. Your rich professional and life path is a solid foundation for your successful”
-4	„Казахстана Нурсултан Назарбаев призвал к прекращению насилия в Мьянме, передаёт МИА Казинформ. Нурсултан подчеркнул, что от террористических атак страдают страны мусульманского мира. От их террористических атак”	„Nursultan Nazarbayev of Kazakhstan called for an end to the violence in Myanmar, passes MIA Kazinform. Nursultan stressed that the countries of the Muslim world are suffering from terrorist attacks, and from their terrorist attacks”
	Sagintayev, Kazakhstan	
5	„об активном участии отечественных предпринимателей в социальных и инфраструктурных проектах. По словам премьер-министра Бакытжана, бизнес должен оценить столь значительное упрощение процедуры и поддержку проектов ГЧП на правительственном уровне”	„on the active participation of domestic entrepreneurs in social and infrastructure projects. According to Prime Minister Bakytzhan, business should assess such a significant simplification of the procedure and support for PPP projects at the governmental level”
-3	„ситуацией по защите прав потребителей в стране, передает корреспондент Zakon.kz. Премьер-министр РК Бакытжан поручил правительству ужесточить контроль в сфере защиты прав потребителей. Поручаю Министерству национальной экономики”	„The situation concerning consumer rights protection in the country, Zakon.kz reports, Prime Minister of the Republic of Kazakhstan Bakytzhan, instructed the government to tighten control in the sphere of consumer rights protection, „I entrust the Ministry of National Economy”
	Kasymov, Kazakhstan	
3	„2000 ценных экспонатов и произведения мастеров кисти, в том числе меч казахского хана Кенесары. Мероприятие продолжилось концертной программой с участием артистов областной филармонии, которые исполнили традиционные казахские”	„2000 valuable exhibits and works by artists of the brush, including the sword of the Kazakh khan Kenesary. The event was continued by a concert program with the participation of the artists of the regional philharmonic society who performed traditional Kazakh”
-3	„Судьи пересмотрели своё решение по протесту прокурора и отменили свой запрет – рассказал Калмуханбет – добавив, что приглашение на свадьбу вынужден отклонить. Если вы нашли ошибку в”	„The judges reconsidered their decision on the prosecutor’s protest and canceled their ban,” Kalmukhanbet said, adding that he had to reject the invitation to the wedding.”

	Morawiecki, Poland	
14	„mniej niż na Zachodzie często emigrują Piąta pułapka to pułapka niskich marż przeciętnych produktów wyjaśniał Wicepremier mówił o filarach rozwoju gospodarczego Pierwszy filar rozwoju gospodarczego Polski to filar innowacyjnych firm”	„we emigrate less than in the west. The fifth trap is a trap of low margins of average products, as explained Deputy Prime Minister, who spoke about the pillars of economic development. The first pillar of Poland's economic development is the pillar of innovative companies”
-10	„prognozować jaki będzie stan deficytu budżetowego na koniec roku Na pewno jakiś deficyt wystąpi ocenił Wicepremier sądzi że deficyt sektora finansów publicznych będzie na koniec tego roku niższy niż zaplanowane”	„forecasting what will be the state budget deficit at the end of the year. Some deficit will definitely be assessed, Deputy Prime Minister believes; he thinks that the deficit of the public finance sector will be lower than planned at the end of this year”
	Ziobro, Poland	
10	„tygodnia związane z niespodziewanym zatrzymaniem reformy sądownictwa przez prezydenta Andrzeja Dudę W mocnym wywiadzie Zbigniew jasno podkreśla że głowa państwa postawiła się przed wyborem historyczna wielkość albo groteska Tymczasem owe”	„week associated with the unexpected withhold of the reform of the judiciary by President Andrzej Duda. In a strong interview, Zbigniew clearly emphasized that the head of state set before the historic choice of size or grotesque. Meanwhile, these”
-13	„jeśli notariusze dopuścili się tutaj nieprawidłowości w zależności od ich skali poniosą również konsekwencje oświadczył Podejrzani mają zarzuty działania w grupie przestępczej oszustw oraz lichwy nie przyznali się do zarzutów”	„If the notaries allowed for irregularities to appear depending on their scale, they will also bear consequences – he stated. The alleged suspects are charged with acting in the criminal fraud group and usury; they did not admit to the charges”

* as this paper presents the results of machine learning analysis of text, it was natural to use Google Translate for the translation. The author, who is fluent in both Polish and Russian, confirms that the Google translation was of good quality

Source: Internet portals, authors' calculations and Google Translate

The first three lines in the tables present, respectively: average human score for all texts, average human score for texts labeled as positive by the computer and average human score for texts labeled as negative by the computer. In each case, the average human assessment confirms the computer results. Texts labeled by computer as positive/negative are judged as positive/negative by human coders and receive the score of above/below 3 on average. The fourth row presents correlation coefficients between the computer assessment and human judgment, which are positive and significant at a 5% confidence level in all cases, and in the case of president Nazarbayev, the correlation is close to one. Interestingly, the correlations for Kazakhstan are higher, which implies that computer and human judgments are closer for the Kazakhstani politicians, than for the Polish ones. The last row presents the number of coders that classified positive and negative documents the same way the computer did, i.e. gave a higher average score for positive sentiment documents than for negative sentiment ones. The results show that almost all coders agreed with the computer, but there were some exceptions. It shows that when humans are presented with short texts (total KWIC length is 30 words), sometimes – very rarely in this case – they disagree between themselves in their assessment.

The presented results show that, in the case of the most positive and the most negative documents, which determine the politician's overall sentiment score, the computer assessment and human coders' judgments are similar.

Finally, table 7 below presents the KWICs with the most positive and the most negative computer sentiment assessment for each politician covered by the validation procedure, in the original language and with an English language translation.

While the computer and humans agree about the positive and negative sentiment polarity of short texts, the examples presented in table 7 show that there are several problems when one treats the text as a "bag of words", ignoring word sequences, part of speech recognition, negation or sarcasm. For example, the most negative KWIC with the name Nazarbayev is the president of Kazakhstan statement criticizing violence in Myanmar. While the word "violence", "terrorism" and "attack" have a clear negative sentiment polarity, the entire text does not criticize the president as such. However, one can argue that the fact that the politician's name is surrounded by many negative words may result in a negative perception about this politician. Another problem is related to the "sector bias". If a given politician's area of professional activity is related to negative events (e.g. combating crime), such as in the case of justice minister Ziobro or interior minister Kasymov, it may lead to a more negative computer assessment than in the case of the sports or culture ministers, who often appear in text related to national teams' wins or to leisure. However, new or improved methods of automated text analysis are being developed, so many of the above-mentioned deficiencies can be reduced or eliminated in the future. Moreover, the size of the sector bias will probably be similar across countries, so it should not significantly impact the overall assessment of press freedom based on automated text analysis.

Conclusions

This paper presents the automated political sentiment analysis of 5707 texts in Russian, sourced from four influential portals in Kazakhstan, and 8407 texts from three such portals in Poland, in the Polish language, covering politics, economics, social, legal and community affairs. The text collection took place between 22 June and 20 September 2017. A computer detected all appearances of the names of the country's president, ruling party leader and all cabinet ministers, and counted how many positive and negative words appear in the 15-word vicinity of a given name. The results show that the political news in Kazakhstan contain more negative sentiment polarity words than the news in Poland. This is surprising given the fact that, in all important freedom rankings, Poland is defined as a "free" country, and Kazakhstan as a "not free" country. The results were positively validated by comparing the computer assessment with human judgment. Furthermore, the results were validated and reinforced by analyzing a large corpus of almost 50,000 texts published by informburo.kz in Kazakhstan between April 2015 and July 2017.

The discussion on the methodology applied by four of the most internationally recognized press freedom rankings shows that human subjective judgment caries a large weight in the overall press freedom assessment. With the rapid development of the automated text analysis in many languages, these freedom rankings can benefit from including such automated text analysis methods into the ranking methodology. The benefits would be twofold: (1) using new, vast amounts of text data will make rankings more reliable, and (2) possible human perception bias will be reduced, as the ranking will be based more on facts, and less on subjective human opinions.

This paper documents that the applied "bag of words" model has some drawbacks, but fast developments of new and improved text mining techniques will allow for a more precise

text analysis. There are many possible practical applications of the presented research method and results presented in this paper. Counting the number of appearances of names of politicians in most influential media can reveal the real political power structure, including the formal and informal role that a given politician plays in shaping national agenda. More importantly, political power assessment can be performed even without prior knowledge on the country's political scene. This approach can be useful for foreign investors or hedge funds considering direct or portfolio investment in developing countries, where in-depth expert knowledge is not available or is very expensive. Another possibility is informing both the public and the government about the cabinet members' opinion formed in the media. And unlike opinions formed by humans (experts or journalists), automated text mining analysis is politically unbiased. As such, political sentiment results can influence personal decisions during cabinet reshuffles. This approach can also be used to investigate product or company brand perception in the media.

References

Alpaydin, E. (2016). "Machine Learning: The New AI", MIT Press.

Bakken, P. F., Bratlie, T. A., Marco, C., & Gulla, J. A. (2016). Political News Sentiment Analysis for Under-resourced Languages. In *COLING* (pp. 2989–2996).

Benoit, K. (2017). "quanteda: Quantitative Analysis of Textual Data", 15 August 2017.

Bosco, C., Patti, V., Bolioli, A. (2013). "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT, Knowledge-Based Approaches to Concept-Level Sentiment Analysis", *IEEE Intelligent Systems.*

Carbonell, J. (1979). "Subjective Understanding: Computer Models of Belief Systems. PhD thesis, Yale.

Ceron, A., Curini, L., Iacus, S. M. (2015). "Using sentiment analysis to monitor electoral campaigns: Method matters evidence from the United States and Italy". *Social Science Computer Review*, 33 (1), 3–20.

Ecker, A. (2017). Estimating policy positions using social network data: cross-validating position estimates of political parties and individual legislators in the Polish parliament. *Social Science Computer Review*, 35 (1), 53–67.

Fortuny, E. J., Smedt, T. D., Martens, D. & Daelemans, W. (2012). " Media coverage in times of political crisis: A text mining approach", *Expert Systems with Applications* 39 (2012) 11616–11622.

Franch, F. (2013). (Wisdom of the Crowds)^2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics*, 10, 57–71. doi:10.1080/19331681.2012.705080.

Gogołek, W., Jaruga, D., Kowalik, K. & Celiński, P. (2015). Z badań nad wykorzystaniem rafinacji informacji sieciowej Wybory prezydenckie i parlamentarne 2015. *Studia Medioznawcze*, 3 (62), 31–40. (In Polish).

González-Bailón, S., Morales, G. D. F., Mendoza, M., Khan, N. & Castillo, C. (2014). Cable news coverage and online news stories: A large-scale comparison of digital media content. In: *Annual Meeting of the International Communication Association* (ICA). Harris, Z. (1954). "Distributional Structure". Word. 10 (2/3): 146–62.

Loukachevitch N., Levchik A., 2016. "Creating a General Russian Sentiment Lexicon". In: *Proceedings of Language Resources and Evaluation Conference LREC-2016.*

McAfee, A., Brynjolfsson, E. (2017). "Machine, Platform, Crowd: Harnessing Our Digital Future", W. W. Norton & Company.

Melville, P., Gryc, W., Lawrence, R. D. (2009). "Sentiment analysis of blogs by combining lexical knowldge with text classification". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1275–1284). ACM.

Ngai, E. W. T., Lee, P. T. Y., (2016). A Review of the literature on Applications of Text Mining in Policy Making. In: *PACIS* (p. 343).

Niculae, V., Suen, C., Zhang, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015), May. Quotus: The structure of political media coverage as revealed by quoting patterns. In: *Proceedings of the 24th International Conference on World Wide Web (pp. 798–808). International World Wide Web Conferences Steering Committee*.

Ogrodniczuk, M., Kopeć, M. (2017). "Lexical Correction of Polish Twitter Political Data". In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 115–125).

Piryani, R., Madhavi, D. & Singh, V. K. (2017). "Analytical mapping of opinion mining and sentiment analysis research during 2000–2015". *Information Processing & Management*, 53 (1), 122–150. Ravi, K., & Ravi, V. (2015). "A survey on opinion mining and sentiment analysis: tasks, approaches and applications". *Knowledge-Based Systems*, 89, 14–46.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R.V. (2014). "Early detection of emerging political topics on Twitter and the impact on concept-level sentiment analysis", *Knowledge-Based Systems* 69 (2014): 24–33.

Sindhwani, V., Melville, P. (2008). "Document-word co-regularization for semi-supervised sentiment analysis", In: *Eighth IEEE International Conference on Data Mining*, 1025–1030, December 2008.

Sobkowicz, P., Sobkowicz, A. (2012). Two-year study of emotion and communication patterns in a highly polarized political discussion forum. *Social Science Computer Review*, 30 (4), 448–469.

Sobkowicz, P., Kaschesky, M. & Bouchard,G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web, *Government Information Quarterly* 29 (2012): 470–479.

Taddy, M. (2013). "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression", *Technometrics*.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2010). "Predicting elections with twitter: What 140 characters reveal about political sentiment". *Icwsm*, *10* (1), 178–185.

Wild, F. (2015). lsa: Latent Semantic Analysis, 8 May 2015, Wilks, Y., Bien, J. (1984). Beliefs, points of view and multiple environments, In *Proceedings of the international NATO symposium on artificial and human intelligence*, pp. 147–171, USA, New York, NY: Elsevier North-Holland, Inc.

Zaśko-Zielińska, M., Piasecki, M. & Szpakowicz, S. (2015) A LargeWordnet-based Sentiment Lexicon for Polish, *Proceedings of the International Conference Recent Advances in Natural Language Processing* (RANLP'2015), pp. 721–730.