# Replication Study: Wikipedia traffic data and electoral prediction

## Project Report

**Jonas Stettner (01/1152625)**

Universität Konstanz

Social Media Data Analysis
Department of Politics and Public Administration

Evaluated by Prof. David Garcia

Konstanz, 2023

# Table of Contents

# 1 Motivation

This project focuses on the use of Wikipedia traffic as a digital trace of internet users seeking information. Wikipedia traffic occurs when users access a Wikipedia page, resulting in a logged request to the server. This straightforward metric of tracking the number of views on specific Wikipedia pages over time can be leveraged for quantitative analysis to test different hypotheses related to information-seeking behavior and its implications. Implications can be of political nature when the subject of information-seeking is a social phenomenon like elections. In the paper titled "Wikipedia traffic data and electoral prediction: towards theoretically informed models" [Yasseri and Bright, 2016] the authors are investigating such implications. The objective of this project is to replicate the parts of findings of the paper while also extending the application of the methods explored in the paper to more recent data.

Yasseri et al. are developing a theory on the relationship between views on election-related pages on Wikipedia and election outcomes. Following the rational choice paradigm, the theory is based on the assumption that voters are deciding for the party whose policies are maximizing their benefit. To do so, it is considered rational to seek information about the election and party policies. The authors cite studies indicating that a significant proportion of adults across different countries rely on Wikipedia as an information source [Rainie, 2020, Ofcom, 2015]. It can be assumed that this still holds. One piece of evidence is that Wikipedia is ranked as the 7th most visited website in the world in August 2023 [Similarweb, 2023]. The authors conclude that views on Wikipedia pages related to an election before the election itself are mainly due to voters seeking information and that one can therefore use page views as a predictor of election outcomes.

Yasseri et al. bring forth two main hypotheses: 1. Page view numbers of general election pages are a predictor of election turnout and 2. Page view numbers of party pages are a predictor of party results. However, they theorize that there are factors that may moderate the prediction strength when it comes to party results, such as voters being more likely to seek information on new political parties and swing voters being more likely to seek information if they are considering changing their vote. Moreover, the authors argue that coverage of parties in the traditional media could weaken the prediction strength of Wikipedia page views. The paper's findings indicate that while Wikipedia may not provide detailed information on the exact results of votes, it does offer useful insights into changes in overall voter turnout and shifts in the percentage of votes obtained by specific parties. These results are based on outcomes of the European Parliament elections in 2009 and 2014 and related Wikipedia page view statistics and traditional news coverage from the time of the elections.

In this project, I am extending the data to the 2019 European Parliament elections, while also gathering page view statistics on the previous two elections, because the authors did not include this data with their paper. My main objective is to replicate testing one the author's first hypotheses about Wikipedia page views being a predictor for election turnout.

# 2 Background

The paper by Yasseri et al. was written in the wake of similar attempts to use digital traces to predict social outcomes. Distinctions of these attempts can be made by data sources and the nature of the predicted social phenomenon. A data source that is comparable to Wikipedia page views, because it originates from people looking for information, is web search data. For example, a study that the authors mention explores using web search data to predict consumer behavior across different domains [Goel et al., 2010]. While there are domain-specific differences in prediction strength, the paper's findings suggest that this generally produces significant results. Search data originating from Google searches has also been applied to predict doctor visits for influenza-like illness. This produced results with high accuracy [Ginsberg et al., 2009] for a limited time. A project by Google, called Google Flu Trends, attempted to provide this as a web service but famously failed in 2013. It has been hypothesized that the predictive power of Google search data is limited due to the nature of Google's search algorithm that biases results [Lazer et al., 2014]. While Wikipedia itself does not use algorithms that bias page views in the same way, their traffic could be affected indirectly by internet users being forwarded from search results.

Another data source used to predict social phenomena frequently both in academia and in commercial contexts is social media. In a recent literature review, Rousidis et al. distinguish phenomena predicted by social media data by assigning them to the fields of Finance, Marketing, and "Sociopolitical" [Rousidis et al., 2020]. The sociopolitical domain, which elections belong to as well, poses the most challenges for prediction; it has witnessed both successful prediction outcomes and notable failures. Another finding of Rousidis et al. is that among the data sources used as predictors, Twitter is the most popular (77% of 43 studies). Yasseri et al. cite a study that claims to have found a statistically significant correlation between the presence of tweets mentioning candidates and their subsequent electoral performance [DiGrazia et al., 2013]. However, it has been argued by other authors that using Twitter data for prediction in the sociopolitical domain, especially for predicting elections is problematic or even fruitless. Suggested reasons include the existence of unanswered questions about the nature of political conversations on social media and representativity issues [Gayo-Avello et al., 2011]. This has to be taken into account when using Wikipedia page views as a predictor. Although articles on Wikipedia are the results of a social process of multiple people contributing, Wikipedia page views should not be influenced by this in the case of the study to be replicated, as the relevant articles already exist, and it is not relevant how people interact with them. Representation issues however could influence the prediction in case Wikipedia pages are visited by a sample of people not representative of the electorate. A third reason for biased social media data has also been mentioned by Yasseri et al. Following their proposed rational choice approach, they theorize that page views could be biased by voters seeking information on new political parties more frequently than on already established parties. A similar phenomenon has been implicitly described by a study [Jungherr et al., 2011] attempting to replicate an earlier study [Tumasjan et al., 2010] that claimed to be able to successfully predict the results of the German Election of 2009 using Twitter data. The original study only included the larger German parties, while the replication study also included the small "Pirate Party". This resulted in a predicted win for the Pirate Party, which did not happen in reality. The cause of the overprediction for this party may be the fact that this party was fairly new at that time or that it was controversial, causing it to be mentioned more than other parties without it having a proportional effect on election outcomes.

The frequent visits to Wikipedia pages during elections can be analyzed as a trend in the context of existing research on social media. Crane et al. introduced a model that can be utilized to classify the impact of an election on page views as a specific type of trend [Crane and Sornette, 2008]. Applying this model, the social system is Wikipedia, and the external event is the election. Based on visualizations by Yasseri et al. (Figure 1), the peak of the trend, which corresponds to the days with the highest page views, accounts for more than 80% of the total page views during the analyzed time period. According to Crane et al., this trend qualifies as Class 1 and is considered exogenous subcritical. While it is debatable whether Wikipedia can be considered a social system apart from the creation of articles, classifying the trend dynamics itself as exogenous subcritical is reasonable as there is less reason to research the specifics of the election once it has been held, resulting in a fast decay of page views to a significantly lower level compared to the peak.

One motivation for the study by Yasseri et al. is a described lack of papers providing theories containing explanations for why digital traces have predictive power [Lazer et al., 2009], which they are counteracting by providing a rational choice theory of information seeking. The authors explanation of this theory is limited, and they do not provide sources for what they are basing their theory on. Brief literature research reveals that rational choice theory has mainly been used to investigate the decision to vote itself, which led to a paradox known as the "voting paradox", describing the result that voting is irrational when the individual wants to optimize personal utility [Bendor et al., 2003, Martorana, 2011]. One paper claims that "not much could be learned by taking a purely rational choice approach" when searching for explanations on why people seek political information during elections [Ohr and Schrott, 2001]. Instead, Ohr et al. propose a mixed approach that considers the social context in which voters are situated, as well as external and internal expectations that drive information-seeking behavior. They emphasize that information seeking serves as a rational tool to fulfill the aforementioned expections but also to enable the expressive component of voting and to engage with the entertaining aspects of elections. The proposed theory shares the same outcome as the theory by Yasseri et al., which is the decision to seek information. However, it presents an alternative and more comprehensive framework for understanding the predictive nature of Wikipedia page views in relation to election outcomes.

After Yasseri et al., two other papers [Salem and Stephany, 2021, Smith and Gustafson, 2017] have used Wikipedia page views to predict elections. Both studies do not find that Wikipedia pageviews can independently predict electoral outcomes. Instead, they propose using Wikipedia pageviews as a complementary measure to enhance predictions made by other models, as it can explain variance that conventional information sources cannot. Yasseri et al. set the groundwork for this by acknowledging that new parties and swing voters bias the page views.

# 3 Data

This project aims to analyze the relationship between Wikipedia page views and turnout in the EU elections of 2009, 2014, and 2019. Included countries are the same as in the paper by Yasseri et al. To achieve this, two key metrics are required: Wikipedia page views for the relevant Wikipedia pages on the elections in the included countries, and the election turnout data. To obtain these metrics, the project retrieves data from two sources: Wikipedia Page View Dumps and the official EU website containing election results. The Wikipedia page views are accessed through downloadable files, as the API Wikimedia provides for this purpose only provides data from 2015 onwards.

These files contain page views from all Wikimedia projects in all languages and can be quite large. In 2019, for example, they are 400-500 MB uncompressed and 4 GB uncompressed. The data before December 2011 has a different format both in content and on the time basis the files exist at. This necessitates separate data processing steps to handle these differences. Another challenge is that article names change over time and the page view files don't include something like an identifier that is consistent over time. To get the page names, the translation function on the English page for an election is used. I considered implementing the MediaWiki Action API to find out whether article names changed and then to automatically correct the names, but this would also have been beyond the scope of this project. Unfortunately, the authors did not include the names of the articles with their paper, but they also write that there were difficulties in finding the names of the articles. One choice is between an election article on the general election and the article of the election in the corresponding country. Like the authors, I chose the article with the largest number of page views.

To process the page view dump files, the project uses the Python framework and cloud computing provider modal.com. Intermediate results are stored on a remote S3 server. Files corresponding to 14 days before and after the election date are downloaded and processed concurrently, handling 100 files at a time. The time span of 28 days is chosen to limit the amount of data to be processed as it would have been too resource and time-consuming for the scope of this project otherwise. The files are then decompressed in chunks and filtered line by line to include only statistics for articles within the Wikipedia project. Each filtered chunk is then written to a parquet file. To query the election article page views, a manually compiled dictionary containing the names of the election articles in the included at the election dates is used. One SQL query per country per file is made using an in-memory duckDB database using python code in the following manner:

```
temp = duckdb.query(f"""
SELECT *
FROM '{input_filepath}'
WHERE  article_title LIKE '{page_name}' AND wikicode = '{wikicode}'
```

The results are stored in separate Pandas dataframes, with each row representing the hourly page views for the Wikipedia election page in a specific country. As all the steps are performed concurrently, the next step involves concatenating these dataframes. Moreover, the data is stacked to transform each row into a daily count, and then pivoted so that each column corresponds to a country and contains a time series of page views. Additionally, a new column is added to contain page views that are normalized relative to the defined time span, ensuring that the time series are comparable.

# 4 Methods

In this replication study, I utilize methods that overlap with those employed in the paper by Yasseri et al. The objective is to analyze the relationship between Wikipedia page views and turnout in the EU elections by employing the Pearson Correlation Coefficient. In the analysis, I consider the relative change in page views as the independent variable and the relative change in turnout between two elections as the dependent variable. While Yasseri et al. do not provide explicit details on how they calculated the "volume of attention in the build-up phase" used to determine the relative change in page views, I assume it refers to the summation of all page views within a predetermined timeframe preceding the election date. The formula for relativ change is obviously:

$$\text{Relative Change} = \frac{\text{new value} - \text{initial value}}{\text{initial value}}$$

Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations, but in this project this is automated using the `scipy.stats.pearsonr` function. The Pearson Correlation Coefficient is a commonly used statistical measure that helps assess the strength and direction of relationships between variables. It is valuable for understanding the level of association between two variables and can be utilized to uncover patterns or trends within the data. The strength of the relationship, indicated by how close the value is to 1 (or -1), can validate or refute the hypothesis that Wikipedia page views are a predictor for turnout.

# 5 Results

Comparing the lineplot of page views from 2009 as shown in the paper by Yasseri et al. (Figure 1) to the lineplot recreated in Figure 1 of this report, it is evident that the two lineplots exhibit close similarity. The peak in page views aligns with the date of the 2009 European Parliament Elections, indicating a clear correlation between the election and the increased interest in related Wikipedia articles. However, when visualizing the data from 2014 and 2019 in the same way, it becomes apparent that the patterns differ from those observed in the 2009 data. The peaks and fluctuations in page views do not align as closely with the election dates in these years.

Taking a closer look at the values of Hungary, a country where the peak does not correspond to the election date, it becomes apparent that there could be errors in the data or its retrieval, as page views are at around zero after three seemingly random peaks till the 15th of June. Further analysis has to be performed under the assumption that the data is erroneous.

The correlation analysis conducted on the relative changes of page views and turnout from 2009 to 2014 and from 2014 to 2019 did not yield statistically significant results. For the time period of 2009 to 2014, the correlation coefficient was calculated to be -0.2884. However, with a p-value of 0.3392, the correlation is not statistically significant. Similarly, for the time period of 2014 to 2019, the correlation coefficient

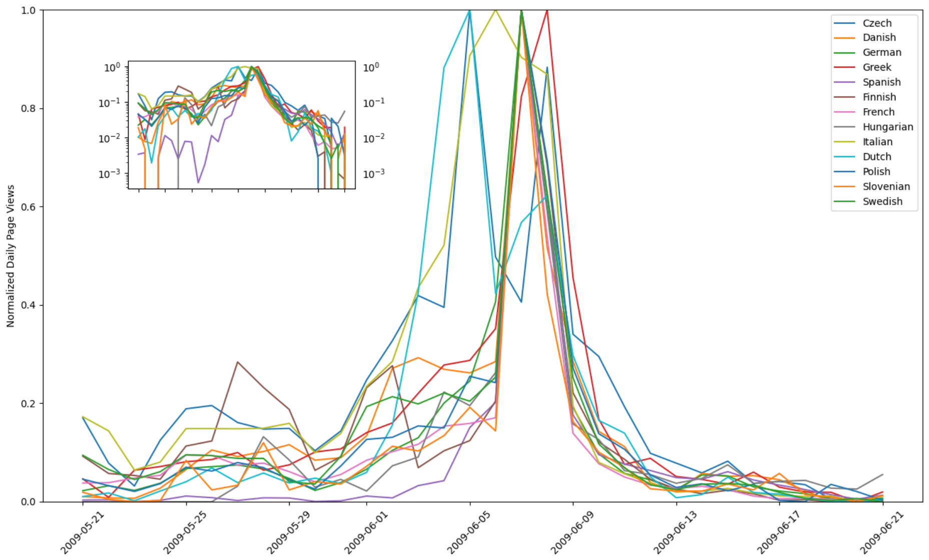

**Figure 1** 'Normalized Wikipedia Election Page Views two weeks before and after the 2009 European Parliament Elections'

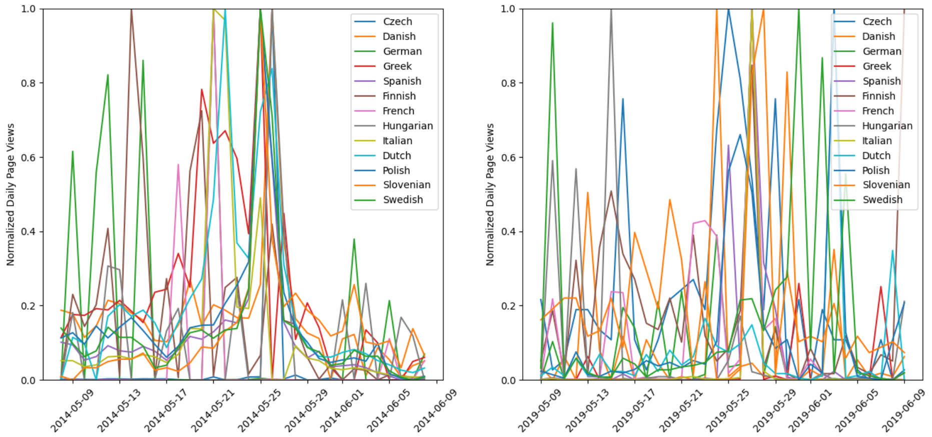

**Figure 2** 'Normalized Wikipedia Election Page Views two weeks before and after the 2014 (left) and 2019 (right) European Parliament Elections'

was found to be 0.3662. However, the corresponding p-value of 0.2184 suggests that this correlation is also not statistically significant. Based on these findings, there is no strong evidence to support a significant correlation between the relative changes in page views and turnout during these time periods.

This is supported by visualizing the respective relative changes in a scatter plot visible in Figure 3.

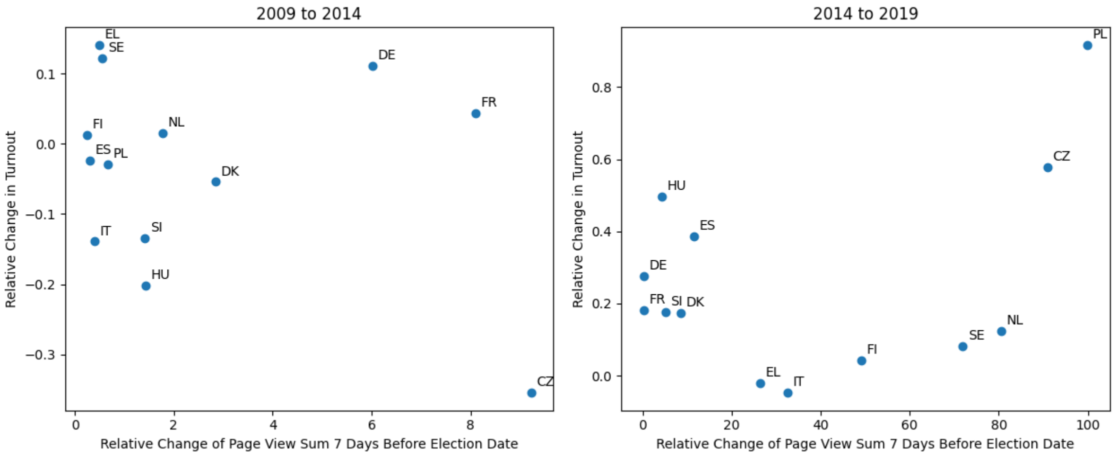

**Figure 3** "Scatterplot of Relative Changes in Page Views and Turnout"

# 6 Discussion

Unfortunately, there are no conclusive results from this study as the data retrieval process most likely contains errors. For these reasons, the reproduction of the paper by Yasseri et al. was not successful.

However, two conclusions can still be drawn from this circumstance. Firstly, if scientific papers truly want to be replicable, they should either include all of the data, or describe the data retrieval process in more detail. A git repository is the ideal way to share code.

Moreover, The Wikipedia API should be used to retrieve data instead of manually downloading page view dump files, as APIs already include filter options that make it more probable that the data is not erroneous. Also, data quality before the introduction of the page view API is not ensured, as data was compiled by different people in different formats [wikimedia.org, nd].

# Bibliography

[Bendor et al., 2003] Bendor, J., Diermeier, D., and Ting, M. (2003). A Behavioral Model of Turnout. *American Political Science Review*, 97(2):261–280.

[Crane and Sornette, 2008] Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. U.S.A.*, 105(41):15649–15653.

[DiGrazia et al., 2013] DiGrazia, J., McKelvey, K., Bollen, J., and Rojas, F. (2013). More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PLoS One*, 8(11):e79449.

[Gayo-Avello et al., 2011] Gayo-Avello, D., Metaxas, P., and Mustafaraj, E. (2011). Limits of Electoral Predictions Using Twitter. *ICWSM*, 5(1):490–493.

[Ginsberg et al., 2009] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

[Goel et al., 2010] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci. U.S.A.*, 107(41):17486–17490.

[Jungherr et al., 2011] Jungherr, A., Jürgens, P., and Schoen, H. (2011). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. “Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment”. *Social Science Computer Review*, 30(2):229–234.

[Lazer et al., 2014] Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science (New York, N.Y.)*, 343(6176):1203–5.

[Lazer et al., 2009] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational Social Science. *Science*, 323(5915):721–723.

[Martorana, 2011] Martorana, M. F. (2011). Voting Behaviour in a dynamic perspective: a survey. *University Library of Munich, Germany.*

[Ofcom, 2015] Ofcom (2015). Adults’ media use and attitudes report 2015. [Online; accessed 22. Aug. 2023].

[Ohr and Schrott, 2001] Ohr, D. and Schrott, P. R. (2001). Campaigns and Information Seeking: Evidence from a German State Election. *European Journal of Communication*, 16(4):419–449.

[Rainie, 2020] Rainie, L. (2020). Wikipedia users. *Pew Research Center: Internet, Science & Tech.*

[Rousidis et al., 2020] Rousidis, D., Koukaras, P., and Tjortjis, C. (2020). Social media prediction: a literature review. *Multimed. Tools Appl.*, 79(9):6279–6311.

[Salem and Stephany, 2021] Salem, H. and Stephany, F. (2021). Wikipedia: a challenger's best friend? Utilizing information-seeking behaviour patterns to predict US congressional elections. *Information, Communication & Society*, pages 174–200.

[Similarweb, 2023] Similarweb (2023). Top websites ranking. [Online; accessed 22. Aug. 2023].

[Smith and Gustafson, 2017] Smith, B. K. and Gustafson, A. (2017). Using Wikipedia to Predict Election Outcomes: Online Behavior as a Predictor of Voting. *Public Opin. Q.*, 81(3):714–735.

[Tumasjan et al., 2010] Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, 4(1):178–185.

[wikimedia.org, nd] wikimedia.org (n.d.). Wikistats: Page View complete dumps. `https://dumps.wikimedia.org/other/pageview_complete/readme.html`. [Online; accessed 26. Sep. 2023].

[Yasseri and Bright, 2016] Yasseri, T. and Bright, J. (2016). Wikipedia traffic data and electoral prediction: towards theoretically informed models. *EPJ Data Sci.*, 5(1):1–15.