sciendo

Social Communication
Volume 1 (2021), pp. 110-123
DOI: 10.2478/sc-2021-0011

POLITICAL ATTITUDES OF VOTERS ON TWITTER IN THE SECOND ROUND OF THE POLISH PRESIDENTIAL ELECTIONS 2015

Rafał Piotr Paradowski[1]

Abstract

This study aims to answer the question of whether and how the voting attitudes of Polish Twitter users correlate with the election results. It also attempts to understand the online mechanisms of expressing political preferences. The data sample consisted of 8698 tweets attributed to 3508 users concerning attitudes towards the two candidates in the second round of the 2015 presidential election in Poland. Research included semantic analysis and word count techniques. Both approaches yielded similar results and were extremely close to the official post-election outcome – smallest offset amounted to less than 0.1. Moreover, experimental exploration of tweets, users' behaviour, interactions and dynamics of tweet activity was conducted.

Key words: Twitter, sentiment analysis, presidential elections, political discourse.

1. Introduction

Recent years have made Twitter – a microblogging platform – a popular tool for conducting research aimed on predicting election results. They are based on an assumption that all expressions and messages of a typical user on Twitter can be quantified and translated to a political vote or at least could mark trends in public discourse.

The idea of this research was to verify whether it is possible to gain accurate information about voting attitudes of Twitter users, and to understand some of the online mechanisms of expressing political preferences. The study exploits most of previously used techniques in exploration of tweets. It also aims to transfer certain standards used in English text analysis to the research of internet communication conducted in Polish.

Research focuses on the last days of presidential elections campaign in Poland in 2015 as they were the best chance for gaining strictly polarized political opinions of Polish users on Twitter in recent years. This type of elections is general, nationwide and is conducted as a simple majority system, which always results in higher voter turnout and greater public interest. The years 2014-2015 comprised a period of particularly increased activity of institutional and private users on Twitter due to four general elections held at that time. Nevertheless, it was the presidential election 2015 that was the first

1 Rafał Piotr Paradowski, Uniwersytet Humanistyczno-Przyrodniczy im. Jana Długosza w Częstochowie, ORCID: 0000-0001-6267-2652

such large test of communication possibilities in the internet political discourse.

2. Twitter as a tool for political polls

Twitter is an American microblogging service and social network where registered users can post short, publicly accessible text and image messages known as "tweets". The website allows various interactions between users. The service was launched in 2006, and today is considered an opinion-forming forum where social and political issues are eagerly discussed (Clement, 2020). Twitter allows to freely share opinions almost in real time, which along with a large number of daily users of 152 million – as of the last quarter of 2019 (Clement, 2020) – and over 500 million posts every day – as estimated for June 2018 (Chandio, Sah, 2020) – make it the most reliable platform for any social research based on content analysis (Chandio, Sah, 2020; Rodak 2017). In addition, Twitter is one of the most important tools for communicating with public opinion. Researchers emphasize its importance for shaping the image both in long-term processes at the level of organizations and public figures (Adamik-Szysiak, 2014; Belford, Greene, & Cross, 2016), and in the case of specific social campaigns (Oliński, Szamrowski, 2019). The service is also used as a source of information for journalists while the users' content is utilized for quoting in press articles, radio and TV broadcasts (von Nordheim, Boczek, Koppers, 2018).

The popularity of Twitter amongst Polish society is lower compared to many Western European countries. The average monthly number of active users (as of October 2020) was: in Poland – 1.47 million, in Ireland – 1.9 million, in Sweden – 2.05 million, in Germany – 6.08 million, in Spain – 9.02 million, and in Great Britain – 20.24 million (Degenhard J., 2020a).

3. Related works

3.1. Ontological perspective

Among the many research questions available on Twitter, the most important are those related to electoral attitudes. They are mainly aimed at creating analytical models that enable predicting election results with reliability higher than traditional polls. The majority of these studies are focused on general elections: parliamentary or presidential, and mainly English-language tweets are analysed (Jain, Kumar, 2017, p. 21; Chauhan, Sikka, Sharma, 2020, pp. 10–15).

An important theoretical supplement to the above researches are those studies undertaken in order to better understand the phenomena of internet political discourse, based on knowledge about the communication processes and political marketing practices (Adamik-Szysiak, 2014; Belford, Greene, & Cross, 2016; Gorwa, 2017). One of those studies indicated the low autonomy of users in creating electoral opinions (Atluri et al., 2017) while another proved the possibility of discovering the intention of the sender – the politician – by exploring the language of their messages (Breeze, 2020; Johnson, Jin, Goldwasser, 2017). To compare, tweets were treated as a source of valid data in research on social campaigns and their impact on audience reactions (Oliński, Szamrowski, 2019) or in the studies related to the distribution and influence of disinformation on actual political and social attitudes, also in the context of public safety (Chandio, Sah, 2020; Colliver et al., 2018).

The findings of many researchers referring to predicting elections on Twitter indicate some methodological and analytical limitations, obstacles or untapped opportunities, which should be considered:

- awareness of the lack of knowledge about the demographic structure of users in

the context of interpreting results and, in addition, frequent omission of the importance of neutral votes (Gayo-Avello, Metaxas, & Mustafaraj, 2011),

- the role of incumbency on stimulating "obvious" predictions, using clear definitions of what is a "vote" and justifying the degree of credibility of the data collected (Gayo-Avello, 2012),
- the existence of the phenomenon called "vocal minority & silent majority", which stands for the relative small group of users who may dominate the stream with their content and thus disproportionately influence public opinion (Furnas, 2012; Gayo-Avello, 2012),
- using user normalization (one Twitter profile = one vote) and preferring analysis only of leading candidates and parties as more predictable (Salunkhe et al., 2017),
- the importance of selecting appropriate keywords and hashtags (Jain, Kumar, 2017),
- possible impact of events taking place during the campaign on the attitudes of voters (Xie, Liu, Wu , Tan, 2018),
- selection of user accounts responsible for irrelevant, fraudulent and spam content, as well as the use of any available optional data that may help to monitor changes in user behaviour (e.g. geolocation data, URL links) (Chauhan, Sharma, Sikka, 2020).

3.2. Review of techniques

Research of Twitter data in general is focused on text analysis. The most commonly used research techniques are:

- counting the number of occurrences of a given word, phrase, hashtags, users etc. in the entire database (volumetric approach), which is to reflect the popularity of a given entity and mapping priorities in communication system,
- simplified sentiment analysis, which allows to classify the entire user's statement, most often in separate categories: positive, negative or neutral,
- complex sentiment analysis or topic modelling, done by identification and categorization, that usually comes down to: 1) assigning labels (codes) to individual content and users in accordance with a set pattern (supervised learning) or to 2) inductive grouping of data showing similarity (unsupervised learning),
- analysing the Twitter communication network, e.g. through observed profiles or repeated interactions with individual users.

The above techniques can be used simultaneously in various configurations, with the allowed combination of quantitative and qualitative tools. In addition, they can be supported by lexicon-based approach or more complex natural language processing tools using machine learning algorithms (Deho et al., 2018; Kharde, Sonawane, 2016). In the context of related studies on voters' attitudes, all methodological trending models can be grouped into two dominant categories: sentiment analysis or volumetric approach combined with sentiment analysis, and the other: that stand for standalone or mixed types of volumetric and social network analysis (Chauhan, Sikka, Sharma, 2020, p. 19).

4. The method

4.1. Database construction

Against the background of all the 2014-2015 general elections, the 2015 presidential campaign was characterised by a particular interest from Twitter users and significant engagement from candidates in communication efforts. The most numerous and easiest to search for were the tweets published immediately before the second round

of the presidential campaign 2015, when opinions were crystallising and split between two candidates. The research was therefore limited to tweets mentioning Bronisław Komorowski (the incumbent president) and Andrzej Duda (the challenger).

The study did not predict, but only allowed to check the relationship between the analysed electoral attitudes and the election results post factum. The data was gathered among 2018-02-24 and 2018-05-01 using Twitter's advanced search functionality. Tweets covered the period 2015-05-15 – 2015-05-22, the final days of the presidential campaign, which included two televised debates. A total of over 70.000 tweets in Polish were collected using eight combinations of queries containing only surnames or candidates' first and second names. The three most complete and best quality datasets were selected (Table 1), consisting of 10.074 entries, which after cleaning, merging and removing duplicates yielded 8698 unique tweets attributed to 3508 users. Graphics were replaced by links, but these were not taken into account during the analysis.

Table 1. Datasets selected to analysis.

Dataset ID	Query type	Searched words	Language limitation	No. of tweets
K-H-2	hashtags	#komorowski	PL	2328
D-H-2	hashtags	#duda	PL	2687
DK-A-4	all of words	duda komorowski	PL	5059

Source: the author's own study.

4.2. Semantic analysis and labelling

First, a random review of the data was conducted to develop a coding system and train the researcher in semantic analysis. A pilot study was also performed at this stage to verify the accuracy of the coding. A sample of 99 tweets was independently labelled by the researcher and 23 testers (not trained on sample tweets earlier). The degree of coding inconsistency was only 12.52% – understood as the assignment by testers the labels opposite to the researcher (e.g. "for" instead of "against"). The non-compliant results were analysed in detail in order to improve the algorithms used in labelling subsequent posts.

Eventually, the following list of labels was established: 1) for Komorowski, 2) against Komorowski, 3) for Duda, 4) against Duda, 5) unspecified (opinion impossible to classify), 6) spam, vulgarisms and other excluded tweets (e.g. betting, polls not supported by own opinion, entries denying the sense of the elections, inexplicit advertisements, accidental foreign language tweets), 7) tweets of press agencies, news services, organisations or institutions.

Coding of the entire database was carried out manually by one trained researcher in two phases. In the first stage, a single label was assigned to each tweet. During the work, simple algorithms were created to semi-automatically search and code the remaining posts (e.g. according to repetitive content or phrases). In the second phase, each user's results were generalised to single codes that best identified their contributions (one user had only one vote). Tweets by politicians and journalists, as long as they originated from their personal profiles, were classified as content of ordinary citizens. Entries generated from business accounts were treated in a similar way – only if the content was a personal opinion. The two-phase coding resulted in two sets of data for analysis: 8698 categorised tweets and 3508 users with assigned attitudes towards the candidates.

4.3. Word count

Research also included a technique known as "word count" (likewise "word frequency", "share volume of tweets"). It was mainly based on counting the occurrences of candidates' names: "Komorowski" and "Duda". No other variations of words (declensions) were used, leaving them in their basic form. The capitation of the words was not significant. Although, it was allowed to count the occurrences of these names within the hashtags. The assumption was made that a supporter of a given candidate would use his or her name in the nominative form.

5. Results

5.1. Creditability

Effort was made to ensure that content of a spam nature, vulgarities, tweets from news agencies or potential trolls do not have a significant impact on the presented research results. Their share was determined at 18.43% in the set of tweets and 11.69% in the set of users (table 2). In the first group undesired entries were classified collectively as spam (without details). But in the second coding phase, two types of accounts were distinguished within that category: 1) media and related organisations, 2) other excluded.

In the first group 60.68% of the data had ascribed electoral preferences; another 20.89% consisted of entries with ambiguous opinions about the candidates. In the user set, on the other hand, clear electoral preferences were assigned to 61.23% of profiles. Accounts for which it was not possible to determine attitudes towards any of the candidates accounted to 27.08% of cases. This category increased by more than six percentage points to the corresponding set of tweets, because some users demonstrated ambivalent attitudes by publishing tweets about one candidate or similar opinions about both politicians.

Table 2. Distribution of labels in tweets and users.

	Set of tweets (after 1st stage of coding)		Set of users (after 2nd stage of coding)	
(Code) Label	**n**	**%**	**n**	**%**
(1) for Komorowski	960	11.04	550	15.68
(2) against Komorowski	1483	17.05	440	12.54
(3) for Duda	1151	13.23	669	19.07
(4) against Duda	1684	19.36	489	13.94
(5) unspecified	1817	20.89	950	27.08
(6) spam & other excluded	1603	18.43	234	6.67
(7) media & organisations			176	5.02
TOTAL	8698	100.00	3508	100.00

Source: the author's own study.

5.2. Tweets timeline & users activity

In the last week of the election campaign two TV debates were organised, which significantly stimulated the discussion around the candidates. They took place on 17th May and 21st May 2015. On the days of the debates and immediately after them, there

was a dynamic increase in engagement in Twitter discussions (Table 3). The first debate was accompanied by a slightly higher number of active users (+33.72%), their tweets (+26.45%) and retweets (+12.45%) on the day of the event. In contrast, the second confrontation was characterised by a higher number of responses (+11.70%) and a large decrease in the number of likes generated (-40.04%). The ratio of replies to the number of tweets was respectively: 0.52 for the period 17-18 May and 0.68 on 21-22 May. It can be assumed that likes are the form of activity that requires the least commitment from users (apart from mere browsing); the opposite of this would be having to reply or comment on a status.

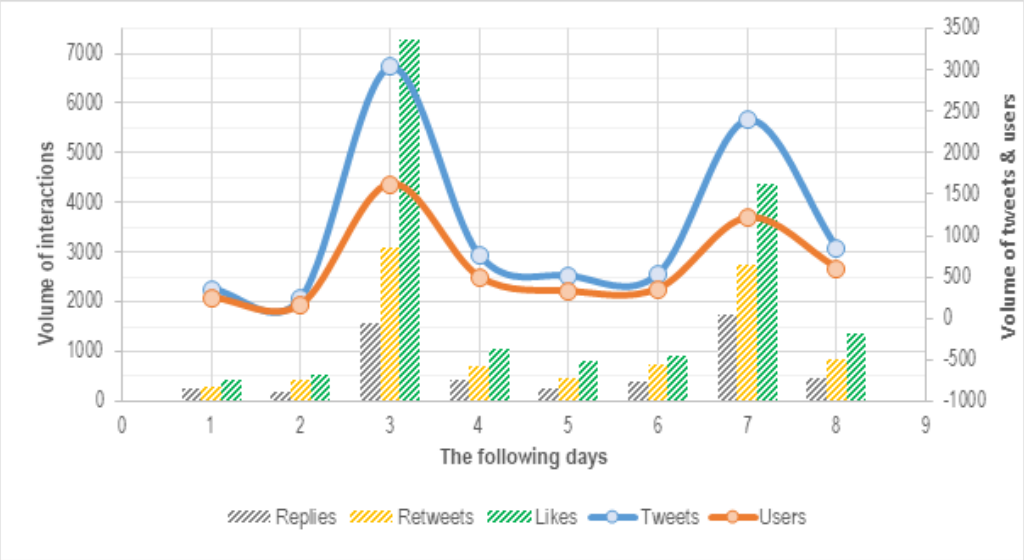

Figure 1. Timeline of tweets and users' interactions.
Source: the author's own study.

5.3. Vocal minority, silent majority

Due to the study of a thematically and time-limited data sample, it is impossible to differentiate between the categories of users who tweet little, on average and a lot. Despite this, after excluding spam and media content some patterns have been observed (Table 3):

1. Accounts publishing only 1-4 posts in the indicated period, which corresponded to 91.63% of users, were responsible for generating 55.72% of all tweets (Table 4).

2. Users who might be suspected of conducting deliberate mass communication activities on Twitter generated a disproportionate number of tweets, but were a small group. For example, 16.04% of the content (1171 tweets) in the sample came from only 31 accounts (1% of users) that published with a frequency of more than 20 entries.

It should be also noted that accounts publishing the most were those classified collectively as media & organisations, and which were excluded from this part of the analysis. Users from that group tweeted most frequently, and their activity was almost three times higher (6.21) than that of other users (2.37).

Table 3. Distribution of tweets by users and volume.

Tweets per period	Users			Volume of tweets		
	n	%	% cum.	n	%	% cum.
1 - 2	2496	80.57	80.57	2986	40.90	40.90
3 - 4	326	10.52	91.09	1082	14.82	55.72
5 - 6	103	3.32	94.42	553	7.57	63.29
7 - 8	53	1.71	96.13	388	5.31	68.61
9 - 10	34	1.10	97.22	323	4.42	73.03
11 - 15	35	1.13	98.35	450	6.16	79.19
16 - 20	20	0.65	99.00	348	4.77	83.96
21 - 30	17	0.55	99.55	433	5.93	89.89
31 - 50	10	0.32	99.87	410	5.62	95.51
51 - 98	4	0.13	100.00	328	4.49	100.00
TOTAL	3098	100.00	–	7301	100.00	–
EXCLUDED:						
(6) spam & other	234	–	–	304	–	–
(7) media & organisations	176	–	–	1093	–	–

Source: the author's own study.

5.4. Engagement and interactions of users.

By comparing the measures of position and dispersion of data by users assigned to the seven categories, it can be observed that (Table 4):

1. Opponents of a candidate tweeted slightly more often (2.89–2.95) than his supporters (2.28–2.59).

2. The supporters of Duda were in the majority (19.07%), although they published less frequently and fewer than the supporters of Komorowski (in ratio 2.28:2.59).

3. Media and organisation-related accounts were the most active, although their activity varied the most (the largest standard deviation and high average tweets per user), which depended on the specific entity/author.

Table 4. Distribution of labelled users followed by statistical measures.

(Code) Label	Classified users	% of classified users	Mean tweet frequency	Median	Standard deviation	Minimum tweets per user	Maximum tweets per user
(1) for Komorowski	550	15.68	2.59	1	4.37	1	60
(2) against Komorowski	440	12.54	2.95	1	7.44	1	96
(3) for Duda	669	19.07	2.28	1	2.71	1	20
(4) against Duda	489	13.94	2.89	1	6.29	1	98

(Code) Label	Classified users	% of classified users	Mean tweet frequency	Median	Standard deviation	Minimum tweets per user	Maximum tweets per user
(5) unspecified	950	27.08	1.72	1	2.89	1	49
(6) spam & other excluded	234	6.67	1.30	1	1.22	1	17
(7) media & organisations	176	5.02	6.21	2	13.63	1	153
TOTAL	3508	100.00					

Source: the author's own study.

The distribution of interactions according to the code assigned to the tweet allowed the behavioural characteristics of all users participating in this discourse to be expanded. As part of this analysis, an auxiliary measure was introduced in the form of the average number of interactions per blog entry within a particular label (abbr. A-NI/T). It is important to note that:

1. The distribution of likes attributed to tweets indicating support or opposition to a particular candidate (codes 1-4) was similar to the voting result.

2. In the set of tweets, the largest number of interactions (35.16%) was generated by entries classified collectively as spam and media, having the highest ratio of interactions per tweet (A-NI/T index: 7.09).

3. Blog entries criticising Komorowski obtained more interactions per tweet than those opposing Duda (in a ratio of 3:2). At the same time, the first one had almost twice as many retweets (19.04:9.95), which may indicate that internet users were more willing to share negative opinions about the incumbent president.

4. The contrary tendency was observed among the positive tweets; there were fewer of retweets for Komorowski and more for Duda (9.73:16.04).

5. Tweets praising the incumbent president generated more polemics as opposed to entries arguing for his rival (14.1:9.3).

6. The smallest total number of interactions had "unspecified" tweets and those entries directed against Duda.

Table 5. Interactions by labelled tweets.

(Code) Label	Set of tweets								
	Replies		Retweets		Likes		TOTAL		A-NI/T
	n	%	n	%	n	%	n	%	
(1) for Komorowski	744	14.08	907	9.73	2265	13.50	3916	13.50	4.08
(2) against Komorowski	486	9.20	1774	19.04	2301	13.72	4561	13.72	3.08
(3) for Duda	491	9.29	1495	16.04	2425	14.46	4411	14.46	3.83
(4) against Duda	599	11.34	927	9.95	1920	11.44	3446	11.44	2.05
(5) unspecified	703	13.31	1010	10.84	1966	11.72	3679	11.72	2.02
(6) spam & other excluded	2260	42.78	3205	34.40	5899	35.16	11364	35.16	7.09
(7) media & organisations									
TOTAL	5283	100.00	9318	100.00	16776	100.00	31677	100.00	3.6 (average)

Source: the author's own study.

The above analysis was followed by an exploration of interactions in the set of

users. Analogous to the previous set, an auxiliary measure was also introduced – the average number of interactions generated by one user's tweets within a particular label (abbr. A-NI/U). This mathematical construct had indirect character, but allowed to illustrate the degree of audience involvement in the content published from specific types of accounts. The observations are as follows:

1. The disparity between the numbers of responses gathered in total by the entries of users declaring support for a particular candidate widened in relation to the same categories in the set of tweets and amounted to 1.9:1 (26.1:13.99).

2. In general, users were more than twice as likely to share tweets from accounts criticising Komorowski than those hostile to Duda (ratio 17.15:7.95).

3. The most interactions per user were obtained by accounts labelled as "media & organisations" (A-NI/U index: 34.99), exceeding the average by almost four times.

4. Accounts categorised as supporters of the incumbent president also achieved an above-average engagement rate (>8.9).

5. In contrast, the least number of interactions per user were observed among content generated by accounts with an undefined voting stance (A-NI/U index: 3.97).

6. In terms of the number of profiles criticising each of the candidates, the number of interactions (A-NI/U) associated with the total content of the critics of the incumbent president was nearly 50% higher than that of Duda's opponents (ratio 8.69:5.8).

Table 6. Interactions according to content generated by labelled users.

(Code) Label	Set of tweets								
	Replies		Retweets		Likes		TOTAL		A-NI/U
	n	%	n	%	n	%	n	%	
(1) for Komorowski	1379	26.10	1653	17.74	4748	28.30	7780	28.30	14.15
(2) against Komorowski	407	7.70	1598	17.15	1818	10.84	3823	10.84	8.69
(3) for Duda	739	13.99	1903	20.42	3180	18.96	5822	18.96	8.70
(4) against Duda	465	8.80	741	7.95	1629	9.71	2835	9.71	5.80
(5) unspecified	638	12.08	1025	11.00	2105	12.55	3768	12.55	3.97
(6) spam & other excluded	184	3.48	420	4.51	586	3.49	1190	3.49	5.09
(7) media & organisations	1471	27.84	1978	21.23	2710	16.15	6159	16.15	34.99
TOTAL	5283	100.00	9318	100.00	16776	100.00	31377	100.00	8.9 (average)

Source: the author's own study.

5.5. Political attitudes of users and the real election results

According to the official results of the presidential election, Andrzej Duda won with 51.55% of the valid votes. Bronisław Komorowski got 48.45% of the valid votes. In turn, invalid votes accounted for 1.47% and the turnout was 55.34%.

5.5.1. Word count

Using the "word count" technique, it was established that the word "Komorowski" appeared 6843 times and "Duda" 7223 times in the entire database. The percentage of "votes" counted in this way was 48.65% to 51.35% in favour of the challenger. It was very close to the official election result (offset 0.2). However, this analysis included all tweets, i.e. those coded as unspecified, excluded and authored by news agencies. After limiting the collection to only relevant entries for or against a particular candidate (5207 tweets),

the differences in politicians' support widened even further (offset 0.36) – the word "Komorowski" occurred 3884 times (48.09%) and "Duda" 4192 (51.91%).

Table 7. Share volume of words – surnames of candidates.

Searched words	Frequency of words				Official results
	Set of tweets w/o exclusions n = 8698		Set of tweets with exclusions n = 5207 (coded as 1–4)		
	n	%	n	%	%
Komorowski	6843	48.65	3884	48.09	48.45
Duda	7223	51.35	4192	51.91	51.55
TOTAL	14066	100.00	8076	100.00	100.00

Source: the author's own study.

5.5.2. Positive/negative ratio

Juxtaposing the number of opposing opinions about a particular candidate, separately in the sets of tweets and users, allowed verification of previously observed correlations. Key findings from this analysis include:

1. In the sets of tweets and users, the higher ratio of positive to negative opinions was always obtained by Duda. Moreover, the relationship between these indicators of both candidates (0.65:0.68 and 1.25:1.37) was very close to the distribution of votes in the election (48.45:51.55).

2. In the user group, Duda attracted significantly more of his own positive electorate (669) than votes from those declaring only dislike for the incumbent president (440). In turn, Komorowski gained a similar number of univocal supporters of his own candidacy (550) as those with a negative attitude towards Duda (489).

3. The volume of negative tweets against Duda (1684) and the number of users identified as his critics (489) were higher than the corresponding values for Komorowski (1483 and 440 respectively).

Table 8. Positive/negative ratio.

	Opinions in set of tweets				Opinions in set of users			
Candidate	positive	negative	total	pos / neg	positive	negative	total	pos / neg
Komorowski	960	1483	2443	0.65	550	440	990	1.25
Duda	1151	1684	2835	0.68	669	489	1158	1.37
TOTAL	2111	3167	5278	-	1219	929	2148	-

Source: the author's own study.

5.5.3. Votes for, against & non-voters

The labels assigned to individual profiles, illustrating the electoral attitude of a user, were summed up in two categories (table 9). The number of votes thus obtained was 48.37% for the incumbent president and 51.63% for the newcomer, and was very close to the official election outcome (offset less than 0.1).

At the same time, an attempt was made to calculate attendance based on the set

of users. Proposed formula excluded data classified as "media & organisations" from the calculation. However, it included accounts with unspecified political views and users responsible for spam in the group of non-voters. The turnout calculated in this way was far from real – the difference between them amounted 9.13 percentage points.

Table 9. Twitter votes count vs. real election outcome.

Vote for	Formula	n	%	Official results [%]
Komorowski	(1) for Komorowski & (4) against Duda	1039	48.37	48.45
Duda	(3) for Duda & (2) against Komorowski	1109	51.63	51.55
Total voters		2148	100.00	-
Non-voters	(5) unspecified & (6) spam and other	1184	-	-
Voter turnout [%]	$\frac{\sum((1),(2),(3),(4))}{\sum((1),(2),(3),(4),\ (5),\ (6))} \times 100\%$	64.47		55.34

Source: the author's own study.

Further investigation of the set of users showed that the most reliable distribution of behaviour occurred among those accounts that tweeted no more than 20 times in the sample (table 10). They represented 99% of users and were responsible for almost 84% of all content (after excluding spam and media accounts; see table 3). Based on those findings, a closer look was taken at this group. The data in the set of users were organised into 4 inseparable categories with tweet frequency: 1-4, 1-10, 1-20 and >21. This analytical model showed that the subgroup publishing the most entries behaved extremely differently from the others, however, it was too small (1% of users) to significantly affect the result calculated on the basis of the set of users. Nevertheless, mentioned subgroup of profiles generated 16.04% of the content, which could strongly influence the election analysis based solely on the volume of tweets. A simple comparison of the number of entries (according to analogous formulas as in table 9) gave a very even result, with a marginal advantage of the incumbent president (2644 tweets – 50.09%) over Duda (2634 – 49.91%).

Table 10. Twitter votes count depending on tweet rate and volume of tweets.

Vote for	Official results [%]	Set of users								Set of tweets	
		tweet rate 1-4		tweet rate 1-10		tweet rate 1-20		tweet rate: >21		n	%
		n	%	n	%	n	%	n	%		
Komorowski	48.45	915	47.98	1000	48.24	1022	48.14	17	68	2644	50.09
Duda	51.55	992	52.02	1073	51.76	1101	51.86	8	32	2634	49.91
Total voters	–	1907	100.00	2073	100.00	2123	100.00	25	100.00	5278	100.00
Non-voters	–	1146	–	1172	–	1178	–	6	–	–	–
Voter turnout [%]	55.34	62.46		63.88		64.31		–		–	

Source: the author's own study.

5.5.4. Likes as voting predictors

Also an in-depth analysis was carried out related to the possibility of using likes as an indicator of electoral outcome (table 11). The results derived from the first formula were more internally reliable and each time gave victory to Duda. The dispersion from the election outcome ranged from -1.2 to +3.1 percentage points. On the other hand, the calculations made in accordance with the second formula resulted in greater variation in general. However, under the former model, there were two results correlating very closely with official outcome with offsets: 0.16 (in set of tweets) and 0.28 (among users publishing 1-4 tweets).

Table 11. Distribution of likes as votes for a given candidate.

Vote for	Formula	Official results [%]	Set of users						Set of tweets	
			tweet rate 1-4		tweet rate 1-10		tweet rate 1-20		n	%
			n	%	n	%	n	%		
Komorowski	(1) for K. & (4) against D.	48.45%	2157	49.64	2870	46.65	3323	45.35	4185	46.96
Duda	(3) for D. & (2) against K.	51.55%	2188	50.36	3282	53.35	4005	54.65	4726	53.04
Komorowski	(1) for K.	48.45%	1654	48.17	2025	42.43	2130	40.11	2265	48.29
Duda	(3) for D.	51.55%	1780	51.83	2748	57.57	3180	59.89	2425	51.71

Source: the author's own study.

6. Conclusions

Analyses conducted on a data sample proved that there was a strong correlation between political attitudes presented by Twitter users and the election outcome. Both methods, semantic analysis and word count, yielded similar results and gave the victory to Duda, which indeed happened. The offset to official post-election score amounted less than 0.1 using the first technique and 0.2 with the second one. At the same time, it was not possible to reproduce the correct voter turnout – estimated between 62.46% and 64.47% (actual turnout 55.34%). It should be stressed, however, that it is much easier to conceptualise a supporter or opponent of a politician than it is to state the conditions that must be fulfilled by a person who refrains from casting any vote.

Study also showed that both TV election debates resulted in different user activity within the collected data. The first debate evoked very strong engagement but less controversial reactions among the sampled authors and audience. The second confrontation was characterised by a higher number of responses (+11.70%) with large decrease in the likes generated (-40.04%) that might mirrored some change in the discourse by provoking more discussion instead of the typical, less involving reactions.

Analyses confirmed the existence of a vocal minority and silent majority. Investigation showed that the most reliable distribution of behaviour occurred among those accounts that tweeted no more than 20 times in the sample (after excluding spam & media profiles). The electoral results obtained from these users were characterised by a very low offset: 0.21–0.47. This group represented 99% of the accounts, and was responsible for almost 84% of the content. The remaining 1% of users behaved extremely differently from the rest (due to twice as many supporters of Komorowski). Their activity would have been responsible for distorting the result of a calculation based only on counting labelled tweets, giving a marginal victory to the incumbent president (50.09%).

Also the volume and distribution of interactions of the users in general proved to be helpful in the study. It transpired that the analysis of likes has always given victory to Duda. Although, the best results were obtained under the codes indicating support for

particular candidates in the set of tweets (offset 0.16) and in the set consisting of accounts publishing 1–4 entries (offset 0.28). In addition, it was possible to capture other differences in how users responded to tweets for or against a particular candidate. For example, users were more than twice as likely to share tweets from accounts criticising Komorowski than those hostile to Duda. Entries praising the incumbent president generated around 50% more polemics as opposed to tweets arguing for his rival. The smallest volume of interactions had "unspecified" content and those entries directed against Duda. On the contrary, accounts categorised as supporters of the incumbent president achieved an above-average engagement rate.

The survey also used the common indicator: positive/negative ratio. The relationship between these indicators of both candidates was very close to the distribution of votes in the actual election (offset 0.45 and 0.75). It was proved that, in the set of users, Duda attracted significantly more of his own positive electorate than votes from those declaring only dislike for Komorowski. Moreover, in both datasets, the challenger obtained more positive entries and votes from his followers than the incumbent president.

7. Discussion

The adopted methodological strategy highlighted two issues. The first was the role of a pilot study – the construction of the coding system and the primary algorithms for the classification. The second was the need to continually expand the dictionary with non-obvious hashtags, repetitive phrases or shared URLs that spoke clearly for or against a particular candidate. The analytical model allowed modification of codes, which increased the accuracy of tweet classification.

Another key to the success of such studies seems to be the use of effective techniques to eliminate irrelevant and "polluted" content. At the same time, it is worth considering excluding the vocal minority from the analyses, and focusing on the silent majority. In defining these two groups, it may be helpful to refer to the frequency of entries published by a single user. Moreover, examining the structure of general interactions in tweets can further improve the validity of the analysis. Special attention should be paid to the likes.

It is certain that the political discourse on Twitter is not conducted in the same way all the time. The study focused only on the last few days of presidential campaign and was not able to gather or process all the dynamics that could occur in long-term. Despite this, it was possible to observe the impact of television campaigns on users' activity. In the future, it would be worth to learn about possible changes in users' attitudes over time. The study also lacked an analysis of users openly contesting the elections.

The study had its weaknesses. First of all, the database constituted only a sample of the content published as part of the pre-election discourse. It may not have covered those tweets that had been possibly deleted by the time they were retrieved for the study. Secondly, analytical models used were often based on excluding parts of the data or restricting it to certain categories. From the initial 70.000 tweets, only 8698 unique tweets assigned to 3508 users were selected for classification. Some observations were based on smaller subgroups. Although such procedures allowed for an in-depth, more qualitative look at the studied phenomena; at the same time, they raised doubts over the representativeness of such. Thirdly, the study was limited to only two candidates in the final phase of the election campaign. It is uncertain whether presented methodological approaches would work if used with more numerous research object or other types of elections.

References

ADAMIK-SZYSIAK, M. (2014). Twitter in Communication Strategies of the Leaders of the Polish Political Par-

ties. Kwartalnik Naukowy OAP UW „e-Politikon", 9, 109–131.

ATLURI, V., CHUN, S. A., VAIDYA, J., YAQUB, U. (2017). Analysis of political discourse on twitter in the context of the 2016 US presidential elections. Government Information Quarterly, 34(4), 613–626. https://doi.org/10.1016/j.giq.2017.11.001

BELFORD, M., GREENE, D., & CROSS, J. P. (2016). Tweeting Europe: A text-analytic approach to unveiling the content of political actors' Twitter activities in the European Parliament. 6th Annual General Conference of the European Political Science Association (EPSA'16), 44.

BREEZE, R. (2020). Exploring populist styles of political discourse in Twitter. World Englishes, 39(4), 550–567. https://doi.org/10.1111/weng.12496

CHANDIO, M. M., SAH, M. (2020). Brexit Twitter Sentiment Analysis: Changing Opinions about Brexit and UK Politicians. In: L. C. Jain, S.-L. Peng, B. Alhadidi, S. Pal (Eds.), Intelligent Computing Paradigm and Cutting-edge Technologies (V. 9, p. 1–11). Springer International Publishing. https://doi.org/10.1007/978-3-030-38501-9_1

CHAUHAN, P., SHARMA, N., & SIKKA, G. (2020). The emergence of social media data and sentiment analysis in election prediction. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-020-02423-y

CLEMENT, J. (2020, July 24). Twitter: most users by country. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

COLLIVER, C., POMERANTSEV, P., APPLEBAUM, A., & BIRDWELL, J. (2018). Smearing Sweden. International Influence Campaigns in the 2018 Swedish Election.

DEGENHARD, J. (2020, October 12). Twitter users in Europe 2020, by country. https://www.statista.com/forecasts/1168954/twitter-users-in-europe-by-country

DEHO, O. B., AGANGIBA, W. A., ARYEH, F. L., & ANSAH, J. A. (2018). Sentiment Analysis with Word Embedding. 2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST), 1–4. https://doi.org/10.1109/ICASTECH.2018.8506717

FURNAS, A. (2012). You Can't Use Twitter to Predict Election Results. The Atlantic, 5.

GAYO-AVELLO, D. (2012). „I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"—A Balanced Survey on Election Prediction using Twitter Data. ArXiv:1204.6441 [Physics]. http://arxiv.org/abs/1204.6441

GAYO-AVELLO, D., METAXAS, P., & MUSTAFARAJ, E. (2011). Limits of Electoral Predictions Using Twitter. ICWSM.

GORWA, ROBERT. (2017). Computational Propaganda in Poland: False Amplifiers and the Digital Public Sphere. Computational Propaganda Project Working Paper Series. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Poland.pdf

JAIN, V. K., KUMAR, SH. (2017). Towards Prediction of Election Outcomes Using Social Media. International Journal of Intelligent Systems and Applications, 9(12), 20–28. doi: 10.5815/ijisa.2017.12.03

JOHNSON, K. M., JIN, D., GOLDWASSER, D. (2017). Modelling of Political Discourse Framing on Twitter. Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017). https://www.cs.purdue.edu/homes/dgoldwas/downloads/papers/JJG_icwsm_2017.pdf

KHARDE, V. A., & SONAWANE, S. S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, 139(11), 5–15. https://doi.org/10.5120/ijca2016908625

LIU, R., YAO, X., GUO, C., & WEI, X. (2020). Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach. Annals of GIS, 1–14. https://doi.org/10.1080/19475683.2020.1829704

OLIŃSKI, M., SZAMROWSKI, P. (2019). The Use of the Twitter in Public Benefit Organisations in Poland: How Communicative Function of Tweets Translates Into Audience Reaction? Central European Economic Journal, 5(52), 10–24. https://doi.org/10.1515/ceej-2018-0009

RODAK, O. (2017). Twitter jako przedmiot badań socjologicznych i źródło danych społecznych: Perspektywa konstruktywistyczna. Studia Socjologiczne, 3(226), 209–236.

SALUNKHE, P., SURNAR, A., & SONAWANE, S. (2017). A Review: Prediction of Election Using Twitter Sentiment Analysis. International Journal of Advanced Research in Computer Engineering & Technology, 06(05), 723–725.

VON NORDHEIM, G., BOCZEK, K., KOPPERS, L. (2018). Sourcing the Sources: An analysis of the use of Twitter and Facebook as a journalistic source over 10 years in The New York Times, The Guardian, and Süddeutsche Zeitung. Digital Journalism, 6(7), 807–828. https://doi.org/10.1080/21670811.2018.1490658