Journal of Big Data

CrossMark

Predicting referendum results in the Big Data Era

Amaryllis Mavragani[1*] and Konstantinos P. Tsagarakis[2]

*Correspondence:
amaryllis.mavragani1@stir.
ac.uk
[1] Department of Computing
Science and Mathematics,
Faculty of Natural Sciences,
University of Stirling,
Stirling FK9 4LA, UK
Full list of author information
is available at the end of the
article

Abstract

In addressing the challenge of Big Data Analytics, what has been of notable significance is the analysis of online search traffic data in order to analyze and predict human behavior. Over the last decade, since the establishment of the most popular such tool, *Google Trends*, the use of online data has been proven valuable in various research fields, including -but not limited to- medicine, economics, politics, the environment, and behavior. In the field of politics, given the inability of poll agencies to always well approximate voting intentions and results over the past years, what is imperative is to find new methods of predicting elections and referendum outcomes. This paper aims at presenting a methodology of predicting referendum results using Google Trends; a method applied and verified in six separate occasions: the 2014 Scottish Referendum, the 2015 Greek Referendum, the 2016 UK Referendum, the 2016 Hungarian Referendum, the 2016 Italian Referendum, and the 2017 Turkish Referendum. Said referendums were of importance for the respective country and the EU as well, and received wide international attention. Google Trends has been empirically verified to be a tool that can accurately measure behavioral changes as it takes into account the users' revealed and not the stated preferences. Thus we argue that, in the time of intelligence excess, Google Trends can well address the analysis of social changes that the internet brings.

Keywords: Big data, Elections, Google Trends, Internet behavior, Nowcasting, Online behavior, Online queries, Politics, Prediction, Referendum, Voting

Introduction

Big Data are large data volumes characterized by the 8 Vs: 'Volume' [1], Variety, Velocity [2], Variability, Value, Volatility, Validity, and Veracity [3]. A popular way to access and use this vast amount of information is the analysis of online search queries [4, 5], with the most notable research on the subject taking into account data from *Google Trends* [6]; a tool that measures variations in the online interest. The validity of Google Trends' data has been undoubtedly shown [7], while many studies have suggested that they are valuable, accurate, and of high benefit for forecastings [8–13], predictions [14] and nowcastings [15], in analyzing online interest [8], and in decision making [3].

Google Trends as a tool is highly integrated in research, with notable contributions in the fields of health and medicine [16–23], economics and finance [4, 24–27], and the environment [28]. The use of Google Trends in political science in general and in predicting election and referendum outcomes in specific is showing promising results so

far. According to recent research [14], what can be observed is an increased use of online data in polls and measuring voting intention. As Internet penetration is significantly growing in western countries and the significance of the Internet cannot be doubted [29], the monitoring and analysis of online search traffic data can be proven valuable in nowcasting election and referendum races. Burnap et al. [30] suggest that changes in online behavior can be accurately depicted on Internet data, while political campaigners increasingly use social media and online platforms [31]. Up to this point, the field of predicting referendum results using online search traffic data has not been much explored, though Google Trends' data have been employed to predict the results of the 2015 Greek Referendum [14].

Finding new ways of predicting the outcome of elections and referendum is crucial, as the poll agencies have not always been successful in the recent past to well approximate voting intentions and results. During the short pre voting period of the critical 2015 Greek Referendum, we monitored the Google queries on the YES and NO keywords, so as to predict the voting intentions and the referendum outcome. Our results better approximated the official referendum outcome than poll agencies, the latter suggesting that the YES and NO difference was very small, or that YES was to win the race [14]. This method has also been applied to five other referendum races concerning crucial EU or constitutional matters and received international attention over the past years, namely the 2014 Scottish Referendum, the 2016 UK Referendum, the 2016 Hungarian Referendum, the 2016 Italian Referendum, and the 2017 Turkish Referendum.

This paper aims at presenting the methodology of how to predict and nowcast referendums with the use of online search traffic data; a methodology tested and validated in six occasions. This methodology aims at becoming a point of reference for the next generation of polling, while showing that Google Trends is a valid tool in nowcastings and predictions, with great potential. The rest of the paper is structured as follows: The "Research methodology" section consists of the method of predicting referendum results, while the "Results" section consists of the presentation of the results of six referendums over the past years (2014–2017) in the European Continent. The "Discussion" section consists of the discussion of the results, while the "Conclusions" section consists of our conclusions and future research suggestions.

Research methodology

Google Trends allows the monitoring of the change in the online interest in a term in a country or region over a selected time frame, e.g. a range of years, 1 year, 90 days, 30 days, 7 days, 4 h, 1 h, or a specified time period. It also allows the multiple relative comparisons of a term in different regions, or the comparison of various terms in one region, while a feature offers the opportunity to compare different terms in different regions. Data, depending on the time frame selected, are either monthly, weekly, hourly, or even of 1 min intervals.

Data from Google Trends are downloaded online in '.csv' format and are normalized over the selected period, so as to allow the easy comparison between the respective

examined terms. The adjustment of the data, as reported by Google, is as follows: "*Search results are proportionate to the time and location of a query: Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity. Otherwise places with the most search volume would always be ranked highest. The resulting numbers are then scaled on a range of 0 to 100 based on a topic's proportion to all searches on all topics. Different regions that show the same number of searches for a term will not always have the same total search volumes.*" [32].

In general, data from Google can be accurate for measuring the public's interest if the terms are carefully selected [33]. Preis et al. [5] for example, in developing the Future Orientation Index, only used Arabic numerals as the selected keywords, as they are universal, so spelling or translation errors would not produce bias. In a referendum, it is imperative that the terms selected are the ones to give an accurate result. Thus the wording of the referendum plays a significant role in predicting the outcome. In most cases, the question of a referendum is answered with a simple YES or NO, though, for example, in the 2016 UK Referendum, the wording did not allow for this, deeming this race more complicated to predict. In this case, the available answers were '*Remain a member of the European Union*' and '*Leave the European Union*,' so we selected the terms "Remain" and "Leave" in order to predict the result. For the rest of the examined cases, we used the translation of the YES and NO keywords in each respective language, that is 'YES' and 'NO' in Scotland, 'NAI' and 'OXI' in Greece, 'IGEN' and 'NEM' in Hungary, 'SI' and 'NO' in Italy, and 'EVET' and 'HAYIR' in Turkey. Note that the keyword choice is not case sensitive in Google Trends.

For predicting the referendums, we mainly use weekly and monthly data (daily intervals), apart from the case of Greece, where daily and hourly data were used due to the very short pre-voting period of 9 days. Apart from the 2014 Scottish Referendum and the 2015 Greek Referendum, data were downloaded from the field 'Campaigns and elections' in Google Trends that eliminates noisy data, i.e. hits that are not attributed to the examined event. If data in this category are not sufficient due to low interest on the subject (as, for example, in the Hungarian Referendum), the search can be widened to include hits in the field 'Politics'.

The time-frames for the data used in this study are as follows:

1. *Scottish Independence Referendum (2014)*: (a) "from September 10th to 17th, and (b) "from August 17th to September 17th".
2. *UK European Union Membership Referendum (2016):* (a) May 24th to June 23rd, and (b) June 16th to 23rd.
3. *Hungarian Migrant Quota Referendum (2016):* (a) from September 2nd to October 2nd, (b) from September 25th to October 2nd, (c) from October 1st at 9 a.m. to October 2nd at 9 a.m., and (d) on October 2nd from 4 to 8 p.m. (4-h).

4. *Italian Constitutional Referendum (2016):* (a) September 3rd to November 30th, (b) October 28th to November 26th, (c) November 19th to November 26th, and (d) November 26th to December 3rd.

5. *Turkish Constitutional Referendum (2017):* For France and Germany: (a) from March 9th to April 9th and (b) from April 2nd to April 9th, while for the Netherlands (a) from March 5th to April 5th, and (b) from March 29th to April 5th.

For the *i*-th set of data, i.e. monthly, weekly, 4-h, hourly, the hits' averages for the YES (and Remain) $\left(Y_{t_i}\right)$ and NO (and Leave) (N_{t_i}) keywords are calculated, and are then percentized as $Y_{t_{pi}} = \frac{Y_{t_i}}{Y_{t_i}+N_{t_i}}$ and $N_{t_{pi}} = \frac{N_{t_i}}{Y_{t_i}+N_{t_i}}$ for YES and NO, respectively, with $Y_{t_{pi}} + N_{t_{pi}} = 1$.
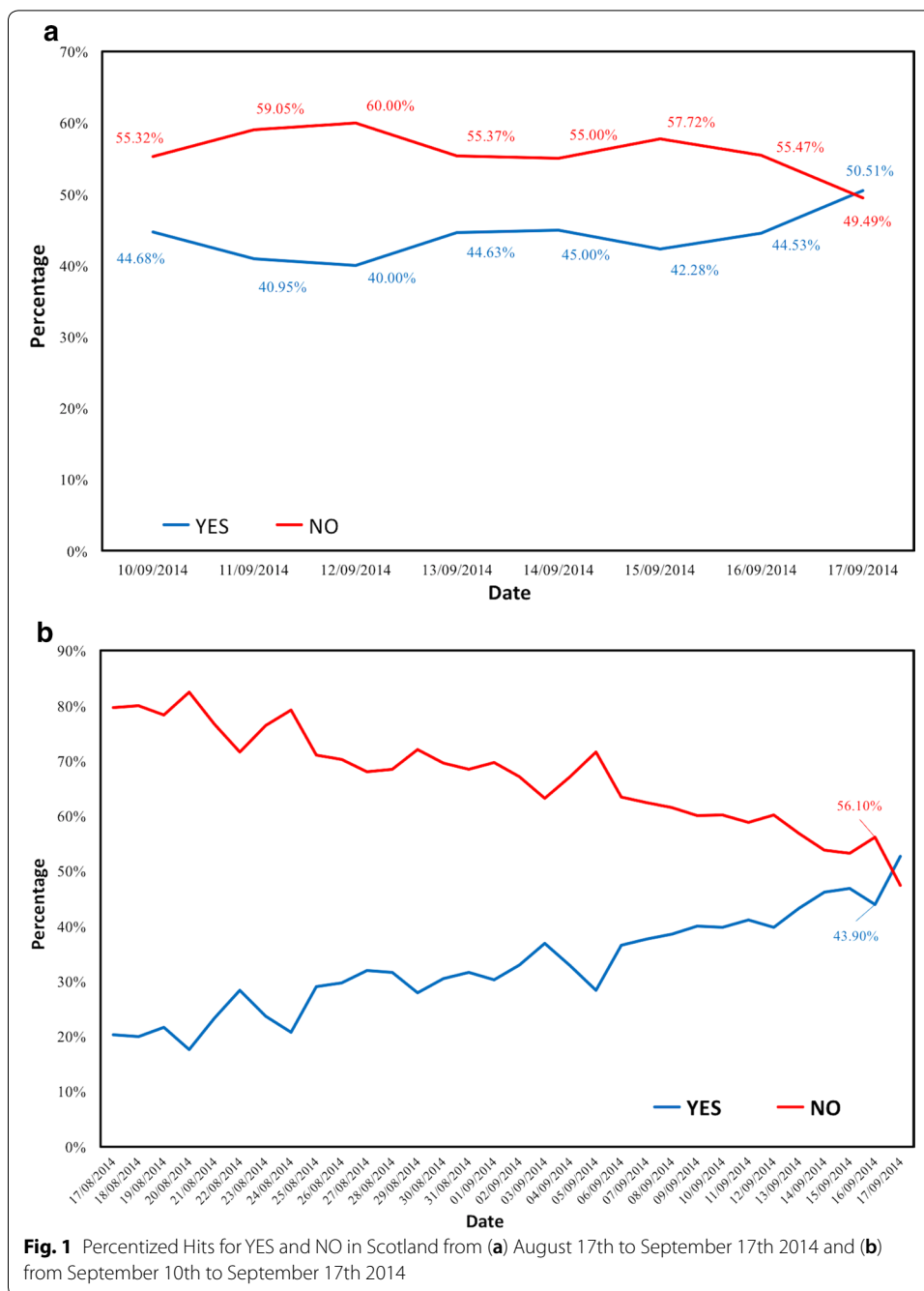
Results

This section consists of the results of six referendums in the European Continent over the last years (2014–2017), namely the 2014 Scottish Independence Referendum, the 2015 Greek Bailout Referendum, the 2016 UK European Union Membership Referendum, the 2016 Hungarian Migrant Quota Referendum, the 2016 Italian Constitutional Referendum, and the 2017 Turkish Constitutional Referendum.

Scottish independence referendum (2014)

On September 18th 2014, the Scottish people were asked to vote on whether or not they wished for Scotland to remain part of the United Kingdom or become an independent country. 84.6% of the eligible voting population showed up to cast their vote, with 55.3% voting for remaining in the UK, i.e. voted for NO, while YES received a 44.7% [34].

Most polls during the last month before the race suggested that NO was on the lead with about 5% difference, not counting the undecided [35–40], though the undecided percentage varied. Using data from Google Trends for the last week before the referendum race, i.e. from September 10th to 17th (Fig. 1a), the average percentized hits for YES were 44.07%, while the NO hits were at 55.93%. Though for the last month before the race, i.e. from August 17th to September 17th, the percentized average of the NO hits were at 64.47% (Fig. 1b), the percentized daily NO hits for the last two days before the race, i.e. on the 15th and 16th, were at 53.17% and 56.10%, respectively. As is evident, using Google Trends data a close approximation of the final referendum result was possible, with the percentized average of the hits for the last week before the referendum being almost the same as the official result.

Fig. 1 Percentized Hits for YES and NO in Scotland from (**a**) August 17th to September 17th 2014 and (**b**) from September 10th to September 17th 2014

Greek bailout referendum (2015)

In 2015, Mavragani and Tsagarakis [14] well approximated the voting intentions and results of the Greek Referendum on July 5th; a crucial referendum on the rescue plan proposed by the EU [41] that received wide national and international media attention [42]. Despite the very short pre-voting period of 9 days, by applying the proposed method we accurately predicted that NO was on the lead, despite that official voting intention polls suggesting that YES and NO were very close to one another, even at a 0.5% difference [43].

The NO vote officially received a total of 61.31% [44]. The percentized hits for NO in Google Trends from June 27th (20:00) to July 4th (20:00) were at 55.99%, while the NO hits from Saturday the 4th (20:00) to Sunday the 5th (20:00) were at 58.20%, as reported by Mavragani and Tsagarakis [14]. The poll agencies' results predictions varied for the NO vote: 54.50% [45], 54% [46], and 49% [47]. It is evident that the method of using percentized Google Trends' data better approximated the official results and this shows great potential in predicting and nowcasting referendum voting intentions and results.
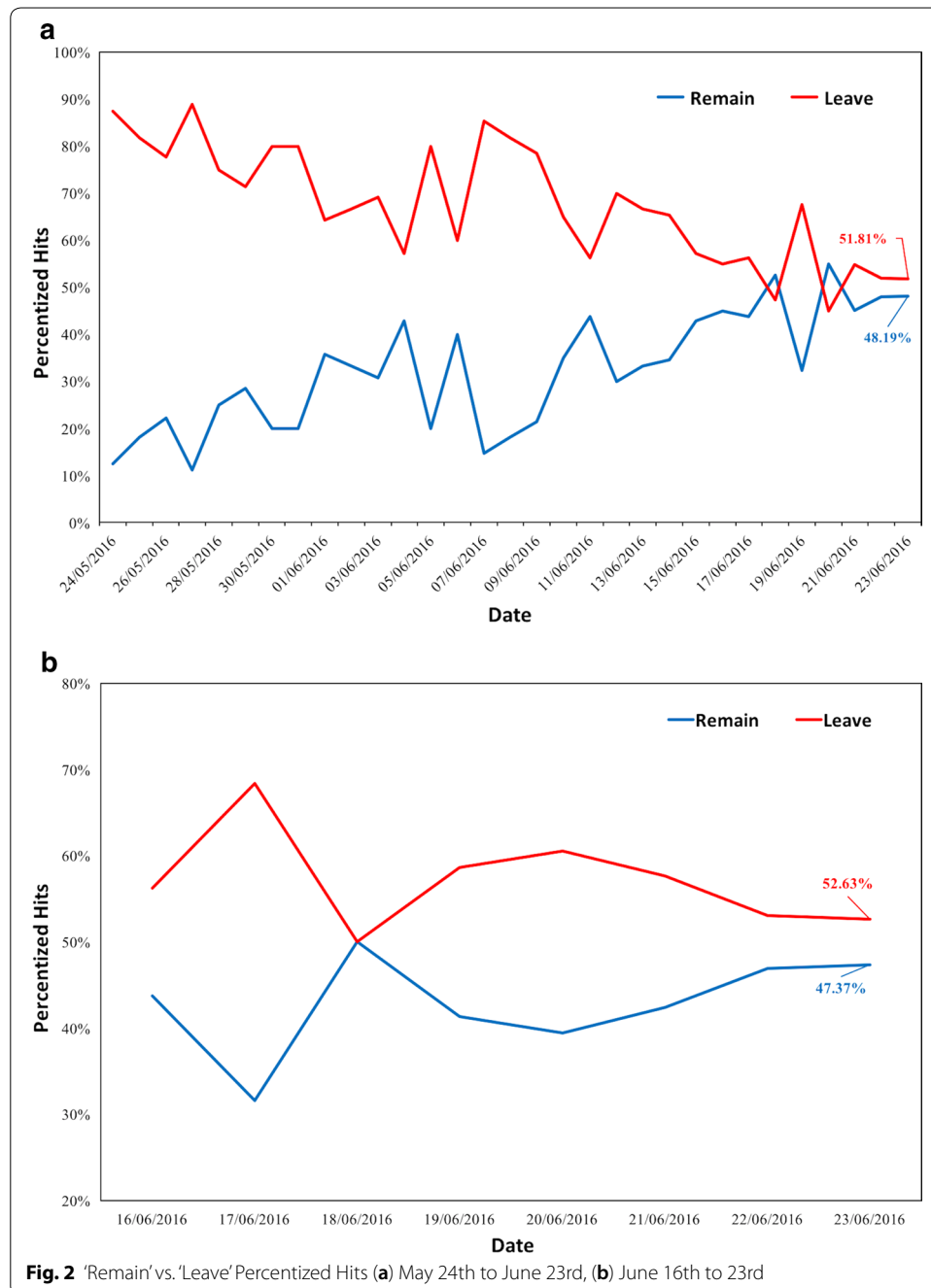
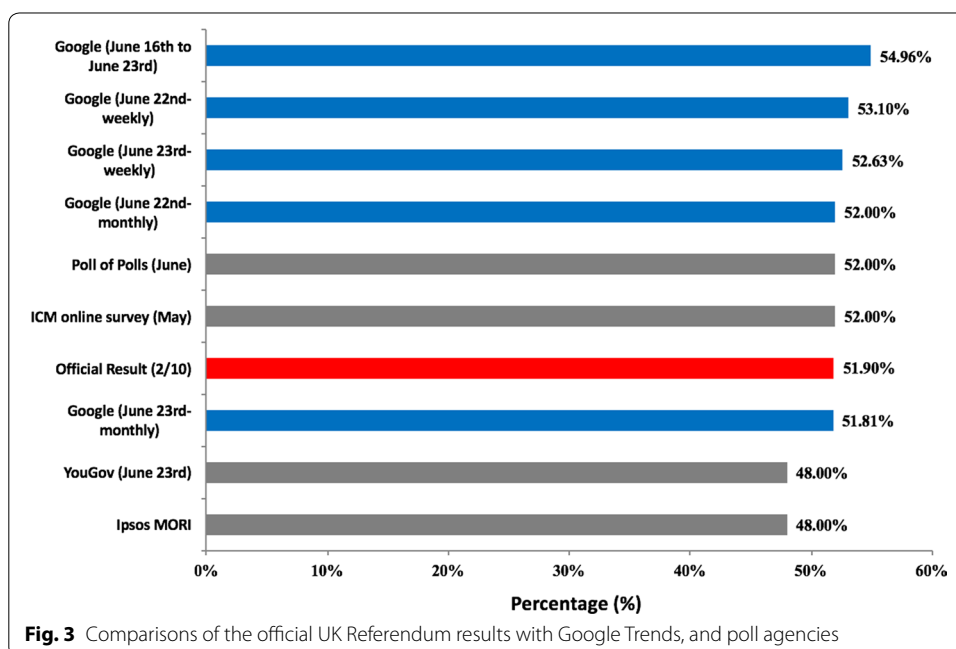**Fig. 2** 'Remain' vs. 'Leave' Percentized Hits (**a**) May 24th to June 23rd, (**b**) June 16th to 23rd

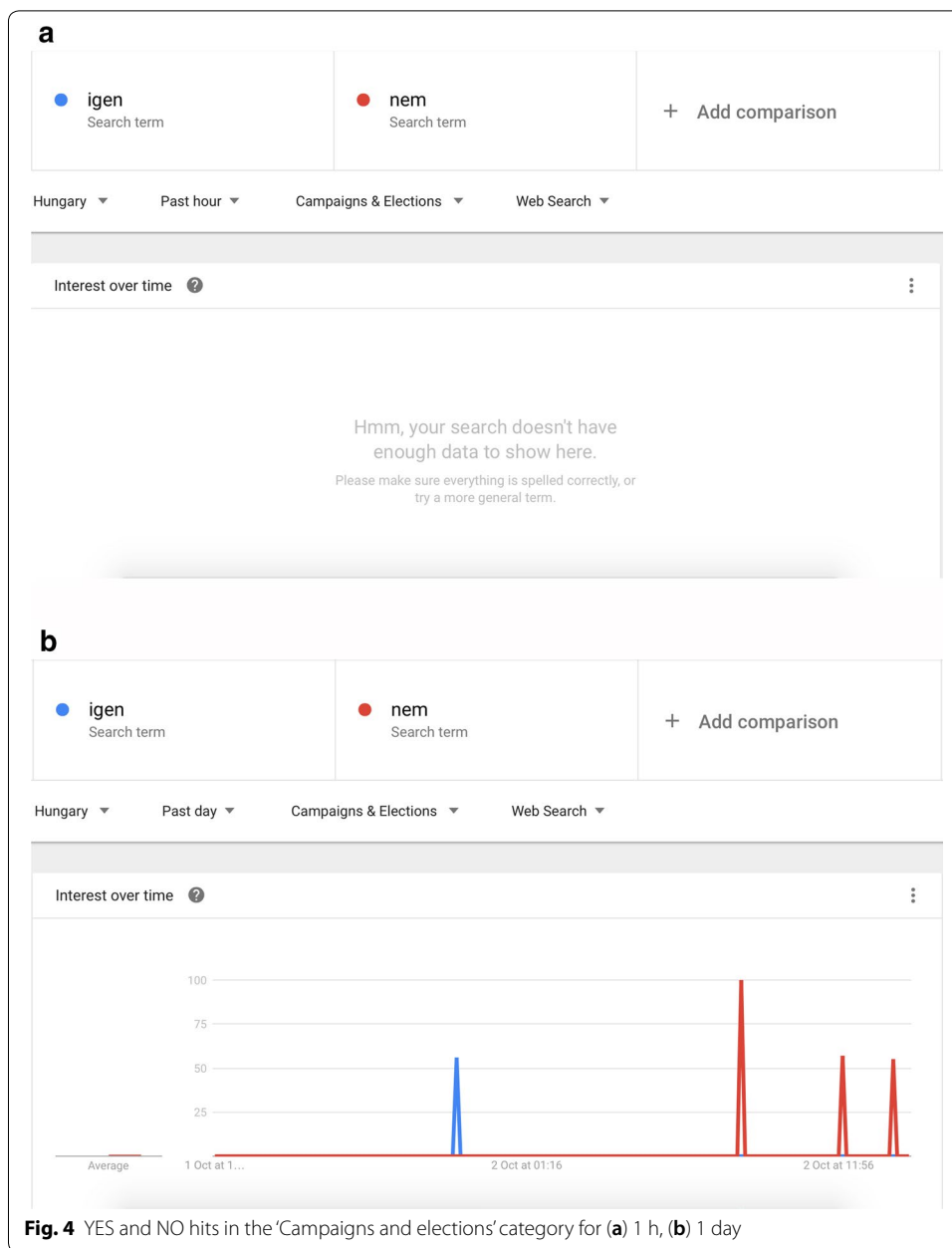### UK European union membership referendum (2016)

The 2016 UK Referendum on whether the UK was to remain a member of or leave the European Union was held on June 23rd. The opinion polls during the pre-voting period were contradicting, with several results suggesting that Remain was on the lead [48, 49], while others suggested that Leave was ahead [50]. The UK Referendum received wide international attention—media [51, 52] and scientific wise [53–55]—during the pre-voting period. Online pollings that were conducted suggested that Leave was leading, while traditional ones suggested a head-to-head race [56].

The official results put 'Leave' at 51.9% and 'Remain' at 48.1% [57]. Analyzing online search queries for the month before the Referendum, we observed that there was a rise in Remain during the last days before June 23rd, but Leave was almost at all points above Remain, as shown in Fig. 2, which consists of the percentized hits in the remain and leave keywords (a) from May 24th to June 21st, and (b) from June 16th to June 23rd.

A shift in favor of the Remain camp was observed following the murder of Jo Cox, a Labor Party MP and supporter of Remain, on the 16th [58]. As shown in Fig. 2a, the only point where Remain is ahead of Leave in Google Trends is the day after Cox's murder, at 52.63%, indicating that what the polls suggested was also immediately depicted in online searches. For the last day of the Referendum, the percentized averages of the monthly Google Trends' data put Remain at 48.19% and Leave at 51.81%, as shown in Fig. 2, while the averages of the daily percentized hits for the month before the race put NO at 60.65%. The average percentized YES and NO hits for the week before were at 45.04% and 54.96%, respectively, while the percentized hits for Remain and Leave on the last day were at 47.37% and 52.63%, respectively.

Contrary to what many polls on voting intentions and the result prediction published on the closing of the ballot boxes suggested, Google Trends clearly showed that 'Leave' was on the lead and was going to win the race. Figure 3 shows the comparisons for the

Fig. 3 Comparisons of the official UK Referendum results with Google Trends, and poll agencies

Fig. 4 YES and NO hits in the 'Campaigns and elections' category for (**a**) 1 h, (**b**) 1 day

term 'Leave' in Google Trends for the last 2 days of the referendum, i.e. the 22nd and 23rd—from monthly (May 24th to June 23rd) and weekly (June 16th to 23rd) datasets-and the weekly percentized average based on Google Trends, compared to the official results and the poll agencies' predictions in descending order.

Hungarian migrant quota referendum (2016)

The Hungarian referendum on immigration policy addresses a different methodological question than the rest. How is the low turnout on the day of the race depicted in online searches during the pre-voting period? While monitoring the Hungarian Referendum, apart from once again providing a good approximation of the outcome, what

Fig. 5 Comparisons of the official Hungarian Referendum results with Google Trends, poll agencies, and the exit poll
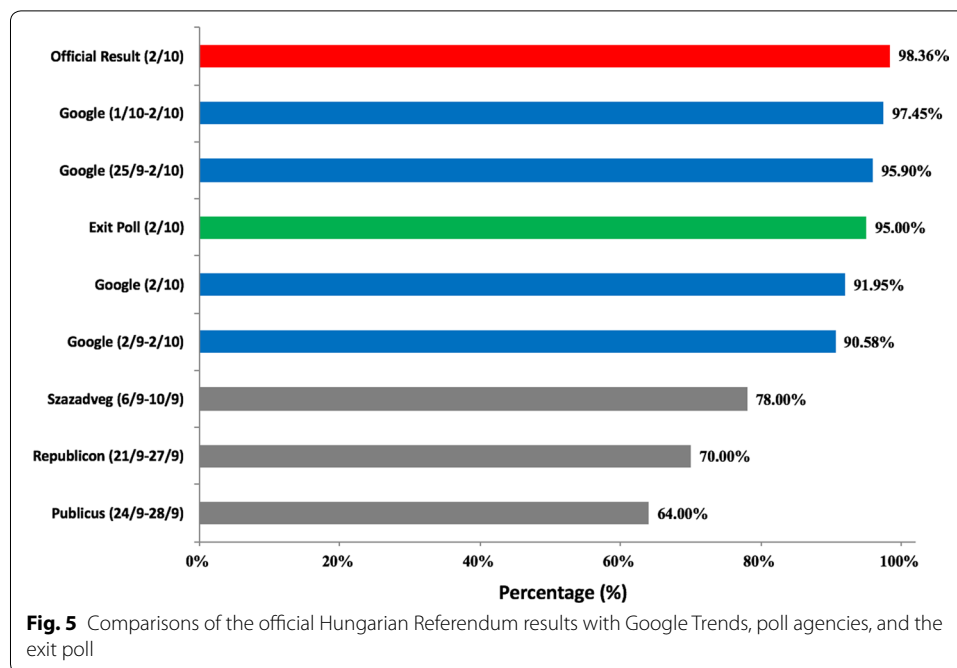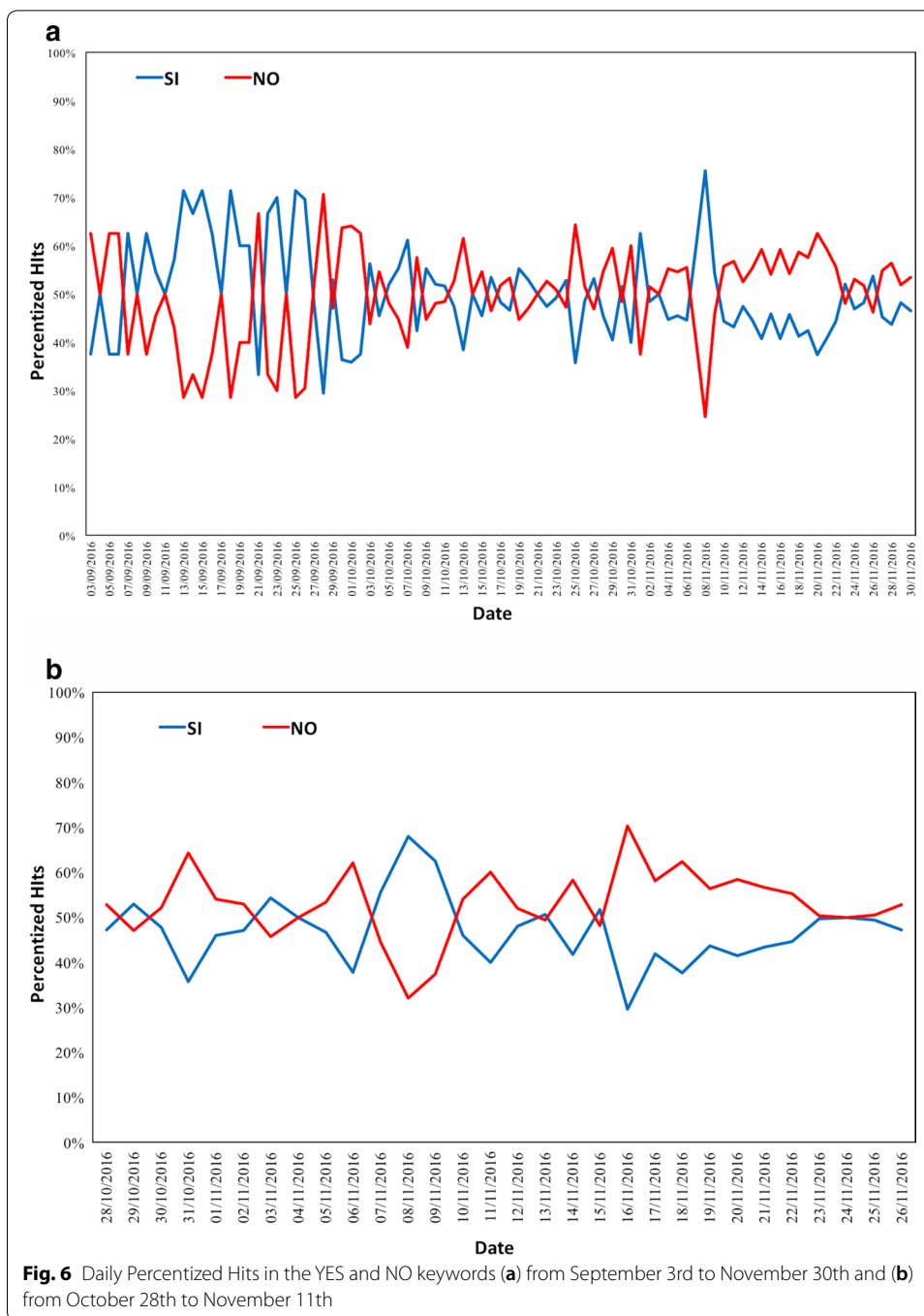
was of notable significance was the low volumes of data on the YES and NO (IGEN and NEM, respectively) keywords that the citizens of Hungary searched for on the web. This was later depicted on the day of the Referendum race, with more than 55% of the eligible voting population not showing up to vote [59].

Figure 4 shows that in the 'Campaings and Elections' field of search in Google Trends, the data based on the hits in the YES and NO searches are not sufficient to provide results. This indicates that the Hungarian voters were not interested enough in the Referendum to search for it online. Based on the very low turnout on the day of the Referendum race on October 2nd 2016—only 44.04% [59]—we conclude that this aforementioned low turnout is also depicted on Google queries.

Thus, in order to approximate the voting predictions and referendum outcome, we widened the search on the YES and NO keywords to include all hits related to politics, i.e. in the 'Politics' field in Google Trends, of which 'Campaigns and elections' is a subfield. Figure 5 depicts the comparisons of the official result for NO, the poll agencies' predictions, and the YES and NO Google percentized hits for 4 sets of data during the pre-voting period, i.e. (a) from September 2nd to October 2nd (monthly), (b) from September 25th to October 2nd (weekly), (c) from October 1st at 9 a.m. to October 2nd at 9 a.m. (daily), and (d) on October 2nd from 4 to 8 p.m. (4-h), in descending order.

It is evident that this method of predicting referendums has been valid in closely approximating the official results—that gave NO a 98.36% [59]—with a 95.90% percentized average for the last week before the race. We once again see the poll agencies not providing accurate results, giving NO at several points during the last month before the race 64% [60], 70% [61], and 78% [62], with only the exit poll closely approximating the results, giving NO a 95% [63].

Fig. 6 Daily Percentized Hits in the YES and NO keywords (**a**) from September 3rd to November 30th and (**b**) from October 28th to November 11th

Italian constitutional referendum (2016)

The Italian Constitutional Referendum took place on December 4th 2016, becoming a highly discussed referendum race, as it brought to light the overall euroskepticism evident in the EU countries over the past few years. As the referendum wording allowed for a YES/NO answer, the selected terms for analysis were 'SI' and 'NO', translating from Italian into 'YES' and 'NO', respectively. The official Referendum results put NO at 59.12% and YES at 40.88%, with the turnout being 65.47% [64].

Fig. 7 Hourly percentized hits for YES and NO for (**a**) November 19th to 26th and (**b**) November 26th to December 3rd
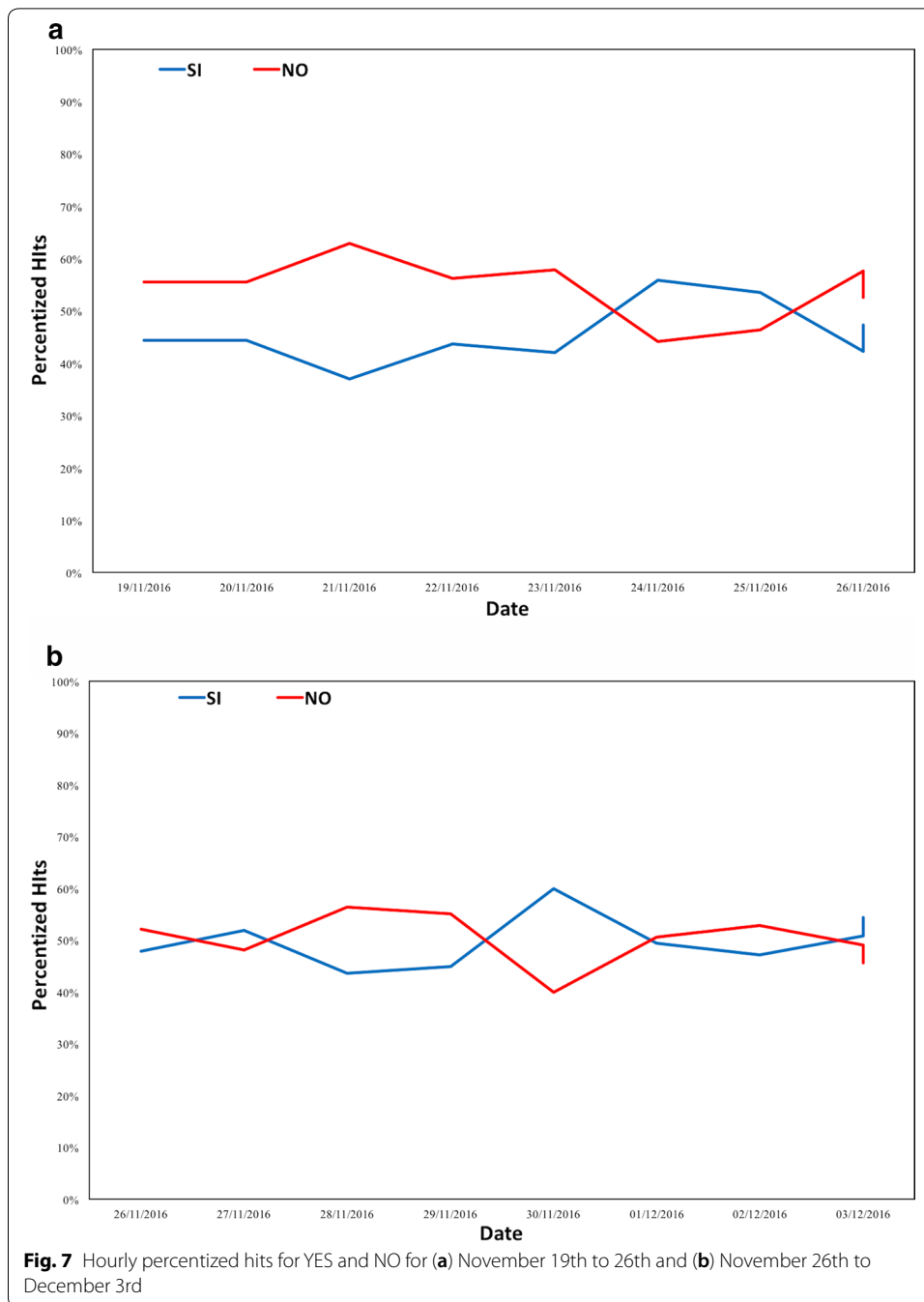
Figure 6 depicts the percentized hits for the YES and NO keywords (a) over the trimester before the referendum race, i.e. from September 3rd to November 30th, and (b) from October 28th to November 11th (up to a week before the race). The monthly averages for September put YES at 54.28% and NO at 45.72%; in October, YES and NO were at 48% and 52%, respectively, while for November, YES was at 47.39% and NO at 52.61%. The monthly average for YES and NO from October 28th to November 26th are 47.70% and 52.30%, respectively.

Fig. 8 Comparison of the official Italian Referendum results with Google Trends and poll agencies
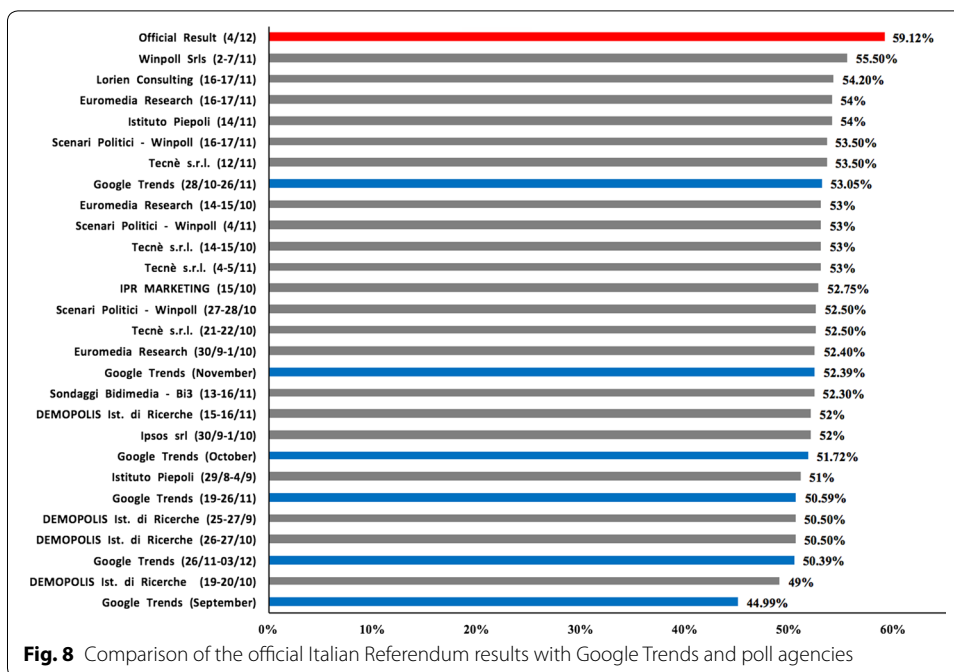
Figure 7 shows the percentized hits on the YES and NO keywords from the week from November 19th to November 26th and for the week from November 26th to December 3rd (hourly data). For the week starting 19, the average of the percentized hits for YES and NO were 49.21% and 50.79%, respectively, while for the week starting 26, the percentized hits were 49.57% and 50.43%, respectively.
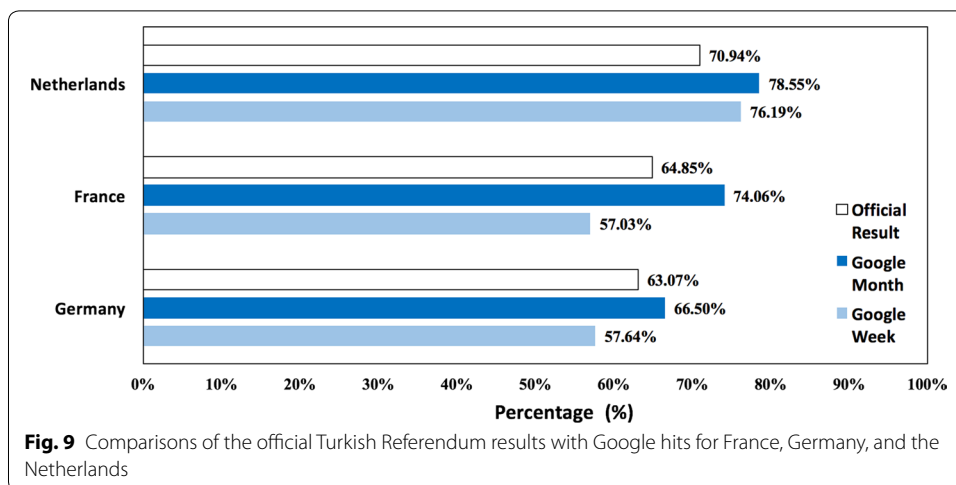
Figure 8 consists of the comparisons of the official results of the referendum race with the predictions using Google Trends and the poll agencies' [65] predictions during the pre-voting period. What is observed is that Google Trends' best approximation was calculated for the last month before the race at 53.05%. Though these approximations are some points below the actual result of NO, many poll agencies also predicted the NO vote to be somewhere between 55.50% and 49%, as shown in Fig. 8.

The predictions using data from Google Trends at some points gave better and at some points worse predictions of the results compared to traditional poll agencies. Despite the difference in percentage from the official result, we see that Google Trends data managed to approximate the result in a similar manner as poll agencies, and on the right side of the result, i.e. NO.

Turkish constitutional referendum (2017)

The President of Turkey, Recep Tayip Erdogan, announced that a referendum was to take place in order for the people of Turkey to decide on whether or not they agree with a set of 18 proposed constitutional amendments.

This Referendum, following the attempted coup of July 2016, was of high national and international significance, as the Turkish people were to also vote for or against the new proposed Presidential System. The wording of this referendum allowed for a simple YES

Fig. 9 Comparisons of the official Turkish Referendum results with Google hits for France, Germany, and the Netherlands

or NO answer, thus the terms selected are 'EVET' and 'HAYIR', translating from Turkish into YES and NO, respectively.

Given the conflicts between Turkey and several European countries, it is of significance to examine the online search queries in the YES and NO keywords in EU countries that have a large population of Turks allowed to vote in the respective country. Thus, in order to approximate the Turkish Referendum results, we selected the countries with the most population of Turks eligible to vote and with a turnout of more than 100,000, i.e. Germany, and France, where the Turkish population were eligible to vote until April 9th [66], and the Netherlands, voting on April 5th [67].

Figure 9 depicts the percentized hits' averages of Google Trends' data for the last week and month of the pre-voting periods compared to the official results for YES [68] in Germany (653,502 voters), France (140,741 voters), and the Netherlands (116,543 voters).

The final official results of the Referendum put YES at 51.18% and NO at 48.82%, while the overseas overall results, including border gates, put YES at 59.09% [68]. In Turkey, with a turnout of 85.46% on the day of the race [69], opinion polls results were contradicting over which option, YES or NO, was on the lead during the month before the race, with YES being put at 46.25% [70], 56.5% [71], 44.47% [72], 59.4% [73], 59% [74], 60.8% [75], and 46.1% [76]. What is evident by the above is that this referendum's outcome was hard to predict. The Turkish Referendum has also resulted in national and international dispute, as ballot papers not bearing the official seal were decided to be valid on the vote count [69].

The difficulty of predicting this referendum race in Turkey was also apparent in the NO searches in the country during the pre-voting period. Though Google Trends' data were useful in approximating the referendum results in the three examined European countries with Turkish population eligible to vote, the volumes of the online search queries for 'HAYIR' (translating into NO) in Turkey were at times extremely low. During the pre-voting period, EVET was way ahead of HAYIR, at some points even reaching very high percentages. Though we cannot argue that this is a result of data bias or manipulation,Internet monitoring and censorship have been suggested to be an issue of high significance in Turkey [77].

To elaborate, NO supporters in Turkey were viewed as "*siding with the coup-plotters*" [78], and at some cases prosecuted or fired [79]. Thus it could be the case that many were afraid to openly express their opinion, therefore poll agencies' voting intention results were so diverse. Internet penetration in Turkey is only 53.7% [80] and scores significantly low in freedom of the press, with 65 (the worst being 100) in the index 'Press Freedom' in 2015 [81], and 71 in 2016 [80], categorized as a 'Not Free' country, while continuously declining since 2010.

Censorship and restriction of freedoms have been reported in general, especially after the attempted coup in July 2016, where social media -such as Facebook, Twitter, Youtube, Whatsapp, and Instagram—have been reported blocked, censored, or restricted in various occasions [82–84]. Thus the 2017 Turkish Constitutional Referendum provides an excellent example of a limitation of this method of predicting referendum results, as analyzing online search queries cannot be applied in regions with low freedom of speech and media or governmental Internet monitoring.
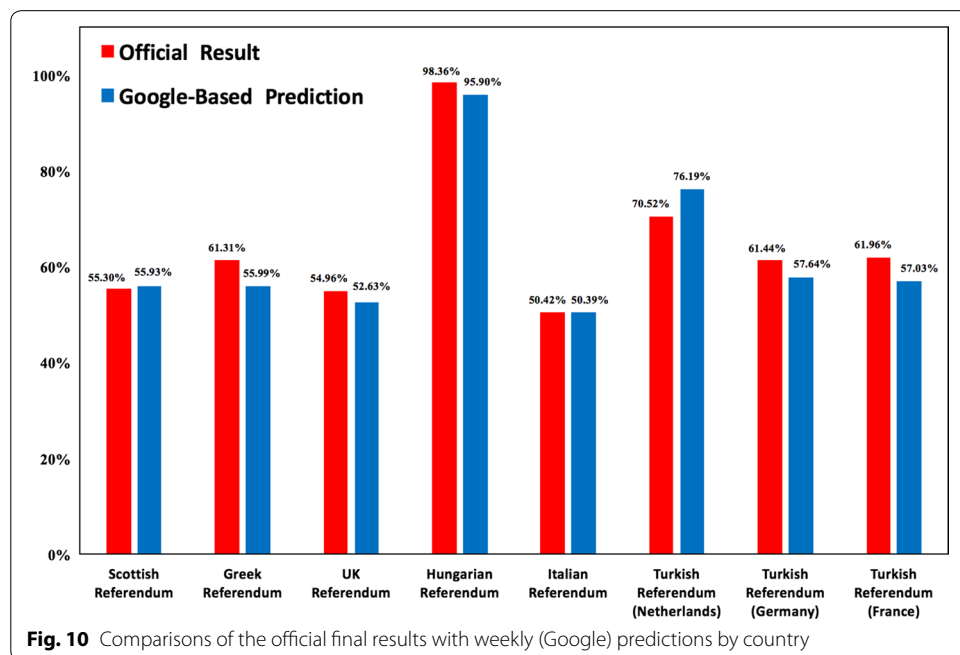
Discussion

This paper consisted of the presentation and validation of a novel method of predicting referendum results by monitoring online search queries from Google, tested and verified in six different occasions from 2014 to 2017, i.e. the 2014 Scottish Independence Referendum, the 2015 Greek Bailout Referendum, the 2016 UK European Union Membership Referendum, the 2016 Hungarian Migrant Quota Referendum, the 2016 Italian Constitutional Referendum, and the 2017 Turkish Constitutional Referendum. The results in this study closely approximated the respective official referendum results, and in some cases better than traditional polls. Based on the results of this paper and the analysis of online search queries in general, what is observed is that we have entered the era where Internet has brought significant changes in terms of monitoring, analyzing, and predicting human behavior.

In the EU, many referendums about EU matters have been conducted in the past, though during the last two decades the NO option seems to be becoming more popular [85]. This is not always attributed as an actual answer to the referendum question, but as a means of expressing dissatisfaction to the respective government [86]. Lately, poll agencies more than often fail to well approximate referendum and election outcomes. Therefore, new methods of predicting voting intentions and results have been examined [87–89], with online surveys taking the upturn.

The collection of referendums in this paper is notable for two reasons. At first, said referendums were significant in terms of policy and EU matters. Secondly, they present cases that will assist in future research using online queries as a polling tool, dealing with various special circumstances that arose. To elaborate, the very short pre-voting period in Greece was a good example of how to nowcast a referendum. In this case, the data had to be downloaded in very short time frames, i.e. every hour, every 4-h, and every day so as to extract robust results. In Hungary, the low turnout and interest in the referendum was depicted in the searches in Google during the pre-voting period. Furthermore, what was notable in the 2016 UK Referendum, was the effect of the murder of Labour MP Jo Cox, where the opinion shifted towards the Remain camp in the next couple of days after the murder. The UK Referendum

Table 1 Mean comparisons for the YES and NO percentized hits

Referendum	YES (%)	NO (%)	|*t*-statistic|	*p*-value
Scottish (monthly)	32.98	67.02	15.456	< 0.001
Scottish (weekly)	44.07	55.93	7.361	< 0.001
UK (monthly)	32.39	67.61	11.128	< 0.001
UK (weekly)	42.85	57.15	4.992	< 0.001
Italian 19/11–23/11	49.41	50.59	1.400	0.162
Italian 26/11–3/12	49.61	50.39	1.121	0.263
Italian 28/10–26/11	46.95	53.05	3.111	< 0.01
Italian 3/11–30/11	47.26	52.74	2.755	< 0.01
Italian 3/9–30/11	50.17	49.83	0.232	0.817
Turkish–France (monthly)	74.06	25.94	6.736	< 0.001
Turkish-France (weekly)	57.03	42.97	0.528	0.609
Turkish–Germany (monthly)	66.50	33.50	7.278	< 0.001
Turkish–Germany (weekly)	57.64	42.36	2.244	< 0.05
Turkish–Netherlands (monthly)	78.55	21.45	8.708	< 0.001
Turkish–Netherlands (weekly)	76.19	23.81	3.633	< 0.01

Fig. 10 Comparisons of the official final results with weekly (Google) predictions by country

was also significant in terms of the wording that did not allow for a simple YES/NO answer, thus the selection of keywords for monitoring the interest was not trivial. The most interesting case, though, was that of the Turkish Referendum, in a country where Internet restrictions are a serious issue. In Turkey, the use of Google Trends' data were not useful in predicting the result. Despite that, they provided good approximations for the three EU countries where significantly large Turkish population voted, i.e. in France, Germany, and the Netherlands. We thus conclude that this

method cannot be applied in regions with low levels of freedom of media, Internet, and speech, or with low Internet penetration.

Table 1 consists of the averages of the weekly and monthly percentized hits for YES and NO (Remain-Leave for the UK) for the examined referendum races, and their respective statistical significance for mean comparisons. Such comparisons can provide beforehand the result regarding the side of the outcome, i.e. whether the public shifts towards YES or NO, while the proximity with the final percentage is increasing as the time span approximates the ballot closing time, as discussed above.

As is evident, the results for the Scottish and the UK referendums exhibit high statistical significance on the final outcome. Regarding the Turkish Referendum, all examined comparisons are on the same side as the official results and are statistically significant, with the exception of the weekly results for France. Regarding the Italian referendum results, two of the compared datasets are statistically significant, while the rest do not statistically prove the 'NO' response.

As is suggested in Fig. 10—depicting the comparisons of the official results in each examined country with the predictions using Google Trends' data for the last week of the pre-voting period—the analysis of online search queries is a valid method of approximating the result of a referendum and the voting intentions. Note that in the Greek and the Hungarian Referendum, our results using Google Trends' data were the closest approximation to the final result, better than the poll agencies' predictions using traditional methods. Also note that for the UK Referendum, the result refers to the percentized hits of the last day of the weekly time series.

As we validated that the results of using Google Trends for voting intentions and result predictions are accurate and, in many cases, better than official polls, we argue that this method will be adopted in the near future by poll agencies and political researchers in this field, as it tackles the economic risk, the cost, and the uncertainty barriers of poll taking.

However, some limitations do exist. At first, this method could not be applied in regions where the use of the Internet is restricted, in the ones having low scorings in freedom of press, or in those with low internet penetration, where data manipulation is possible, i.e. intentional googling of a certain outcome. Furthermore, the sample can not be proven representative. Not all internet users vote and not all voters use the Internet to search for the respective referendum keywords, thus not each hit can be linked to referendum voting, and in no way is it to be implied that such a 1–1 correspondence exists. In addition, a limitation of the tool is that it has been observed that data retrieved on different time points for the same time-frame may slightly vary. Despite the above, many studies in various subjects have shown that indeed Google data can be used to analyze or predict behavioral variations [4, 5, 8, 13, 16, 23, 26, 90–92] and that empirical relationships exist between online search traffic data and human behavior [93, 94].

Conclusions

In this paper, we presented a novel methodology for predicting voting intentions and referendum results using online search traffic data from Google, which was tested and validated in six referendum races. Said referendums, i.e. the 2014 Scottish Referendum, the 2015 Greek Referendum, the 2016 UK Referendum, the 2016 Hungarian Referendum,

the 2016 Italian Referendum, and the 2017 Turkish Referendum, were of significance for the European Union and received wide national and international attention. Employing data from Google Trends, we estimated the respective referendum results using data on the (translated) YES and NO keywords. Our results exhibited good performance, while, in some cases, were more accurate than official polls.

Since online behavioral changes can be measured by online data [30], the potential benefit of using Google Trends' data as a poll taking method is high, especially as traditional polls did not always manage to accurately predict the outcome, and, in many cases, were not even on the winning side of the results. Google Trends has been shown to be a credible means of examining behavioral changes, as it uses the revealed and not the stated preferences of the users. Thus in regions were Internet is widely accessible and not restricted, online data are useful in analyzing and predicting human behavior in many research topics.

This method of poll taking, based on empirical relationships, could be of interest to political researchers, as it is a valid analyzer of human behavior, indicating that Internet data can give insight to behavioral variations towards political matters and election races. Future research on the subject could focus on developing more sophisticated models using online data, as well as the combination of various online sources, or the combination of online with traditional survey data. Overall, the above suggest that monitoring Internet data in general and Google Trends data in specific will be a polling tool to address the challenge of Big Data in the future.

Authors' information
Amaryllis Mavragani is a Ph.D. Candidate at the Department of Computing Science and Mathematics, University of Stirling. She holds a B.Sc. in Mathematics from the University of Crete and an M.Sc. from Democritus University of Thrace. Her research interests include Big Data, Mathematical Modeling, Infodemiology, Internet Behavior, and Public Health.

   Konstantinos P. Tsaragakis is a Professor at the Department of Environmental Engineering, Democritus University of Thrace. He holds a Civil Engineering Diploma from Democritus University of Thrace, a BA in Economics from the University of Crete, and a Ph.D in Public Health from the University of Leeds. His research interests include Wastewater Management, Environmental Economics, Energy Economics, Environmental Policy, Big Data, Public Awareness and Behavior, Quantitative Methods, and Techno-economic Project Analysis.

Author details
[1] Department of Computing Science and Mathematics, Faculty of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK. [2] Business and Environmental Technology Economics Lab, Department of Environmental Engineering, Democritus University of Thrace, 67100 Xanthi, Greece.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Hilbert M, Lopez P. The World's technological capacity to store, communicate, and compute information. Science. 2011;332:60–5.
2. Chen CL, Zhang CY. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf Sci. 2014;275:314–47.
3. Al Nuaimi E, Al Neyadi H, Mohamed N, Al-Jaroodi J. Applications of big data to smart cities. J Int Serv App. 2015;6:25.
4. Preis T, Moat HS, Stanley HE. Quantifying trading behavior in financial markets using Google Trends. Sci Rep. 2013;3:1684.
5. Preis T, Moat HS, Stanley HE, Bishop SR. Quantifying the advantage of looking forward. Sci Rep. 2012;2:350.
6. Google Trends. https://trends.google.com/trends/explore. Accessed 10 Apr 2018.
7. McCallum ML, Bury GW. Public interest in the environment is falling: a response to Ficetola (2013). Biodiv Conserv. 2014;23:1057–62.
8. Jun SP, Park DH. Consumer information search behavior and purchasing decisions: empirical evidence from Korea. Technol Forecast Soc. 2016;31:97–111.
9. Han SC, Chung H, Kang BH. It is time to prepare for the future: forecasting social trends. Computer applications for database, education, and ubiquitous computing. Berlin: Springer; 2012. p. 325–31.
10. Jun SP, Park DH, Yeom J. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. Technol Forecast Soc. 2014;86:237–53.
11. Jun SP, Yeom J, Son JK. A study of the method using search traffic to analyze new technology adoption. Technol Forecast Soc. 2014;81:82–95.
12. Vicente MR, Lopez-Menendez AJ, Perez R. Forecasting unemployment with internet search data: does it help to improve predictions when job destruction is skyrocketing? Technol Forecast Soc. 2015;92:132–9.
13. Vosen S, Schmidt T. Forecasting private consumption: survey-based indicators vs. Google trends. J Forecast. 2011;30:565–78.
14. Mavragani A, Tsagarakis KP. YES or NO: predicting the 2015 Greferendum results using Google Trends. Technol Forecast Soc. 2015;2016(109):1–5.
15. Carrière-Swallow Y, Labbé F. Nowcasting with google trends in an emerging market. J Forecast. 2013;32:289–98.
16. Nuti SV, Wayda B, Ranasinghei I, Wang S, Dreyer RP, Chen SI, Murugiah K. The use of google trends in health care research: a systematic review. PLoS ONE. 2014;9(10):e109583.
17. Zhou X, Ye J, Feng Y. Tuberculosis surveillance by analyzing google trends. IEEE Trans Biomed Eng. 2011;58(8):2247–54.
18. Troelstra SA, Bosdriesz JR, De Boer MR, Kunst AE. Effect of tobacco control policies oninformation seeking for smoking cessation in the Netherlands: a google trends study. PLoS ONE. 2016;11(2):0148489.
19. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, Icardi G, Orsi A. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty. 2015;4(1):54.
20. Wang HW, Chen DR, Yu HW, Chen YM. Forecasting the incidence of dementia and dementia-related outpatient visits with google trends: evidence from Taiwan. J Medi Internet Res. 2015;17(11):e264.
21. Zhang Z, Zheng X, Zeng DD, Leischow SJ. Information seeking regarding tobacco and lung cancer: effects of seasonality. PLoS ONE. 2015;10(3):e0117938.
22. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could google trends be used to predict methamphetamine-related crime? An analysis of search volume data in Switzerland, Germany, and Austria. PLoS ONE. 2016;11(11):0166566.
23. Davidson MW, Haim DA, Radin JM. Using networks to combine big data and traditional surveillance to improve influenza predictions. Sci Rep. 2015;5:8154.
24. Kristoufek L. Power-law correlations in finance-related Google searches, and their crosscorrelations with volatility and traded volume: evidence from the Dow Jones Industrial components. Phys A. 2015;428:194–205.
25. Kristoufek L. Can google trends search queries contribute to risk diversification? Sci Rep. 2013;3:2713.
26. Choi H, Varian H. Predicting the present with Google Trends. Econ Rec. 2012;88:2–9.
27. Kristoufek L. BitCoin meets Google Trends and Wikipedia: quantifying the relationship between phenomena of the Internet era. Sci Rep. 2013;3:3415.
28. McCallum ML, Bury GW. Google search patterns suggest declining interest in the environment. Biodiv Conserv. 2013;22:1355–67.
29. Wagner SA, Vogt S, Kabst R. The future of public participation: empirical analysis from the viewpoint of policymakers. Technol Forecast Soc. 2016;106:65–73.
30. Burnap P, Rana OF, Avis N, Williams M, Housley W, Edwards A, Morgan J, Sloan L. Detecting tension in online communities with computational Twitter analysis. Technol Forecast Soc. 2015;95:96–108.
31. Weber I, Popescu AM, Pennacchiotti M. PLEAD 2013: politics elections and data. In WSDM 13.
32. Google Trends. How Trends data is adjusted. https://support.google.com/trends/answer/4365533. Accessed 10 Apr 2018.
33. Scharkow M, Vogelgesang J. Measuring the public agenda using search engine queries. Inte J of Public Opin R. 2011;23(1):104–13.
34. Scottish Parliament. 2014. http://www.parliament.scot/ResearchBriefingsAndFactsheets/Scottish_Independence_Referendum_2014_Results.pdf. Accessed 10 Apr 2018.

35. TNS. 2014. http://www.tns-bmrb.co.uk/sites/tns-bmrb/files/TNSUK_SOM2014Sep9_DataTables.pdf. Accessed 10 Apr 2018.
36. Panelbase. 2014. http://www.panelbase.com/media/polls/F6022TablesForPublication170914.pdf. Accessed 18 Apr 2017.
37. Survation. 2014. http://survation.com/wp-content/uploads/2014/09/24-hour-scottish-referendum-poll.pdf. Accessed 10 Apr 2018.
38. Ipsos MORI. 2014. https://www.ipsos-mori.com/Assets/Docs/Polls/scotland-referendum-18-september-tables.pdf. Accessed 18 April 2017.
39. YouGov. 2014. https://yougov.co.uk/news/2014/09/18/scotland-no-enters-polling-day-4-ahead/. Accessed 18 Apr 2017.
40. YouGov. 2014. https://yougov.co.uk/news/2014/09/12/scotlands-yes-bandwagon-stalls/. Accessed 10 Apr 2018.
41. Yardley J, Kitsantonis N. Greek prime minister calls for referendum on bailout terms. 2015. http://www.nytimes.com/2015/06/27/business/international/greek-debttalks-enter-final-stages.html. Accessed 27 Feb 2016.
42. BBC. The papers. 2015. http://www.bbc.com/news/blogs-thepapers-33420972. Accessed 27 Feb 2016.
43. Sembhy R. Greek referendum: Polls neck and neck as Tsipras calls for debt relief. 2015. http://www.ibtimes.co.uk/greek-referendum-polls-neck-neck-tsipras-calls-debt-relief-1509232. Accessed 10 Apr 2018.
44. Greek Ministry of Interior. Referendum July 2015. 2015. http://ekloges-prev.singularlogic.eu/r2015/e/public/index.html#{"cls":"main","params":{}}). Accessed 27 Feb 2016.
45. MARC. Results predictions. 2015. http://www.ibtimes.com/greek-referendum-results-2015-opinion-polls-show-no-vote-will-likely-prevail-1996150. Accessed 10 Apr 2018.
46. MRB. Results prediction. 2015. http://www.star.gr/Pages/Politiki.aspx?art=284110&artTitle=dimoskopisi_mrb_mikro_provadisma_tou_ochi_deite_ta_prota_pososta_gia_logariasmo_tou_star. Accessed 27 Feb 2016.
47. Metron Analysis. Results predictions. 2015. http://www.antenna.gr/news/Politics/article/414803/mprosta-to-oxi-stin-teleytaia-dimoskopisi-tis-metron-analysis. Accessed 27 Feb 2016.
48. Ipsos MORI. Brexit poll: Remain takes narrow lead in final EU referendum survey. Evening Standard. 2016. http://www.standard.co.uk/news/politics/brexit-poll-remain-secure-narrow-lead-in-final-eu-referendum-poll-a3278996.html. Accessed 10 Apr 2018.
49. Independent. EU referendum polls: Final ComRes poll shows significant lead for Remain. 2016. http://www.independent.co.uk/news/uk/politics/eu-referendum-poll-final-brexit-comres-yougov-opinium-tns-survey-remain-leave-live-result-a7096316.html. Accessed 10 Apr 2018.
50. ICM. UK voters leaning towards Brexit, Guardian poll reveals. The Guardian. 2016. http://www.theguardian.com/politics/2016/may/31/uk-voters-leaning-towards-brexit-guardian-poll-reveals. Accessed 10 Apr 2018.
51. Barnes P. EU referendum poll tracker. BBC News. 2016. http://www.bbc.com/news/uk-politics-eu-referendum-36271589. Accessed 10 Apr 2018.
52. Erlanger S. 'Brexit': Explaining Britain's Vote on European Union Membership. The New York Times. 2016. http://www.nytimes.com/interactive/2016/world/europe/britain-european-union-brexit.html?_r=0. Accessed 10 Apr 2018.
53. Henderson A, Jeffery C, Lineira R, Scully R, Wincott D, Jones RW. England, Englishness and Brexit. Polit Quart. 2016;87:2.
54. Oliver T. European and international views of Brexit. J Eur Public Policy. 2016;23(9):1321–8.
55. Crafts N. The Impact of EU Membership on UK Economic Performance. Polit Quart. 2016;2016(87):2.
56. Wheeler B, Hunt A. The UK's EU referendum: all you need to know. BBC News. http://www.bbc.com/news/uk-politics-32810887. Accessed 10 Apr 2018.
57. The Electoral Commission. 2016. https://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information. Accessed 10 Apr 2018.
58. Bloomberg. First Brexit Poll Since Jo Cox Killing Has 'Remain' in Lead. 2016. http://www.bloomberg.com/news/articles/2016-06-18/first-brexit-poll-since-jo-cox-killing-shows-remain-in-lead. Accessed 10 Apr 2018.
59. National Election Office. http://www.valasztas.hu/en/ref2016/481/481_0_index.html. Accessed 10 Oct 2016.
60. Publicus. http://www.publicus.hu/blog/kvotanepszavazas_csokkeno_reszveteli_szandek_-_partok_tamogatottsaga_n/. Accessed 10 Oct 2018.
61. Republicon. http://republikon.hu/elemzesek,-kutatasok/160930-nepszavazas-kutatas.aspx. Accessed 10 Apr 2018.
62. Szazadveg. https://888.hu/article-szazadveg-a-szavazok-tobbsege-elutasitja-a-nepszavazas-bojkottjat. Accessed 10 Apr 2018.
63. Nezopont. http://hvg.hu/itthon/20161002_Exit_poll_A_Nezopont_Intezet_32_millio_nem_szavazatra_szamit. Accessed 10 Apr 2018.
64. Italian Ministry of Interior. http://elezioni.interno.it/referendum/scrutini/20161204/FX01000.htm. Accessed 10 Apr 2018.
65. Dipartimento per l'Informazione e l'Editoria. http://www.sondaggipoliticoelettorali.it/ListaSondaggi.aspx?st=SONDAGGI. Accessed 25 Jan 2017.
66. BBC NEWS, 2017. Voting starts in Europe for Turkish referendum. http://www.bbc.com/news/world-europe-39405210. Accessed 10 Apr 2018.
67. Daily Sabah Elections. Turkish nationals abroad start voting on constitutional referendum. Available from: https://www.dailysabah.com/elections/2017/03/28/turkish-nationals-abroad-start-voting-on-constitutional-referendum. Accessed 10 Apr 2017.
68. Anadolu Ajansi, 2017. http://secim.aa.com.tr. Accessed 17 Apr 2017.
69. Anadolu A. 2017. http://aa.com.tr/en/turkey/election-board-head-says-unsealed-ballot-papers-valid/798550. Accessed 17 Apr 2017.
70. AKAM. 2017. http://www.sozcu.com.tr/2017/gundem/iste-son-referandum-anketi-yuzde-53-1787544/. Accessed 17 Apr 2017.
71. MAK. http://www.diken.com.tr/mak-anketi-evet-yuzde-56-5-hayir-yuzde-43-5/. Accessed 17 Apr 2017.

72. AKAM. 2017. http://www.birgun.net/haber-detay/avrasya-kamuoyu-arastirma-referandum-anketini-acikladi-15235 3.html. Accessed 17 Apr 2017.
73. ORC. 2017. http://www.orcarastirma.com.tr/wp-content/uploads/2017/04/16-NİSAN-REFERANDUM-OY-TERCİHLERİ. pdf. Accessed 17 Apr 2017.
74. Qriously. 2017. http://www.jamesinturkey.com/yes-still-ahead-watch-istanbul/. Accessed 17 Apr 2017.
75. A&G. 2017. http://www.milliyet.com.tr/adil-gur-referandum-anketini-siyaset-2432577/. Accessed 17 Apr 2017.
76. THEMIS. 2017. http://www.sozcu.com.tr/2017/gundem/son-dakika-son-referandum-anketi-aciklandi-flas-sonuc lar-var-1790998/. Accessed 17 Apr 2017.
77. Akgul M, Kirlidog M. Internet censorship in Turkey. Int Pol Rev. 2015;4(2):1–22.
78. BBC NEWS. 2017. http://www.bbc.com/news/world-europe-39064657. Accessed 17 Apr 2017.
79. Turkish 'No' voices muffled in Erdoğan's referendum. https://www.politico.eu/article/turkey-constitutional-refer endum-recep-tayyip-erdogan-orhan-pamuk-akp-chp-degirmenci-hurriyet/. Accessed 13 Dec 2017.
80. Freedomhouse. 2016. https://freedomhouse.org/report/freedom-press/2016/turkey. Accessed 17 Apr 2017.
81. Freedomhouse. 2015. https://freedomhouse.org/report/freedom-press/2015/turkey. Accessed 17 Apr 2017.
82. Turkeyblocks. 2017. https://turkeyblocks.org/2016/08/25/social-media-blocked-turkey/. Accessed 17 Apr 2017.
83. Telegraph. 2016. http://www.telegraph.co.uk/technology/2016/12/20/turkey-blocks-access-facebook-twitter-whats app-following-ambassadors/. Accessed 19 Apr 2017.
84. Voanews. 2017. http://www.voanews.com/a/turkey-moves-to-block-internet-access-/3662886.html. Accessed 19 Apr 2017.
85. Qvortrup M. Referendums on Membership and European Integration 1972–2015. Polit Quart. 2016;87:1.
86. Vasilopoulou S. UK Euroscepticism and the Brexit Referendum. Polit Quart. 2016;87:2.
87. Murr AE. The wisdom of crowds: applying Condorcet's jury theorem to forecasting US presidential elections. Int J Fore. 2015;31(3):916–29.
88. Rothchild D. Combining forecasts for elections: accurate, relevant, and timely. Int J Fore. 2015;31(3):952–64.
89. Wang W, Rothchild D, Goel S, Gelman A. Forecasting elections with non-representative polls. Int J Fore. 2015;31(3):980–91.
90. Bragazzi NL, Bacigaluppi S, Robba C, Nardone R, Trinka E, Brigo F. Infodemiology of status epilepticus: a systematic validation of the Google trends-based search queries. Epilepsy Behav. 2016;55:120–3.
91. Mavragani A, Sypsa K, Sampri A, Tsagarakis KP. Quantifying the UK online interest in substances of the EU watch list for water monitoring: diclofenac, estradiol, and the macrolide antibiotics. Water. 2016;8:542.
92. Pollett S, Wood N, Boscardin WJ, Bengtsson H, Schwarcz S, Harriman K, Winter K, Rutherford G. Validating the use of Google trends to enhance pertussis surveillance in California. PLOS Curr Outbreaks. 2015;19:7.
93. Mavragani A, Ochoa G. Forecasting AIDS prevalence in the United States using online search traffic data. J Big Data. 2018;5:17.
94. Mavragani A, Ochoa G. Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. J Big Data. 2018;5:30.