TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

FORECASTING FUTURE EVENTS WITH PUBLICLY ACCESSIBLE ONLINE DATA

A Study on Finnish Parliamentary Elections from 2015 to 2023

Tapio Vepsäläinen

TURUN YLIOPISTON JULKAISUJA ANNALES UNIVERSITATIS TURKUENSIS
SARJA SER. E OSA TOM. 128 | OECONOMICA | TURKU 2025

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

FORECASTING FUTURE EVENTS WITH PUBLICLY ACCESSIBLE ONLINE DATA

A Study on Finnish Parliamentary Elections from 2015 to 2023

Tapio Vepsäläinen

TURUN YLIOPISTON JULKAISUJA ANNALES UNIVERSITATIS TURKUENSIS
SARJA SER. E OSA TOM. 128 | OECONOMICA | TURKU 2025

University of Turku

Turku School of Economics
Department of Management and Entrepreneurship
Information Systems Science
Doctoral Program, Turku School of Economics

Supervised by

Professor Reima Suomi
Department of Management and Entrepreneurship
Turku School of Economics,
University of Turku, Finland

Associate Professor Hongxiu Li
Faculty of Management and Business
Tampere University, Finland

Reviewed by

Professor Marko Joas
Faculty of Social Sciences, Business and Economics
Åbo Akademi, Finland

Professor Dr. Dr. h.c. Jörg Becker
Department of Information Systems
University of Münster, Germany

Opponent

Professor Dr. Dr. h.c. Jörg Becker
Department of Information Systems
University of Münster, Germany

ISBN 978-952-02-0188-3 (PRINT)
ISBN 978-952-02-0189-0 (PDF)
ISSN ISSN 2343-3159 (PRINT)
ISSN ISSN 2343-3167 (ONLINE)
Painosalama, Turku, Finland, 2025

To Leo and Ella, for the joy they bring to my life,
And to Annu, for her unwavering belief in me

UNIVERSITY OF TURKU
Turku School of Economics
Department of Management and Entrepreneurship
Information Systems Science
VEPSÄLÄINEN, TAPIO: Forecasting Future Events with Publicly Accessible Online Data
Doctoral dissertation, 184 pp.
Doctoral Program, Turku School of Economics
December 2025

ABSTRACT

Publicly accessible online data has become an increasingly feasible data source for predictive analytics. This thesis explores how digital footprints, such as social media interactions, can be used in forecasting. Focusing specifically on elections as an example application, the thesis presents a series of models used to forecast the outcome of Finnish parliamentary elections. By evaluating the precision and limitations of the models in electoral forecasting, this research seeks to bridge the gap between data science and information systems research, offering insight into the broader impact of digital data utilization in societal decision-making contexts.

The methodology employed in this study combines predictive modeling and data science techniques. The research integrates publicly available data, such as social media interactions and online content, to train models capable of forecasting electoral results. The approach is built on multiple original studies, each exploring different facets of election prediction. The robustness and practical utility of the predictions are assessed through real-world testing, involving the publication of forecasts prior to elections. In addition, the interpretability of the models is analyzed to understand whether the results align with political theories. Ethical considerations, such as privacy and data ownership, are also carefully examined throughout the study.

The key findings of this study demonstrate the potential of using publicly available online data to forecast election outcomes. The forecasting models evolved significantly during the three election cycles studied (2015, 2019, and 2023). The final model integrates diverse data sources, including social media interactions, electoral history, and candidate attributes. Progressive improvements in accuracy were observed throughout the study, and the models eventually approached the precision of traditional polling methods. The study underscores the incremental benefits of incorporating diverse data types while addressing the challenges associated with data collection and feature selection. Although current models exhibit robust predictive capabilities, their practical applicability compared to opinion polls is limited. However, the results suggest that there is substantial promise for future enhancements.

The research advances the field of election forecasting by introducing a methodology that leverages publicly accessible candidate data alongside social media insights, offering a candidate-level perspective on electoral predictions. This approach not only complements traditional macro-level methods, but also provides insights towards understanding the theoretical foundations of voting behavior. Although the potential of social media as a predictive tool is highlighted, the research acknowledges

existing challenges such as bias, suggesting mitigation strategies, and underscoring the importance of domain knowledge in data-driven research. Practically, the study suggests that hybrid methodologies that combine traditional polling with candidate-specific insights can improve prediction precision. Additionally, it emphasizes the significance of cross-disciplinary understanding and transparent decision-making in refining methodologies for predictive analytics using online data. Overall, the research highlights the need for a holistic approach in utilizing digital data, balancing technical proficiency with ethical and contextual awareness.

KEYWORDS: Machine Learning, Social Media, Elections, Predictive Analytics, Data Science, Election Forecasting

TURUN YLIOPISTO
Turun kauppakorkeakoulu
Johtamisen ja yrittäjyyden laitos
Tietojärjestelmätiede
VEPSÄLÄINEN, TAPIO: Forecasting Future Events with Publicly Accessible Online Data
Väitöskirja, 184 s.
Turun kauppakorkeakoulun tohtoriohjelma
Joulukuu 2025

TIIVISTELMÄ

Yhteiskunnan digitalisaatio sekä teknologian nopea kehitys ovat yhdessä luoneet uusia mahdollisuuksia tutkia ja ennustaa yhteiskunnallisia ilmiöitä. Monet tutkijat ovat ehdottaneet, että digitaalisia aineistoja, kuten median käytön yhteydessä tallentuvia jälkiä, voitaisiin hyödyntää ennustemallien kehityksessä. Tämä väitöskirja tarkastelee ilmiötä vaaliennusteiden avulla. Väitöskirjassa esitellään sarja ennustemalleja, joiden tarkkuutta ja rajoituksia arvioimalla pyritään ymmärtämään verkossa julkisesti saatavilla olevien aineistojen ennustamisessa hyödyntämisen edellytyksiä. Väitöskirja tuo yhteen tietojärjestelmätieteen, politiikan tutkimuksen sekä tietojenkäsittelytieteen näkökulmia, tarjoten arvokkaita näkemyksiä digitaalisten aineistojen hyödyntämisen mahdollisuuksista.

Metodologisesti tutkimus hyödyntää ensisijaisesti määrällistä lähestymistapaa ilmiöiden ennustamiseen. Tutkimuksessa hyödynnetään julkisesti saatavilla olevia aineistoja, kuten ehdokkaiden sosiaalisen median seuraajamääriä, vaalikonevastauksia ja muuta ehdokkaisiin liittyvää verkkosisältöä mallien kouluttamiseen. Väitöskirja on kokoelmatyö, joka yhdistää useamman alkuperäisen artikkelin tulokset. Työhön sisältyvät tieteelliset julkaisut käsittelevät vaalien ennustamisen eri näkökulmista. Tutkimuksen yhteydessä on myös julkaistu artikkeleita yliopiston verkkosivuilla, joissa ennusteita on esitelty ennen vaaleja. Julkaistujen ennusteiden avulla pyritään arvioimaan menetelmän luotettavuutta sekä käytännön hyötyjä. Tutkimuksessa huomioidaan myös eettiset näkökulmat, kuten ehdokkaiden yksityisyys.

Kokonaisuudessaan väitöskirja kertoo monipuolisen tarinan julkisesti saatavilla olevien aineistojen hyödyntämisestä vaalituloksien ennustamiseen. Ennusteet laadittiin kolmien perättäisten eduskuntavaalien yhteydessä (2015, 2019 ja 2023). Tutkimusperiodin aikana lähestymistapa kehittyi jatkuvasti, lopulta keskittyen laajojen aineistojen analysointiin koneoppimismenetelmien avulla. Tulosten tarkastelussa korostuu monipuolisten aineistojen hyödyntämisen edut ja haasteet. Vaikka tutkimusperiodin aikana ennusteiden tarkkuus jäi perinteisistä mielipidemittauksiin pohjautuvista ennusteista, vaikuttaa lähestymistapa varteenotettavalta.

Tutkimus edistää vaalien ennustamiseen liittyvien menetelmien kehitystä, samalla luoden pohjaa laajemmalle aineistojen hyödyntämiseen liittyvälle ymmärrykselle. Vaalien ennustamisessa hyödynnettävät menetelmät keskittyvät yleensä suhteellisen rajattuun kohteeseen. Esimerkiksi eduskuntavaaleissa mielipidemittaukset keskittyvät yleensä arvioimaan puolueiden kannatusta. Väitöskirjassa esitellyt menetelmät tuottavat tietoa jokaisesta ehdolla olevasta ehdokkaasta, ja siten tarjoavat hienojakoi-

semman näkökulman äänestyskäyttäytymisen tarkasteluun kuin mielipidemittaukset.

Väitöskirja pyrkii myös tuomaan esiin lähestymistapaan liittyviä haasteita. Esimerkiksi sosiaalisesta mediasta kerätty aineisto ei vastaa satunnaisotoksella kerättyä aineistoa, eikä sen siten voida ajatella edustavan koko yhteiskuntaa. Erilaisten vinoumien ja haasteiden käsittely edellyttää poikkitieteellistä lähestymistapaa. Viimekädessä väitöskirja korostaa tarvetta kokonaisvaltaiselle lähestymistavalle, joka ottaa huomioon tutkittavan ilmiön kontekstisidonnaisuuden, teknisen osaamisen sekä eettiset näkökulmat.

ASIASANAT: Koneoppiminen, Sosiaalinen media, Vaalitutkimus, Ennakoiva analytiikka, Data-analytiikka, Vaalien ennustaminen

Acknowledgements

My journey with this research topic began more than a decade ago. Throughout the research period, there were several periods of reduced activity as I pursued various professional opportunities outside academia and focused on my hobbies and family life. Despite interruptions, I maintained a commitment to complete this endeavor. After years of persistence and adaptation to the evolving research environment, the dissertation is finally complete.

Completing this dissertation would not have been possible without the guidance, support, and encouragement of numerous individuals who contributed to both my academic development and my personal well-being throughout this journey. I express my sincere gratitude to all those who have played a role in making this achievement possible.

Foremost, my deepest gratitude goes to my supervisor, Professor Reima Suomi, whose patient guidance, extensive experience, and trust in my ability to complete this dissertation were invaluable. The consistent mentorship throughout this journey provided me with a needed foundation to bring this project to completion. Similarly, I express my deepest gratitude to my second supervisor, Associate Professor Hongxiu Li from Tampere University, whose meticulous feedback on our joint projects and willingness to provide assistance whenever needed significantly enhanced the quality of my work.

My sincere appreciation also to the preexaminers, Professor Marko Joas from Åbo Akademi and Professor Jörg Becker from the University of Münster. Their rigorous evaluation and approval of my work from both political science and information systems perspectives confirms that I have succeeded in combining these different academic disciplines. I am additionally honored that Professor Becker has agreed to serve as the esteemed opponent in my doctoral defense, bringing his considerable expertise to the final examination of this work.

Despite my research topic being somewhat unconventional within the field of Information Systems Science, I was consistently welcomed and supported by all of my colleagues. I wish to express my heartfelt gratitude to other researchers I had the opportunity work with, particularly Associate Professor Jonna Järveläinen for her enthusiasm and constructive feedback, Professor Hannu Salmela for his openness to engaging in scholarly discussions and providing guidance when ever needed, and Professor Jukka Heikkilä for providing distinct perspectives to my work. I also

extend my appreciation to all other members of the Information Systems Science community of Turku, past and present, who contributed to a stimulating research environment.

One of the greatest parts of my doctoral journey was getting to know fellow doctoral candidates. There were many amazing individuals with whom I spent time with, and few of them have become lifelong friends. I am especially grateful to Markus Zimmer for his companionship and support throughout this process; his exceptional dedication to academic excellence is truly inspiring. I am also thankful to Juho Vaiste for his peer support and willingness to discuss any aspect of my research journey, as well as Anne-Marie for her consistent moral support during our shared time at the department.

Beyond my immediate academic circle, I was fortunate to encounter several inspiring scholars across the University of Turku. Especially Professor Emeritus Kari Lukka provided intellectual inspiration throughout the journey. His insights into the philosophy of science and pursuit of research excellence greatly influenced my appreciation of philosophical perspectives in research.

The Kilpisjärvi Information Systems Seminars also deserve a special mention. These adventures in Lapland not only provided opportunities to present my work but also created lasting bonds with fellow researchers. I am especially grateful to Reima for organizing these, and to Brita, whose warmth and encouragement at these seminars created a welcoming atmosphere for everyone. Similarly, being part of the team organizing the Well-being in the Information Society conferences provided opportunities to connect with researchers focusing on the societal impacts of technology. I would also like to thank all the team members of TSE Jukola team who shared countless memorable experiences in Jukola relays across Finland. Particularly, Eija, for her dedicated leadership and organizing the team over the years.

Interdisciplinary collaboration outside the University of Turku also played an important role in my research. I am grateful to researchers from the field of political science who welcomed me into interdisciplinary discussions. Thank you, Aleksi and Veikko, for our thoughtful exchanges and for inviting me to participate in IntraComp research seminars, as well as other political scientists who shared their thoughts about my research. These interdisciplinary interactions significantly broadened my perspective and enriched my research approach.

Throughout my doctoral studies, I maintained professional commitments in the software industry. I am deeply grateful to Jarkko for facilitating flexible working arrangements that allowed me to balance academic and professional responsibilities. Without this accommodation, completing this dissertation would have been considerably more challenging.

This dissertation would not have been possible without financial support from various sources. I am particularly grateful to Liikesivistysrahasto and Wallenbergin säätiö for their substantial support. I also wish to thank the University of Turku and

the Turku School of Economics doctoral program for providing a finishing grant that enabled me to complete this dissertation. Additional financial support was provided by Turun kauppakorkeakoulun tukisäätiö through the Anja ja Erkki Toivasen rahasto, the TS-yhtymän stipendirahasto and the Sampo-yhtiöiden stipendirahasto. This diverse funding allowed me to dedicate sufficient time and resources to this research project.

Finally, I wish to express my profound gratitude to my family and friends. To my wife, Annu, whose unwavering support and understanding during this journey have been my foundation; to my parents, who instilled in me the value of education and perseverance; to my children, Leon and Ella, who are always a source of joy and motivation; and to my friends for their support, encouragement, and for their steadfast belief in my abilities. In particular, I am grateful to Ville for sharing his enthusiasm and knowledge of data science and for our valuable discussions.

11.05.2025
Tapio Vepsäläinen

TAPIO VEPSÄLÄINEN
is a researcher and entrepreneur interested in the intersection of data science, information systems, and political science. His research focuses on election forecasting using digital footprints. He is driven by creating real value through digital innovations by applying emerging technologies to real-world challenges. Beyond his academic work, Tapio works as an entrepreneur and consultant in software development, where he leverages artificial intelligence and data science techniques to build innovative digital products and solutions.

Appendices

Table of Contents

Abbreviations

ML	Machine Learning
AI	Artificial Intelligence
DSR	Design Science Research
IS	Information Systems Science
DK	Design Knowledge
RMSE	Root mean squared error
MAE	Mean absolute error
CRISP-DM	Cross-Industry Standard Process for Data Mining
LLM	Large language model
PVEA	Personal Vote Earning Attribute
KESK	Finnish Centre (Keskusta)
VIHR	Greens (Vihreät)
SFP	Swedish People's Party (Svenska folkpartiet)
SDP	Social Democratic Party (Sosialidemokraattinen puolue)
VAS	Left Alliance (Vasemmistoliitto)
KOK	National Coalition Party (Kansallinen Kokoomus)
PS	Finns Party (Perussuomalaiset)
KD	Christian Democrats (Kristillisdemokraatit)
EDA	Exploratory Data Analysis
API	Application programming interface
HTTP	HyperText Transfer Protocol
HTML	Hypertext Markup Language
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
JSON	JavaScript Object Notation
XML	Extensible markup language
SOAP	Simple Object Access Protocol
RESTful	Representational State Transfer
GraphQL	Graph Query Language
JWT	JSON Web Token
SPA	Single-page Application
SSR	Server-side Rendering
GDPR	General Data Protection Regulation

CCPA	California Consumer Privacy Act
PII	Personally identifiable information
VP	Vote-popularity
LPR	List proportional representation
OLPR	Open-list proportional representation
PR	Proportional representation
FPTP	First-past-the-post
IMF	International Monetary Fund
OECD	Organization for Economic Co-operation and Development
SHAP	Shapley Additive exPlanations
STV	Single transferable vote
US	The United States of America
UK	The United Kingdom

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Tapio Vepsäläinen, Hongxiu Li & Reima Suomi. Understanding Facebook and public opinion: Predicting Finnish parliament elections 2015. Government Information Quarterly, 2017; Volume 34, Issue 3, Pages 524-532

II Tapio Vepsäläinen, Hongxiu Li & Reima Suomi. The role of social media platforms in forecasting elections: a comparison of Twitter and Facebook. Digital Government: Research and Practice, 2024; Volume 5, Issue 3, Pages 1 - 15.

III Tapio Vepsäläinen. Predicting Candidate Votes in Multiparty Elections. Submitted to the International Journal of Forecasting.

1 Introduction

1.1 Background and Motivation

Humans have an intrinsic desire to predict future events, a characteristic that has played a pivotal role in our success as a species. While the past cannot be altered, humans are naturally inclined to contemplate and try to shape the future (Baumeister et al., 2016). This tendency for foresight allows people to focus on possible outcomes and guide actions toward achieving favorable results (Baumeister et al., 2016). Forecasting, as a formalized embodiment of this inherent trait, has evolved into a sophisticated field that helps manage uncertainties and optimize decision making around the world (Kuhn and Johnson, 2013).

The integration of forecasting into modern decision-making processes reflects its growing importance in addressing global challenges (Petropoulos et al., 2022). As the world faces complex uncertainties, ranging from pandemics to climate change, the ability to quantify and visualize these uncertainties is essential (Petropoulos et al., 2022). Research and advancing forecasting methodologies is more essential than ever as it equips decision-makers with the tools needed to navigate an increasingly unpredictable future.

The core focus of this thesis is an examination of how publicly accessible digital data can be harnessed to predict the future. Many researchers have proposed that publicly accessible digital data, such as social media interactions, search queries, and online purchasing behaviors, provide valuable insights into human behavior and societal trends, thus improving predictive models (Lazer et al., 2021; Murphy et al., 2014; Salganik, 2019).

In recent years, harnessing digital data for predictive analytics has become increasingly feasible due to recent advances in computational statistics, machine learning, artificial intelligence, and big data technologies (Buyalskaya et al., 2021; Karpatne et al., 2017). These advances, combined with interdisciplinary collaboration, are leading to new ways of studying society (Lazer et al., 2009; Buyalskaya et al., 2021). Consequently, there has been consistent growth in interdisciplinary publications that adopt these approaches, reflecting a trend toward integration of research agendas (Porter and Rafols, 2009; Zuo and Zhao, 2018; Mao et al., 2019).

Information Systems (IS) research has also increasingly addressed the role of extracting insights and knowledge from large-scale data and the use of advanced scientific methods within its disciplinary framework (Agarwal and Dhar, 2014). At

its core, IS is situated at the intersection of information technology, human behavior, and social context (Gregor, 2006; Thatcher et al., 2018). As digital technology has become ubiquitous, the potential scope of IS research has grown significantly. This expansion encompasses not only the technical and methodological aspects of leveraging digital data but also the societal implications and ethical considerations surrounding its use.

The growing interest in the use of publicly available digital data for scientific purposes comes from both the challenges of traditional methods and the opportunities presented by new approaches. Traditional data collection techniques, such as surveys and interviews, often struggle with scale, timeliness, and self-reporting biases (Couper, 2000; Groves, 2011). In contrast, digital platforms generate vast amounts of real-time data, which can be used to address these limitations (Lazer et al., 2021). The use of diverse data streams in predictive analytics has the potential to provide precise and timely insights and to enhance decision making in numerous sectors (Chen et al., 2012). In contemporary discussion, this approach is often encapsulated under the term data science.

Despite the potential to leverage digital tracers for predictive analytics, several challenges remain (Lazer et al., 2021). Access to large volumes of data does not automatically lead to better predictions (Kitchin, 2014). Overfitting can occur when models mistakenly interpret noise as meaningful patterns, resulting in inaccurate conclusions (Hawkins, 2004). Furthermore, results may sometimes be redundant, as similar results could be achieved with much simpler baseline models rather than employing complex algorithms (Goel et al., 2010).

The use of contemporary data science has been criticized for emphasizing predictive modeling over traditional theory building (Agarwal and Dhar, 2014). Although these advanced analytical techniques offer powerful tools for identifying patterns, they may sometimes lack a theoretical and domain-specific foundation, potentially limiting the interpretability and applicability of their findings (Karpatne et al., 2017). For example, a machine learning model might identify patterns in social media usage but may not account for the sociocultural factors that influence these patterns.

Addressing these challenges requires going beyond data collection and analysis. Domain expertise is necessary to appropriately contextualize the data and foster that the algorithms are aligned with specific problems (Salganik, 2019; Martínez-Plumed et al., 2019). The quality of predictions often depends on the quality of the input data, highlighting the importance of filtering and curating the data to minimize errors. A common phrase among the field is ”garbage in, garbage out”(Kilkenny and Robinson, 2018). This involves balancing focus on insights from data with domain knowledge, alongside continuous refinement of analytical methods and consideration of ethical implications regarding data use (Boyd and Crawford, 2012).

The use of theory in data science is gaining attention because it helps form hypotheses and provides a framework for identifying and quantifying important vari-

ables in large datasets (Lazer et al., 2021). This approach aligns with practices in IS research, where both theoretical and practical aspects are valued (Gregor, 2006; Iivari, 2007). Consequently, there is a growing interest in integrating data-scientific approaches into IS research, as these methods can enhance the ability of researchers to generate insightful results (Agarwal and Dhar, 2014).

1.2 Research Objectives

Building on a broader examination of how publicly accessible digital data can predict future actions, this thesis focuses on one particularly compelling application: election forecasting.

Election forecasting benefits from explicit and well-defined outcomes, which provide a clear basis for evaluating the effectiveness of predictive models. Unlike other predictive scenarios that can involve subjective interpretations, election results are definitive, measurable, and time-bound, allowing for precise evaluations and comparisons between predicted and actual outcomes.

Consequently, electoral forecasting stands out as a significant area for using digital footprints in predictive analytics (Brito et al., 2021; Phillips et al., 2017; Skoric et al., 2020).

The overarching research question that guides this study is the following:

"How accurately can publicly available online data predict election outcomes?"

This question seeks not only to measure the accuracy of the prediction, but also to explore the characteristics of the underlying data and the methodological approaches that influence these predictions.

Several factors currently facilitate the study of this research question. Candidates actively use social media profiles, campaign websites, public statements, and other digital platforms to make themselves highly visible (Zittel, 2016; Maddens et al., 2006; Strandberg et al., 2024). In addition, there is a wealth of readily accessible information on previous elections, government performance, economic conditions, and incumbents (Islam, 2006). This abundance of data provides a contextual backdrop for understanding current electoral dynamics.

Predictive analytics emerges as an ideal methodological framework for analyzing this research question. Predictive analytics involves the analysis of large datasets, such as those derived from social media activities, web search trends, government databases, and digital news articles, to identify patterns that can effectively predict future events (Kuhn and Johnson, 2013; Shmueli and Koppius, 2011).

The popularity of data science, and subsequently predictive analysis, has increased significantly in recent years (Abbasi et al., 2023). However, despite its potential, predictive analytics remains underrepresented in the field of IS, where it is rarely utilized in mainstream empirical research (Shmueli and Koppius, 2011). This

thesis aims to bridge this gap by illustrating the applicability of predictive analytics and assessing its impact within a key societal domain.

By focusing on the context of Finnish elections, this research aims to understand the broader implications of the use of digital data in predictive models and to evaluate their effectiveness and limitations in a real-world context. In doing so, the study seeks to answer questions about the ability of digital footprints to reflect societal behaviors and decisions. The study is positioned at the intersection of data science, online data, and election forecasting, utilizing data mining techniques, trend and pattern recognition, and predictive modeling, as illustrated in Figure 1.

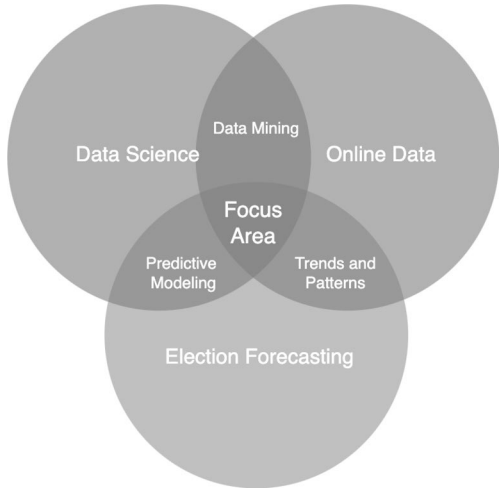

Figure 1. Research Focus

In this study, the development of predictive models for the forecast of Finnish parliamentary elections in 2015, 2019, and 2023 serves as a practical approach. The models presented harness the diverse nature of online data, exploring various types of data and sources to provide an assessment of their predictive power. The aim is to provide information on the success factors and limitations of such models.

By weaving these considerations into the research, the study aims to contribute to the broader discourse on the use of digital data to improve predictive capabilities in various domains, spotlighting the opportunities and challenges inherent in such approaches. Subsequently, this research seeks to inform both academic inquiry and practical applications, ensuring that the predictive models developed are rigorous and relevant to contemporary electoral contexts.

1.3 Structure of the Thesis

This thesis is organized into a series of interconnected chapters designed to provide a comprehensive exploration of how publicly accessible digital data can be utilized

for election forecasting. Rooted in the rapidly advancing fields of machine learning, artificial intelligence, and big data, it focuses on leveraging these technologies in the context of Information Systems (IS) research to predict election outcomes, particularly within the Finnish democratic system.

Chapter 2, "The Data-Driven Research Agenda," lays the foundation by exploring the integration of digital data into research processes. The chapter discusses the interconnection of online and offline data, typical research traditions, and key methodological approaches in data science. This chapter serves as a primer on the current landscape, setting the stage for subsequent analyses and applications.

Chapter 3, "Utilization of Online Data," delves into the sources and characteristics of online data that are pivotal for predictive analytics. The chapter describes how social networks, open data, and other domain-specific resources can be harnessed for research purposes, emphasizing the methods to retrieve and integrate these data types. The chapter also highlights the potential and pitfalls of using such data, including ethical implications and quality considerations.

In Chapter 4, "Forecasting Elections," the focus shifts to applying data-driven methodologies to predict electoral outcomes. This chapter deals with traditional and emerging methodologies in political science, examining their application in election forecasting. It provides a detailed overview of political theories, electoral systems, and methods to leverage digital data, with a specific focus on the Finnish electoral system.

Chapter 5, "Research Methodology," provides insight into the philosophical and methodological considerations underpinning this research. It outlines the approach taken to model election forecasting and evaluates the efficacy of these models through systematic validation techniques.

Chapter 6, "The Journey," chronicles the development and testing of various models used to forecast Finnish parliamentary election outcomes from 2015 to 2023. The chapter includes details about the data collection and preparation processes and reflects on the effectiveness of these models in predicting election results, shedding light on the complexities and challenges faced.

Chapter 7, "Results," presents the findings of the research and discusses the development and application of predictive artifacts and their evaluation. It includes an in-depth analysis of these models and their performance in forecasting electoral outcomes, highlighting the nuances of data preparation and feature selection.

Finally, Chapter 8, "Key Lessons," synthesizes the insights gained throughout the research. The chapter reflects on the broader significance of using digital data in predicting elections and discusses the implications for data-driven research. This chapter emphasizes the balance between technical innovation and theoretical grounding, advocating the combined use of data science and domain expertise to enhance predictive capabilities.

This structured approach ensures a cohesive narrative that bridges theoretical

insights with empirical evidence, facilitating an understanding of how data-driven methods can transform traditional research paradigms in the context of electoral forecasting.

2 The Data-Driven Research Agenda

This chapter focuses on concepts related to the utilization of online data, including foundational research traditions and key concepts. The contemporary discourse surrounding science and technology encompasses many prominent terms such as artificial intelligence, machine learning, and big data (Haenlein and Kaplan, 2019). Understanding the relevance and scope of these subjects is essential to this thesis, as they have sparked considerable academic discussion on the potential and implications of online data utilization. By exploring these elements, the chapter highlights the significant role these technologies play in shaping modern research and societal narratives.

2.1 Connecting Online and Offline

The interconnectedness of the physical and digital worlds posits that online data can be used to analyze and predict real-world phenomena (Lazer et al., 2009). For example, numerous studies suggest that trends on social networks can often reflect and even forecast public opinion, economic indicators, or social movements (Phillips et al., 2017).

On a high level, several formal theories can be identified which acknowledge that the digital and physical worlds are deeply intertwined, such as Actor-Network Theory (ANT) (Muniesa, 2015), Social Construction of Technology (SCOT) (Pinch and Bijker, 1984), and Media Ecology (Strate, 2004). Although these theories are not directly relevant to this thesis, they testify to the existence of a relationship between the digital and physical worlds, influencing areas such as technology adoption, digital identity, and networked communication.

Digital traces are probably the most intriguing type of online data. These traces are a form of observational data, defined as data resulting from passive observation of a social system without direct intervention (Salganik, 2019). The analysis of digital traces presents numerous theoretical challenges, particularly in the realm of causal inference. The inherent limitations of observational data require careful consideration of possible confounding variables and biases, which often complicate the ability to draw robust conclusions (Rosenbaum, 2021).

However, online data are not limited to digital traces. Online data refers to any information that is accessible via the internet, encompassing a wide range of digital

content including user-generated content, transactional data, multimedia files, social media interactions, web pages, databases, and other forms of electronically stored information that can be retrieved and used for various purposes.

Research that uses publicly available online data can be described as re-purposing. Re-purposing involves using data collected by governments and companies for purposes other than those for which it was originally generated (Salganik, 2019). This approach allows for the analysis of large-scale real-world data, but introduces additional considerations regarding data quality and relevance. It also raises ethical implications for the use of data beyond its original context.

A typical example of re-purposed data is Google Flu Trends, which aims to predict flu outbreaks by analyzing Google search queries (Salganik, 2019). The original paper demonstrated how search data on the internet could be used to estimate flu activity in near real time (Ginsberg et al., 2009). This innovative approach allowed researchers to harness data that were originally generated for a completely different purpose, individual's search for health information, to provide valuable public health insights.

After initial hype, the researchers discovered that Google Flu Trends performance was not significantly better than a simple regression model based on recent flu data (Goel et al., 2010). During the 2009 Swine Flu outbreak, it overestimated flu cases due to changes in search behaviors driven by widespread fear. Over time, algorithmic changes by Google, such as suggesting related search terms, led to an overestimation of the prevalence of flu (Salganik, 2019). Despite its challenges, the Google Flu Trends example underscores the potential of re-purposed observational data for meaningful insights across various domains.

The study by Kosinski et al. (2013) demonstrated that Facebook Likes can predict a range of sensitive personal attributes, such as sexual orientation, ethnicity, political views, personality traits, and more (Kosinski et al., 2013). By analyzing Likes data along with demographic profiles and psychometric tests from a large sample of volunteers, the researchers developed a model that accurately predicted individual profiles. This study highlights the potential of online data to reveal personal characteristics and the importance of ethical dimensions.

The connection between the online and offline worlds is well-established through both theoretical frameworks and empirical evidence. Using online data to understand real-world phenomena significantly broadens the scope of IS research. This expansion requires a robust methodological approach that combines data science techniques with theoretical insights from the research domain. Subsequently, researchers must have a strong grounding in relevant domain theories, fostering extensive cross-disciplinary collaboration. This collaboration is essential to integrate technical and theoretical perspectives, enriching the research process and ensuring that the findings are technically sound and contextually meaningful.

2.2 Central Concepts

2.2.1 Research Traditions

There are various research traditions which are used in relation to online data, such as data science, internet research, and data mining. Because there is such a large number of different perspectives, it is sometimes challenging to define the boundaries of these fields clearly.

Data science is often seen as the most widely acknowledged approach (Fuchs, 2017). It is an interdisciplinary field that employs a variety of scientific and computational methods to extract meaningful insights and knowledge from various types of data (Han et al., 2023). Using methods from statistics, computer science, and machine learning, data science aims to discover patterns, trends, and novel insights from data retrieved with techniques such as web crawling, social media monitoring, and open data analysis. Data science is closely related to the concepts of big data, machine learning, and artificial intelligence.

Another popular stream of research is the tradition of internet research. It focuses on the collection and analysis of online data for academic purposes (Markham et al., 2012; Tsatsou, 2016). It often overlaps with fields such as computational social sciences, internet studies, and social network analysis (Hargittai and Sandvig, 2015). Internet research not only uses the internet as a medium for data collection, but also studies the internet itself as a subject. Internet research underscores the importance of deductive reasoning and theoretical frameworks over simple pattern discovery (Tsatsou, 2016). Furthermore, it has a rich tradition in qualitative methods, which allows for a deeper contextual understanding of online phenomena, providing insights that go beyond what purely quantitative approaches can offer.

Data mining, closely related to knowledge discovery from databases (KDD) and data science, refers to the methods used to discover insights from large data sets. It encompasses a wide range of techniques related to the discovery of knowledge from data, big data, artificial intelligence, and machine learning (Han et al., 2023). Primarily quantitative, data mining literature provides extensive information on algorithms and models to retrieve and use online data, but often overlooks domain understanding, focusing primarily on technical aspects (Cao, 2010; Cao et al., 2010).

Although data science and data mining are quantitative and technique-driven, domain knowledge remains crucial. Models like the Cross-Industry Standard Process for Data Mining (CRISP-DM) recognize the importance of business understanding, but this aspect is frequently omitted in the literature (Wirth and Hipp, 2000; Cao, 2010). Domain knowledge plays a fundamental role in the data science process. Human intelligence, which includes intuition, empirical knowledge, and expert evaluation, is essential to interpret data and guide analyses (Cao, 2017). Beyond raw data, domain intelligence incorporates relevant factors and meta-knowledge, significantly influencing problem framing and solution accuracy (Cao, 2017).

2.2.2 Artificial Intelligence

Artificial Intelligence (AI) has long been a field of fascination and debate (Haenlein and Kaplan, 2019). At its core, AI involves creating systems capable of performing tasks that typically require human intelligence. These tasks include problem solving, decision making, understanding natural language, and recognizing patterns. Although AI encompasses a broad range of technologies and applications, it is crucial to delineate its specific relationship to this thesis.

The historical evolution of AI has witnessed several cycles of optimism and skepticism, often referred to as ”AI summers” and ”AI winters” (Haenlein and Kaplan, 2019). Each AI summer has seen advances and increased expectations, while AI winters have been periods of disillusionment and reduced funding. In recent years, the advent of large language models (LLMs) and other advanced AI technologies has once again fueled a substantial amount of buzz surrounding AI. These developments have demonstrated the potential of AI in natural language processing, generative tasks, and more, further solidifying its role as a transformative technology (Kasneci et al., 2023).

An interesting phenomenon related to advances in artificial intelligence is the AI effect. The AI effects suggests that when some features of AI enter the main stream, it is no longer considered very intelligent. Therefore, AI is a very elusive concept (Haenlein and Kaplan, 2019).

2.2.3 Machine Learning

Machine learning (ML) is a subfield of AI that focuses on the development of algorithms that allow computers to learn from and make decisions based on data (Jordan and Mitchell, 2015). There has been a constant buzz around machine learning, particularly due to significant breakthroughs achieved through techniques such as Deep learning. Deep learning, a subset of ML involving neural networks with many layers, has been especially pivotal in advancing fields like computer vision, speech recognition, and natural language processing (LeCun et al., 2015).

Machine learning has progressed dramatically over the past two decades, evolving from a laboratory curiosity to a practical technology in widespread commercial use. Within AI research, machine learning has emerged as the concrete method for developing practical software for computer vision, speech recognition, natural language processing, robot control, and other applications. Many system developers now recognize that for many applications, it can be much easier to train a system by showing it examples of the desired input-output behavior than to program it manually anticipating the desired response for all possible inputs (Jordan and Mitchell, 2015).

In contrast to conventional programming, where explicit instructions are given,

ML models identify patterns and infer rules based on the provided data sets. This capability makes ML particularly powerful for tasks such as predictive analytics, anomaly detection, and personalization (Geron, 2019).

It should be noted that many traditional statistical methods are also considered machine learning algorithms. For example, linear regression, a fundamental statistical analysis technique, is classified as a machine learning algorithm (Geron, 2019). This demonstrates the broad spectrum of methods encompassed within machine learning, ranging from classical statistical approaches to cutting-edge neural networks.

Machine learning systems can be classified according to several criteria (Geron, 2019). The most typical categorization is the level of supervision required by the system. This refers to the extent to which the data being used are labeled or unlabeled and how the learning process is guided.

- **Supervised Learning:** Systems are trained on labeled data, which means that each training example is paired with an output label.
- **Unsupervised Learning:** Systems are fed unlabeled data and must find patterns and relationships within the data set.
- **Semi-supervised Learning:** Uses a combination of labeled and unlabeled data to improve the accuracy of learning.
- **Reinforcement Learning:** Systems learn by interacting with their environment and receiving rewards for performing actions that achieve defined goals.

Another way to classify machine learning systems is based on their learning process. This involves the method by which the models are trained and updated.

- **Online Learning:** Systems learn incrementally by processing data instances sequentially and adapting their models as new data arrive.
- **Batch Learning:** Systems are trained on a batch of data at once, and the models are not updated incrementally.

Finally, machine learning systems can also be categorized by their learning approach, which describes how learning is conceptualized in relation to the data:

- **Instance Learning:** Systems learn by memorizing examples and comparing new data points to known data points.
- **Model Learning:** Systems learn by detecting patterns in training data.

The versatility of machine learning provides powerful tools for a wide range of applications. This thesis leverages machine learning methods to analyze data to uncover patterns and generate insights that drive research conclusions and recommendations. By incorporating different types of machine learning concepts, research can effectively address various aspects of forecasting with online data.

2.2.4 Big Data

Big data is a popular concept among scientists and industry practitioners (Buhl et al., 2013; Kondraganti et al., 2024). Big data is not just a technical term, but also a cultural, technological, and scholarly phenomenon (Boyd and Crawford, 2012). It is a belief that the analysis of large amounts of digital data can uncover previously unknown insights (Boyd and Crawford, 2012; Kitchin, 2014). Some authors go as far as calling it a myth that suggests that wisdom can be derived from pure data in the cloud, bypassing human flaws and limitations (Fuchs, 2017). Although big data is often viewed as a buzzword, the number of scientific publications has gradually increased at least to the year 2021 (Kondraganti et al., 2024).

The technical perspective suggests that big data is differentiated from small data by the volume, velocity, and variety of the data (Boyd and Crawford, 2012; Kitchin, 2014). There exists a great deal of broader conceptual discussion, and the meaning has been expanded over the years to include more characteristics (Salganik, 2019; Kitchin, 2014). However, to avoid going too deep into the definition, the discussion is limited to the original three characteristics.

The **volume** of big data refers to massive amounts of data. Historically, data in the terabyte and petabyte range have been considered big data. However, with advances in storage and processing power, even larger sets of data are now commonly managed. These data are often measured in petabytes, or even exabytes, significantly surpassing the processing capabilities of traditional database systems (Kitchin, 2014). To handle this scale, advanced technologies such as distributed computing are necessary.

The **velocity** of big data refers to the speed with which data are generated, transmitted, and processed (Kitchin, 2014). In today's interconnected world, data is often created in or near real time. This rapid data generation requires advanced systems capable of real-time or near-real-time processing to quickly capture, store, and analyze the information.

The **variety** of big data encompasses the diversity in structured and unstructured formats, originating from multiple sources (Kitchin, 2014).

The data collected from large-scale data systems offers an intriguing representation of larger phenomena, encapsulating valuable insights at specific moments in time, regardless of whether it fits into the academic definition of big data. It is interesting to note that, given the rapid pace of these source systems, the collected data

serve as a unique snapshot of the state of the system at the time of collection. This snapshot may no longer exist in its original form, making it impossible to replicate the data collection.

When contemplating of all this, it is clear that while the physical volume of data used in this research is not immense, its conceptual scope and the breadth of information it encapsulates align it with the larger discourse of big data research. Therefore, the perceived connection between big data research and the presented study arises because the data used is retrieved from large big data systems, and because it follows premises typically presented in big data research. This applies specifically to the social media data that has been accumulated over time from Facebook and Twitter.

2.3 Methodological Approaches in Data Science

Data science is widely applied in both academic and industrial projects, leading to the development of several established frameworks that guide these projects. Data science initiatives often face organizational, sociotechnical, and technical challenges. Among these frameworks, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is the most widely recognized methodology (Schröer et al., 2021; Martínez-Plumed et al., 2019). In addition, there are numerous other frameworks that cater to different project needs (Martinez et al., 2021). However, few methodologies fully address all dimensions of a project, including team dynamics, data management, and information management (Martinez et al., 2021).

In this section, some popular approaches to managing data science initiatives are reviewed. These methodologies, while primarily designed for industry applications, provide useful insights that can be adapted to academic research. These frameworks also provide a good overview of what kind of activities are included in data science projects, such as the study presented here.

2.3.1 Cross-industry Standard Process for Data Mining

The cross-industry standard process for data mining (CRISP-DM) is a widely used methodological framework for data science projects (Schröer et al., 2021; Martínez-Plumed et al., 2019). It is a hierarchical process model consisting of sets of tasks described at four levels of abstraction. The four layers are phases, generic tasks, specialized tasks, and process instances (Wirth and Hipp, 2000).

At the highest level of abstraction, the model comprises six phases: Business understanding, data understanding, data preparation, modeling, evaluation and deployment (Wirth and Hipp, 2000). Within these phases, the framework specifies generic tasks that outline essential activities at each step, along with specialized tasks and process instances to accommodate different project needs. The outline of the process, as well as the iterative nature, is presented in Figure 2.

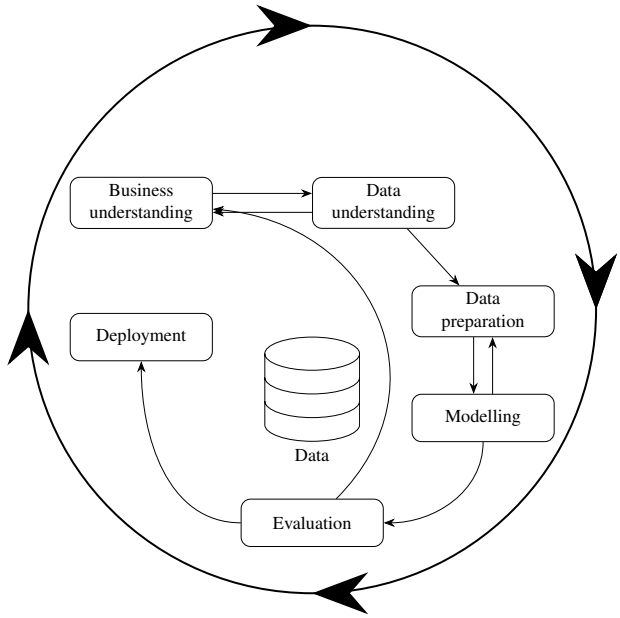

Figure 2. CRISP-DM Diagram (Wirth and Hipp, 2000)

The CRISP-DM process begins with the business understanding phase, where the objectives and requirements of the project are identified from a business perspective and transformed into a definition of data mining problems (Wirth and Hipp, 2000). In this phase, business goals and success criteria are determined, followed by an assessment of resources, requirements, and possible risks. Once the business context is clear, specific data mining goals are established, and a project plan is developed, detailing tools, techniques, and cost-benefit analysis.

The next phase, data understanding, involves collecting initial data and describing its structure and format. Efforts in this phase focus on exploring the data to identify patterns and verifying its quality. Documentation of each task is crucial, resulting in reports that outline the findings and data quality issues, providing a comprehensive overview of the data's current state.

In the data preparation phase, the focus is on selecting, cleaning, and formatting the data to create the final data set. General tasks include selecting data, performing data cleaning, constructing new attributes or records, integrating data from different sources, and formatting the data for subsequent modeling tasks. Each step is documented to ensure clarity and consistency in the preparation process.

During the modeling phase, appropriate modeling techniques are chosen, and test designs are created. Models are built and assessed based on predetermined criteria, with adjustments to parameters made as necessary. Documentation of choices made and assessments conducted ensures transparency in modeling activities, ultimately

leading to a data set ready for evaluation.

The evaluation phase involves assessing the model's alignment with business objectives, taking into account both data mining results and business success criteria. This phase includes a review of the entire process to identify any overlooked factors. The outcome of the evaluation informs decisions on how to proceed, with detailed reports that capture the insights gained from the analysis.

Finally, the deployment phase involves planning for the model's integration into business processes. This includes devising strategies to monitor its performance and maintain the model throughout its lifecycle. A final report is compiled to summarize the project results, providing stakeholders with a clear overview of the project's findings and experiences.

Table 1 outlines the phases and their associated activities. The table links each phase of the process, from business understanding to deployment, with specific tasks and outputs. This overview clarifies the sequence and content of each phase.

The third level, the specialized task level, is the place to describe how actions in the generic tasks should be carried out in situations specific to the project. The fourth level, the process instance, is a record of the actions, decisions, and results of an actual data task engagement. A process instance is organized according to the tasks defined at the higher levels, but represents what actually happened in a particular engagement, rather than what happens in general.

The CRISP-DM framework is suggested to be valuable for planning, documentation, and communication within data science projects. It is deemed helpful for creating specialized processes from generic checklists, which may be particularly advantageous for large projects with multiple stakeholders. Some research indicates that its benefits are more significant in repeatable processes and that defining key performance indicators could improve the control and evaluation of project progress (Wirth and Hipp, 2000). However, more recent studies question its continued relevance, suggesting that while CRISP-DM remains effective for certain types of project, it may not be as suitable for exploratory data science projects, where a more flexible model might be needed (Schröer et al., 2021). Additionally, many studies do not foresee a deployment phase, which poses challenges to the comprehensive applicability of the CRISP-DM framework (Martínez-Plumed et al., 2019).

As the review presents, CRISP-DM is goal-oriented, presupposing a clearly identifiable objective from the outset. However, some authors have observed that data science is often more exploratory in nature (Martínez-Plumed et al., 2019). Consequently, while CRISP-DM provides a strict structure, it is common for data science projects to loosely follow it. Frequently, especially in academic research, the entire deployment stage is missing and generic tasks are ignored (Schröer et al., 2021).

Another common deviation from the framework is to have data preparation as the final step when the prepared data itself constitute the output of the data mining activities (Martínez-Plumed et al., 2019). In addition, data scientists often revisit

Table 1. CRISP-DM Stages and Activities (Wirth and Hipp, 2000)

Stage	Description
Business Understanding	**Determine Business Objectives**: Identify and document business goals, success criteria, and gather background information. **Assess Situation**: Conduct an inventory of resources, understand requirements, constraints, and assess potential risks and contingencies. **Determine Data Mining Goals**: Define data mining objectives and success criteria aligned with business goals. **Produce Project Plan**: Develop a comprehensive plan that includes initial assessment of available tools and techniques, terminology, costs, and benefits.
Data Understanding	**Collect Initial Data**: Gather initial data and document it in an initial data collection report. **Describe Data**: Create a detailed data description report. **Explore Data**: Conduct an exploratory data analysis and produce a data exploration report. **Verify Data Quality**: Assess and document data quality issues in a data quality report.
Data Preparation	**Select Data**: Specify criteria for data selection and exclusion, providing rationale. **Clean Data**: Perform data cleaning and document the process in a data cleaning report. **Construct Data**: Create new attributes or records and document derived attributes, as well as generated records. **Integrate Data**: Merge data from different sources. **Format Data**: Reformat data for modeling, creating a data set and data set description.
Modeling	**Select Modeling Techniques**: Choose modeling techniques and document choices in a modeling technique document. **Generate Test Design**: Develop and document test design, including parameter settings. **Build Model**: Build models and document model descriptions. **Assess Model**: Evaluate models and revise parameter settings based on assessments.
Evaluation	**Evaluate Results**: Assess data mining results against business success criteria and detail findings in an assessment report. **Review Process**: Conduct a review of the modeling process and document findings in a review report. **Determine Next Steps**: Propose actions based on evaluation outcomes and list possible actions.
Deployment	**Plan Deployment**: Develop a detailed deployment plan. **Plan Monitoring and Maintenance**: Create plans for ongoing model monitoring and maintenance, and document these in a monitoring and maintenance plan. **Produce Final Report**: Compile a final report and prepare a final presentation. **Review Project**: Conduct a project review and document experiences in an experience documentation.

previous exploratory activities as new insights emerge during the process. Thus, CRISP-DM emphasizes the process, but its application in data science often reflects the iterative and exploratory nature of the work.

Despite the identified drawbacks, CRISP-DM remains the most recognized methodological guideline in data science today (Martínez-Plumed et al., 2019; Schröer et al., 2021). The wide adaptability suggests that even when applied loosely, it benefits data scientists by providing structure and terminology for various activities, facilitating a better understanding of the current project.

2.3.2 Predictive Analytics

Predictive analytics can be seen as a subfield within data science that focuses on using data-driven techniques to forecast future outcomes. Discussions specifically addressing methodologies in predictive analytics remain limited, with an emphasis on practical applications rather than theoretical frameworks.

A typical workflow for predictive analytics includes several key stages (Shmueli and Koppius, 2011). These stages align rather closely with other established methodologies such as CRISP-DM. The approach begins with clearly defining the purpose of the study and articulating the problem statement (Shmueli and Koppius, 2011; Kuhn and Johnson, 2013). This is followed by designing the study and collecting data, preparing the data by cleaning and transformation, and performing exploratory data analysis (EDA) to explore patterns and relationships (Shmueli and Koppius, 2011; Kuhn and Johnson, 2013).

The next steps involve choosing variables and their form, selecting potential modeling methods, and evaluating and validating models to ensure robustness. Finally, the model is used and findings are reported, effectively communicating insights into actionable decisions.

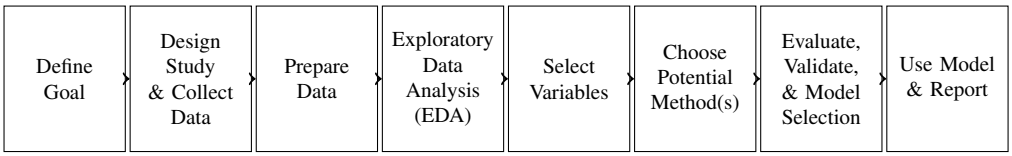

Figure 3. Predictive Analytics Process Stages (Shmueli and Koppius, 2011)

These stages share similarities with CRISP-DM phases, such as business understanding, data preparation, modeling, and deployment, focusing on a structured and iterative approach (Shmueli and Koppius, 2011; Kuhn and Johnson, 2013).

In summary, while predictive analytics aligns closely with broader data science methodologies, its emphasis often lies in practical execution and application, ensuring that model development is both rigorous and relevant to strategic objectives.

2.3.3 Other Industry Frameworks

Other known methodologies used in data science include Knowledge Discovery in Databases (KDD) and the Sample, Explore, Modify, Model, and Assess (SEMMA) methodology, which is developed by the SAS Institute (Martínez-Plumed et al., 2019).

The KDD model involves nine iterative and interactive steps that focus on uncovering hidden knowledge from the data, while emphasizing the need for domain-specific understanding to foster the extraction of patterns that are relevant and actionable (Fayyad et al., 1996). Its iterative nature allows for constant refinement and adaptation, and it serves as the foundational backbone from which CRISP-DM evolved, integrating systematic methodologies from KDD's extensive process.

The SEMMA methodology comprises five distinct stages: sample, explore, modify, model, and assess, and is tightly associated with SAS Enterprise Miner (Azevedo and Santos, 2008). This framework structures the data mining process through these sequential steps, offering a guided analytical process that is primarily used within the commercial software environment. It provides a more linear approach compared to KDD, allowing users to methodically approach data analysis with clear, defined stages.

Both KDD and SEMMA offer alternative structured frameworks for performing data mining tasks, each with their own strengths and areas of emphasis, catering to different project needs and software environments. Furthermore, there are other methodological approaches to data science and data mining that, although less well known, could provide niche solutions or innovative techniques as identified in (Martínez-Plumed et al., 2019).

2.3.4 Emerging Methodologies

As data science continues to be widely adopted, there is an ongoing effort to refine methodologies to better align with real-world applications. One such methodology is Data Science Trajectories (DST).

DST provides an alternative methodological framework that addresses some limitations of traditional approaches such as CRISP-DM. According to its developers, DST offers guidelines that are more suited to the complexities encountered in real-world data science projects (Martínez-Plumed et al., 2019). Unlike CRISP-DM, DST is not structured around a linear process. It places greater emphasis on exploratory activities that are central to effective project execution.

These exploratory activities include goal exploration, which focuses on identifying business objectives that can be achieved through data-driven approaches, data source exploration, which involves discovering novel and valuable data sources, and data value exploration, which seeks to determine the potential value that can be ex-

tracted from the data. Additionally, DST encompasses result exploration, which connects data science insights back to the initial business objectives, narrative exploration, which involves crafting insightful stories from the data through visual or textual means, and product exploration, which is about converting the discovered data value into a service or application that provides distinct benefits to users and customers.

Although there are connections between DST's exploratory activities and the phases of CRISP-DM, the approach of DST is generally more open-ended (Martínez-Plumed et al., 2019). This open-ended nature reflects the iterative and dynamic process of data science, allowing practitioners to revisit and revise earlier steps as new insights are gained and new challenges arise.

Despite the potential, DST is not widely acknowledged, and the real benefits remain unclear.

3 Utilization of Online Data

3.1 Characteristics of Online Data

Online data encompasses various characteristics that influence its utility and limitations in research. This section lists common characteristics, acknowledging that the list is not exhaustive. The characteristics include authenticity, structure, origin, purpose, data quality, representativeness, ownership and accessibility, and dimensions of time and location. For a concise overview, see Table 2.

Authenticity: Primary data is the original, unprocessed information collected directly from the source. Derived data is something that is obtained from processing the original information (Kitchin, 2014, -36). Original data can be used for secondary purposes, such as research, although it was not originally generated for that purpose.

Structure: Data structure defines which analysis methods are applicable. Data can be structured, semi-structured, or unstructured (Kitchin, 2014). Structured data, such as databases and spreadsheets, is highly organized and suitable for quantitative analysis. Semi-structured data, including JavaScript Object Notation (JSON) and extensible markup language (XML) files, contains some organizational tags that facilitate analysis. Unstructured data, such as text, images, or videos, lack inherent organization and must be structured before it can be used for statistical analysis (Han et al., 2023). More than 80% of the worlds data is stored in unstructured formats (Han et al., 2023). There are various tools that can be used to convert unstructured data into suitable formats for statistical analysis, such as LLM. However, some information is lost during the conversion process.

Origin: The origin is often ambiguous, but at a high level it can be defined as user-generated (Krumm et al., 2008), publisher-generated (Chae et al., 2024), hybrid (Chae et al., 2024), or system-generated (Kitchin, 2014). User-generated content is produced by end users who interact with systems without direct commercial interest, such as social media records. Materials produced by professional institutions such as publishers, statistical institutions, and other similar entities are defined here published-generated. System-generated data, also known as exhaust data, are by-products of various processes and interactions that are collected incidentally (Kitchin, 2014). It consists of items such as server logs, network data, sensor recordings, and many other types which are recorded actively but rarely analyzed. Hybrid data are a combination of multiple origins.

Purpose refers to the intention or rationale behind the collection and use of the data (Kitchin, 2014, -36). For primary data, the purpose is typically explicit, as these data are collected to address specific research questions or objectives; for example, survey data might be gathered to understand consumer preferences with the clear intention of informing marketing strategies. Secondary data, on the other hand, is initially collected for a different primary purpose but is later repurposed for new analyses or investigations, such as using government census data to determine the size of the market (Salganik, 2019). Exhaust or by-product data often has an indirect purpose, as it is generated inadvertently during the primary functioning of a system or process and may be analyzed later for insights, like website analytics collected as users interact with a page. Understanding the purpose behind data collection helps determine its relevance, applicability, and potential biases in its analysis and use.

Quality: Most freely available online datasets include some junk and spam (Salganik, 2019). With user-generated content, these issues can be expected, although their prevalence can vary. The level of moderation can help reduce junk content, with fully moderated data often having less. Regardless of detectable data quality issues, it is important to account for biases and errors introduced by system designers (Boyd and Crawford, 2012).

Representativeness: To generalize findings, understanding the demographic context of the data is essential. Big data is essentially non-representative (Salganik, 2019). Publicly available data often lack demographic information due to privacy concerns, even when it is generated by government records. Although this limitation allows for in-sample comparisons, it should always be considered seriously when making any inferences.

Ownership: Another important characteristic is the access control related to the data. Data can be private, partially public, public, or not accessible at all. Some data is publicly available, but might require methods such as web-scraping to access. Even when data are publicly available, ethical concerns and privacy issues might restrict its use. The most interesting sources of big data are private, and accessing them typically requires a contract with the data owner (Salganik, 2019). Furthermore, even if you have access to the data, there may be restrictions on how you can use it (Hargittai and Sandvig, 2015).

Dimensions of time and location: Data can be characterized from both temporal and spatial dimensions (Kitchin, 2014). Temporality refers to the fact that data are often time-bound, which means that they are collected, modified, and potentially rendered obsolete over time. Online data may change or cease to exist, as digital content is inherently unstable (Hargittai and Sandvig, 2015). Online data drifts in three main ways: shifts in who is generating the data, changes in how it is generated, and modifications in system design itself. Drifting implies that the insights derived from the data are transient (Salganik, 2019). Spatiality, on the other hand, refers to the geographical and contextual location of the data. Data might be generated

in different physical locations or within various contexts, which can influence its content and relevance (Kitchin, 2014).

Characteristic	Key Points and Challenges
Authenticity	Primary data is the unprocessed original data, while derived is something that has already been processed.
Structure	Defines which analytical methods are applicable. Types include structured, semi-structured, and unstructured data.
Origin	Data can originate from end-users, publishers, systems, or from any combination of those.
Purpose	Refers to the intention or goal behind collecting or using the data. The purpose is either primary or secondary.
Data Quality	Primary data often has many issues such as missing values or junk/spam, which has to be dealt with before the data can be used for analysis.
Representativeness	Big data is typically non-representative. Researchers have to pay careful attention when making inferences.
Ownership and Accessibility	Data accessibility ranges from private to public, with ethical and legal limitations on its use.
Time & Location	Refers to temporal aspects (data is time-bound and may drift) and spatial aspects (location affects relevance).

Table 2. Characteristics of Online Data

All data, of course, are not equal for data mining and research, which is why understanding the characteristics is vital. Some estimates suggest 80% of the internet traffic consists of video (Patel et al., 2022). Multimedia formats have a very high information density, as they can contain several kinds of data, such as text, audio, visual, and meta-data. Historically, data mining from multimedia formats has been a challenging undertaking (Vijayakumar and Nedunchezhian, 2012). However, recent advances in technology have made it more manageable (Wu et al., 2023).

3.2 Online Data Sources

3.2.1 Social Media

Social networks have arguably been the most relevant data source for this research. Social media, which consist of various social networking services, have become an integral part of society and a central component of contemporary digital life (Van Dijck and Poell, 2013). Social media networks have billions of users worldwide, with the average adult spending more than two hours each day using social networking services (Statista, 2022). These platforms enable users to create and share content, facilitating continuous engagement and interaction among users. In 2024, some notable social networking services included Facebook, YouTube, Instagram, TikTok, Douyin, Weibo, Snapchat, Pinterest, X (formerly Twitter)[1], and Reddit. The widespread usage of these platforms generates enormous amounts of data from everyday interactions. This vast pool of data has been suggested to provide valuable insights for researchers in analyzing trends, understanding public opinion, and studying social phenomena (Golder and Macy, 2014; Savage and Burrows, 2007).

The extensive user base of social media makes it a rich resource for a wide range of research applications. Using data retrieved from social media has gained substantial foothold in academic studies (Aichner et al., 2021). However, the exact definition of social media, as well as the nature of services, has fluctuated over the years (Aichner et al., 2021). This variability makes it difficult to compare studies conducted at different times.

When analyzing published empirical results, researchers should pay attention to when the empirical results were released, how social networks are defined in the context, and the evolution of the platform itself during the years after publication. For example, the role of X in politics has drastically changed due to changes on the platform; currently, it is more difficult to view and participate in discussions on X without registering. Consequently, results obtained in the early 2010s may not be reproducible today. By taking these factors into account, researchers can improve the accuracy of their comparisons and avoid common theoretical or methodological pitfalls (Aichner et al., 2021).

Understanding the origins and nature of social media data is essential for deriving accurate insights. Researchers need to frame their studies within the context of the specific social media landscape they are investigating. Temporal changes and evolving user behaviors can affect the characteristics of the data and their comparability (Salganik, 2019). Researchers should acknowledge the dynamic nature of these platforms and adapt their methodologies accordingly to maintain the validity

[1] A distinction is made between X and Twitter depending on the context. When a study involving Twitter data is referenced, Twitter is specifically discussed. Conversely, if the conversation pertains to a more general discussion of the platform in its current form, it is referred to as X. This differentiation ensures that clarity and relevance are maintained according to the specific focus of each discussion.

of their findings.

Data from social media can be structured, semi-structured, or unstructured. It is often stored in graph databases due to their ability to efficiently handle complex relationships and interconnected data (Han et al., 2023). Structured data, such as user profiles and pages, are highly organized and suitable for quantitative analysis. Semi-structured data includes elements like timestamps and metadata found in posts and messages; these elements contain organizational tags that aid in analysis. Unstructured data includes posts, comments, images, and videos.

There is some variation in the origin of different types of social media data. The most interesting data are usually the generated by the end users, including posts, likes, and comments. Social networks also host large amounts of content generated by professional publishers, such as news articles or promotional content. Some of the data also originate directly from the platform itself, such as metrics like engagement rates, click-through rates, and session durations.

Using social media data for research is always secondary usage. The primary purpose of the data generated is to serve the platform. The available data is generated either through direct user actions, like posting and sharing, or as by-products of user interactions with the platform.

Social media data are often very dirty, as a significant percentage of the data is generated by bots (Salganik, 2019). Poor data quality is a typical issue related to user-generated content. Although moderation can reduce the extent, it cannot completely eliminate it.

It should be also acknowledged that the available data rarely represents the target population. By default, researchers should assume that the data are representative only of the specific subset of users analyzed. It is essential to identify specific factors or evidence that suggest a broader applicability before considering the data as representative of a larger population. The cause of poor representativeness is self-selection bias, leading to non-representative sampling and, consequently, limits the generalizability of the findings (Schoen et al., 2013). Demographic information is often incomplete or anonymized, making it difficult to take corrective measures (Schoen et al., 2013).

The ownership and access to social media data is a controversial topic (Bruns, 2019). Large parts of social media data are public. Therefore, it can often be accessed through web-scraping, but ethical and privacy concerns may still restrict its use. Some parts of the content are usually semi-public, as while anyone can access it, viewing requires registering. Registering requires accepting the terms of service, which typically prohibit any secondary use of the data. The ownership of original content is also a somewhat unclear topic (Bruns, 2019). Meta-data, such as user demographics, are typically out of reach for researchers. Even if the data were accessible, there is no way to verify whether it is accurate.

Social media data are inherently time sensitive, as they are continuously gener-

ated and modified, meaning that data collected at one point can become outdated quickly. This temporality affects the relevancy of the data and requires constant updates for ongoing analysis (Salganik, 2019). Furthermore, the spatial dimension should also be considered given the global nature of social media platforms. Evidence suggests that social media attention tends to be biased towards more populous areas during disasters, potentially skewing the relevance and reach of data in crucial situations (Fan et al., 2020). Furthermore, the global nature of social media services means that followers can be from anywhere, complicating the interpretation of social media metrics. For example, in the context of political campaigns, some candidates may have a substantial portion of their followers from outside their electoral district or even from different countries, influencing the perceived support and reach despite these followers not being eligible voters. Thus, understanding and accounting for these temporal and spatial nuances is essential to perform accurate and relevant social media analysis.

Observational data from social media data has been applied in various research areas. Researchers have attempted to use it to forecast movie box office sales by analyzing online patterns and discussions (Asur and Huberman, 2010). In mental health research, linguistic analysis of Facebook posts has revealed indications of depression (Eichstaedt et al., 2018). Political sentiment analysis has used tweets to predict election outcomes (Tumasjan et al., 2010). Environmental issues have been studied by measuring public engagement with climate change topics using Twitter data (Kirilenko and Stepchenkova, 2014). In financial markets, the results suggest that there was a correlation between the mood expressed on Twitter and the movements of the stock market (Bollen et al., 2021). Global happiness trends have been examined using Twitter data, unveiling patterns of happiness and information spread (Dodds et al., 2011). In transportation research, tweet analysis has detected traffic incidents in real time (Gu et al., 2016). Social media activity has identified signs of depression in mental health monitoring (De Choudhury et al., 2013). Lastly, consumer buzz and social media traffic have been used to predict a firm's market value (Luo and Zhang, 2013).

Despite many promising findings, social media data comes with various caveats. Most of the research is mainly focused on Twitter for data collection (Chauhan et al., 2021; Brito et al., 2021; Gayo-Avello, 2013; Phillips et al., 2017; Rousidis et al., 2020). However, evidence suggests that results from one platform may not be easily generalized to others (Alhabash and Ma, 2017; Brito et al., 2021; Rousidis et al., 2020). Additionally, many studies tend to be short-term and focus on single events, thereby limiting a broader understanding of trends and behaviors over time (Schoen et al., 2013). Furthermore, sampling social media data poses a major challenge for standard statistical methods; a random sample from social media users often does not accurately represent the general population (Gayo-Avello, 2013; Schoen et al., 2013). Long-term studies with repeated data collection are crucial for providing

deeper insights and more reliable representations of social media phenomena (Brito et al., 2021).

3.2.2 Open Data

Publicly available data repositories are valuable resources for researchers. These platforms, maintained by government bodies, research institutions, other organizations, and individuals, provide access to a wide array of data sets intended for public use. Researchers can utilize these repositories to find data relevant to their fields, often without the complexities of direct data extraction from other sources.

Open data, in short, refers to the right to copy, edit, and redistribute data in the original or modified form. Although this definition is sufficient for most discussions related to the topic, the exact definition is nuanced. Open can mean different things for different agencies in the context of intellectual property rights (Kitchin, 2014). Therefore, few organizations have sought to create a precise definition (Kitchin, 2014). The up-to-date definition set forth by the The Open Knowledge Foundation in 2015 seems to be the most relevant, and it can be summarized as:

> ”Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.” Pollock (2015)

Governments have a long history of releasing statistical information (Duncan et al., 1993), and concerns about the confidentiality of personal records have also been present for a long time (Duncan et al., 1993). The International Open Data Charter has gained traction, being adopted by 74 national and local governments and endorsed by 55 organizations and non-state actors (Badiee et al., 2021).

In Finland, government statistics are closely aligned with the concept of open data. Data are released under the Creative Commons 4.0 license (Finland, 2024), and the Act on Openness of Government Activities stipulates that all government actions are public unless otherwise specified (of Justice, 1999). As a result, the data are open by default, which means that ensuring that the public has access unless specific exceptions are applied. Open by default means that policy makers can presume that all data will be published, unless there is a justifiable reason to not (Badiee et al., 2021).

Although the data are technically public, several government bodies still charge for access. As a result, most of the data does not meet the open data definition. Various data products from government bodies come with different licenses, which may dictate how the data can be used and whether they can be redistributed.

Government-released administrative records and statistics are an important source of data for researchers (Salganik, 2019). The primary purpose of the data is to support effective governance, and the collected data typically contains sensitive informa-

tion about citizens. Governments often take several steps to anonymize and process data before publishing them as open data (Salganik, 2019). Therefore, most of the publicly available data released by governments can be described as secondary, although openness and transparency are valid traits of good governance, and some data may be purposefully collected to demonstrate this commitment. However, primary data are rarely released as is due to potential violations of the rights of citizens to privacy.

Data from public repositories are typically structured, mostly in the form of tables and databases, designed for straightforward quantitative analysis. Government data archives often include demographic statistics, economic indicators, public health records, and more. Open data initiatives may also provide access to environmental data, transportation records, and educational statistics, as there are no restrictions in the focus of open data (Kitchin, 2014). These organized data sets are valuable for longitudinal studies and trend analysis.

Despite the structured nature of public repository data, it should not be blindly trusted. Data may be outdated if repositories are not regularly updated, and completeness can vary between different data sets. Additionally, while public data are generally considered reliable, they may still contain inaccuracies or biases introduced during data collection. Especially data from surveys are subject to standard issues related to collecting data. These factors must be taken into account to ensure a rigorous analysis.

3.2.3 Domain-specific Data Sources

Each research domain encompasses a multitude of private organizations and individuals who collect and store vast amounts of data every day. These data, often stored on private devices and databases, may sometimes be publicly available online or accessible upon request. Domain-specific data sources include market data, real estate data, code repositories, online reviews, blogs, news, research publications, and job market data. These data sets are generated and managed by their respective publishers, who maintain control over the data (Kitchin, 2014).

The characteristics of private data vary significantly. One characteristic that can be considered common to all data sources of this type is ownership. Although organizations may make the data publicly viewable to serve interests such as transparency or market facilitation, the owners of the data retain the rights to it. Therefore, private owners can modify the data as they see fit, provided that they follow legal guidelines. Data access limitations pose a major challenge for research, especially from the perspective of reliability, as studies may be difficult to replicate (Boyd and Crawford, 2012). Furthermore, biases present in the data collection process, whether intentional or accidental, are found in the online data as they are found in any other data source (Boyd and Crawford, 2012).

As a case example, we can investigate the voting advice application data used in this research. In Finland, voting advice applications are offered by state-owned media companies and private firms. Voting advice applications are used by citizens to compare political candidates and by candidates to release information about their opinions (Isotalo, 2021). Voting advice applications typically aim to reach as wide audience as possible. The candidate data are therefore publicly available. However, how citizens engage with the application and what they respond to the questionnaires is kept private. The candidate data are structured, and can be retrieved using approaches like web-scraping.

The production of the data is a joint effort by the media companies and the candidates. Candidates select their answers from a given set of questions. The data quality is typically quite good because the questions are moderated, and it is unlikely to contain responses by bots. Furthermore, moderators have direct access to the data and can modify it if necessary. In general, voting advice application data are quite representative of the candidate population, although there is a form of self-selection bias that has to be acknowledged. Some candidates decide to not spend their time answering the surveys.

3.3 Collecting Online Data

3.3.1 Selecting Data Collection Method

There are several methods to collect data online. The most straightforward approach involves accessing structured and curated data intended for research purposes, which can be easily downloaded, such as the government-provided open data repositories mentioned earlier. Alternatively, the internet can facilitate the distribution of surveys, inviting users to participate by completing and submitting responses. Beyond these conventional techniques, there is a vast array of secondary, observational, data available on the internet.

Collecting observational online data is accomplished through two primary methods: using application programming interfaces (APIs) for machine-readable data, or employing web-scraping techniques to extract information from website content aimed at human consumption (Broucke and Baesens, 2018). In traditional web browsing, a browser interacts with a web server by sending a HyperText Transfer Protocol (HTTP) request and receiving a Hypertext Markup Language (HTML) response, which the browser then renders for user interaction (Broucke and Baesens, 2018). This process is designed for human consumption and interaction.

Web-scraping refers to the process of using a machine to gather, store and analyze websites meant for human use (Broucke and Baesens, 2018). In this method, the machine visits websites, copies the content displayed to users, and then converts this content into a structured format. Web-scraping, however, is often more challenging

than using APIs. Social media services, for example, frequently make their format obscure and continually change it to prevent automated scraping (Bruns, 2019). In addition, servers can easily block access to data using various authentication challenges and other security measures, making it difficult to retrieve information reliably. One common method is using Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA). CAPTCHAs are designed to differentiate human users from bots (Moradi and Keyvanpour, 2015).

APIs, on the other hand, provide structured data formats such as JSON or XML, which are easier for machines to process directly (Broucke and Baesens, 2018). Many websites, including large platforms like X, Facebook, LinkedIn, and Google, provide APIs to allow external programs to access their data repositories in a structured way. These APIs are designed to be used by computer programs, facilitating functionalities such as searching and posting messages, fetching lists of friends and their likes, and viewing connections. For example, X's API can be used to easily obtain a list of recently published content, eliminating the need to build a web scraper for this purpose. Data obtained through an API can include additional metadata, allowing a more comprehensive credibility analysis (Dongo et al., 2021).

Resorting to web-scraping is often necessary (Broucke and Baesens, 2018). While there are many ethical and technical challenges related to web-scraping, it is sometimes a necessary evil for digital research (Trezza, 2023). The reasons for web-scraping can be, for example, that there are no APIs or that the use of APIs is limited (Broucke and Baesens, 2018). For example, social media services exposed their APIs widely prior to 2018, but it is these days increasingly difficult to retrieve social media data without using web-scraping (Trezza, 2023; Bruns, 2019). If you can view data in your web browser, you can potentially access and retrieve it through a program. However, accessing data may sometimes require substantial technical know-how. Once accessed, the data can be stored, cleaned, and utilized in any required manner.

In summary, while APIs offer a more straightforward and reliable means of accessing data, subject to availability and functionality, web-scraping serves as an alternative method when APIs are insufficient or unavailable (Broucke and Baesens, 2018).

The next two subsections will review each method more thoroughly, discussing their respective advantages, challenges, and best practices.

3.3.2 Application Programming Interfaces

APIs exist in various forms, including Simple Object Access Protocol (SOAP), Representational State Transfer (RESTful), and Graph Query Language (GraphQL) (Sayago Heredia et al., 2020). SOAP, introduced in the late 1990s, uses XML to encode messages. While technically possible, SOAP is rarely used on browser clients due to its complexity. RESTful APIs are more commonly employed due to their simplicity

and flexibility, utilizing standard HTTP methods. GraphQL, developed internally by Facebook, offers an alternative by allowing clients to request precisely the data they need, thus minimizing the data fetched and reducing the number of server requests.

Authentication and authorization are crucial concepts in web services to ensure secure access to sensitive information (Trnka et al., 2018). Authentication verifies the identity of a user or application, while authorization determines the permissions and access levels granted to authenticated users. These processes are essential for enforcing terms of service and ensuring compliance with agreements (Mustapha et al., 2024).

Different forms of APIs handle authentication and authorization in various ways. SOAP APIs often rely on WS-Security, which includes features for message integrity, confidentiality, and single-sign-on. RESTful APIs commonly use token-based authentication methods, including JSON Web Tokens (JWT). GraphQL APIs also employ similar token mechanisms for secure access. The OAuth standard is widely used across these APIs, providing a robust framework for token-based authentication and delegated access. OAuth allows users to grant third-party applications limited access to their resources without sharing credentials, enhancing both security and user convenience (Hardt, 2012).

When information is accessible anonymously (without authorization), it can be considered public. Public information transmitted via APIs does not require authentication. Many websites fetch and display public information through public APIs. Therefore, APIs can be used to discover and retrieve data similarly to web-scraping (Mustapha et al., 2024). A common misconception surrounding APIs is that they always require authentication. However, if the content is publicly visible without logging in, it can often be accessed through an API (Mustapha et al., 2024). Most APIs use the HTTP protocol, just like websites, and understanding that an HTTP API essentially operates as a web service can be helpful for grasping its functionality.

My personal experience is that API scraping can typically be performed on services that use a single-page application (SPA) architecture. SPAs rely on APIs to dynamically update content without reloading the entire page. If the APIs are publicly exposed, they can be used to retrieve data without the client interface. In contrast, server-side rendering (SSR) generates the entire HTML content on the server side and sends it to the client. SSR does not rely on client-side APIs for data fetching in the same way SPAs do, potentially making direct API access less available in SSR contexts. In summary, SPAs offer more opportunities for API scraping due to their architectural reliance on client-side APIs, whereas SSR environments may limit such access.

APIs are commonly used in academic research for tasks such as retrieving online reviews (Aldabbas et al., 2020). Platforms like Google offer several public APIs that allow retrieving data without authorization. These APIs can be used to fetch information such as user reviews, updates, and ratings. Using APIs, researchers can

efficiently collect large volumes of public data programmatically, eliminating the need for traditional web-scraping methods that involve parsing HTML content.

3.3.3 Web-scraping

Web-scraping involves the use of scripts or software to automatically gather data from the internet, particularly from websites that do not provide direct download links or APIs (Broucke and Baesens, 2018). This technique relies on parsing HTML content to extract the desired information. Various tools and libraries, such as BeautifulSoup, Scrapy, and Selenium, facilitate this process.

The typical web-scraping workflow includes identifying the structure of a website, locating the data of interest within the HTML, sending HTTP requests to web pages, and parsing the resulting HTML to filter out irrelevant content (Broucke and Baesens, 2018). By extracting data programmatically, web-scraping can be especially useful for gathering information from websites that do not provide public APIs.

Web-scraping is a versatile tool with a wide range of applications in different fields. One notable example of web-scraping in practice is The Billion Prices Project, which collects price data from hundreds of online retailers to track inflation daily (Cavallo and Rigobon, 2016). By scraping product and price information from various e-commerce sites, the project constructs a large data set that provides real-time insights into price changes and economic trends (Cavallo and Rigobon, 2016). This example demonstrates the practical application and significant potential of web-scraping for economic research and data analysis.

Businesses also use web-scraping for several purposes, such as competitive analysis and market research. Companies can monitor competitor pricing strategies, product availability, and customer reviews to make informed decisions and optimize their own offerings. In the field of digital marketing, web-scraping is used to analyze social media trends and sentiment, enabling brands to measure public perception and effectively tailor their marketing campaigns.

Web-scraping presents several technical challenges that professionals must navigate to effectively gather data. Anti-scraping mechanisms such as CAPTCHA, IP blocking, and dynamic content loading can thwart basic scraping efforts. Handling JavaScript-rendered content, which requires executing JavaScript on the client side, often necessitates the use of headless browsing tools like Selenium or Puppeteer. Beyond accessing the data, significant effort is needed to clean up and transform the scraped information into a structured and usable format. This process includes parsing HTML, dealing with inconsistent data patterns, and filtering out irrelevant content. Addressing these technical hurdles is essential for successful web-scraping projects.

Web scraping raises specific ethical concerns, including compliance with websites' terms of service and data protection laws, such as the European General Data

Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) in the United States (Broucke and Baesens, 2018). Best practices for ethical scraping involve respecting the robots.txt file to honor a website's crawl instructions and implementing rate limiting to prevent overloading servers. In addition, scraping is considered more friendly during off-peak hours to minimize disruption. Web-scraping has also been at the center of high-profile legal battles, such as the LinkedIn vs. HiQ Labs case, which highlights the ongoing legal challenges and ambiguities surrounding the practice (Lawrence and Ehle, 2019).

3.4 Using Online Data in Research

3.4.1 Example Applications

Recent advances in information and communication technology have increased the availability of data of diverse origins, enabling the integration of a wide variety of datasets. Internet, in essence, acts as a vast and widespread repository that encompasses numerous data sources and formats (Wang et al., 2018). Online data plays an increasingly important role in informed decision-making, providing a rich foundation of information and knowledge that researchers and decision makers can access and leverage (Wang et al., 2018).

Different forms of online data have been extensively used in research, as evidenced by numerous relevant research papers. For example, search data has been used to study disease outbreaks, with a relationship found between flu activity and search volumes (Ginsberg et al., 2009). Similar results were found for COVID-19 (Lampos et al., 2021), suggesting that search data could serve as complementary data for public health surveillance. Search data has also been used to predict consumer activities, such as purchasing music or video games (Goel et al., 2010).

Another widely used data source are online reviews. Empirical results indicate a relationship between sales performance and reviews, as they impact consumer purchasing decisions (Floyd et al., 2014). In the hospitality industry, online reviews are used to predict overall satisfaction, with data from platforms such as TripAdvisor and Yelp being analyzed to gauge customer experiences and preferences (Zhao et al., 2019).

Furthermore, web-based code hosting services, such as GitHub, have been used to study software development processes and related activities. Researchers analyze repositories to understand collaboration patterns, code quality, and project outcomes (Kalliamvakou et al., 2014). Similarly, the source code and related repositories of projects such as WebKit and OpenStack have been examined to investigate the dynamics of collaboration within the open-source ecosystem (Teixeira, 2018).

In addition, integration of information from real estate transactions, public school ratings, and mortgage rates has been used to forecast housing prices (Park and Bae,

2015). This combination of different data types allows for a more comprehensive understanding of the factors influencing the real estate market. Similarly, in the stock market, the use of news data has been used to predict stock market movements (Fung et al., 2003), highlighting the wide-ranging applicability of open data in economic forecasting.

3.4.2 Quality Issues

Data quality is a fundamental concern in all areas of research. As noted earlier, many freely available online datasets often exhibit considerable quality issues (Salganik, 2019). The quality of data can vary significantly depending on the source, and it always requires careful scrutiny to ensure the accuracy and validity of research outcomes.

In data science, the effectiveness of a system can be compromised when the training data is insufficient in size or lacks representativeness. Noisy or irrelevant features can adversely affect modeling accuracy, leading to what is commonly known as the ”garbage in, garbage out” phenomenon (Kilkenny and Robinson, 2018). Getting the right balance is essential. A model should not be too simplistic, leading to underfitting, nor too complex, leading to overfitting (Geron, 2019; Hastie et al., 2009).

Data analysts typically invest a substantial amount of time diagnosing data quality issues (Kandel et al., 2011). Data are considered to be of quality if they meet the requirements of their intended use. Numerous factors define data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability (Han et al., 2023). Real-world data often tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines aim to fill in missing values, smooth out noise, identify outliers, and correct inconsistencies.

The underlying issue is that a model is a simplified version of the observations. These simplifications aim to discard superfluous details that are unlikely to generalize. To decide what data to discard and what data to keep, the analyst has to make assumptions. For instance, a linear model assumes that the data is fundamentally linear, meaning that the deviations between individual data points and the straight line are merely noise that can be safely ignored. Making these assumptions helps to focus on the critical aspects of the data that are most likely to be relevant for future predictions, but also introduces the risk of oversimplifying complex relationships (Geron, 2019).

An examination of past studies using data from what was formerly known as Twitter highlights the challenges posed by its unstructured and biased data nature (Gayo-Avello, 2012). An example case is a study by Tumasjan et al. (Tumasjan et al., 2011), where researchers attempted to predict election outcomes based on Twitter data. They initially claimed that the volume of messages that mention political parties could reliably predict election results. However, it was later discovered

that the authors had selectively included data that fit their model while excluding data that did not. For example, they chose not to include mentions of the German Pirate Party, which would have been ranked as the winners if included (Jungherr et al., 2012).

3.5 Integration of Several Data Sources

3.5.1 Objectives and Advantages of Data Integration

The primary objective of data integration is to extract valuable insights that cannot be achieved by analyzing individual data sets (Gligorijević and Pržulj, 2015). The consolidation of diverse datasets is suggested to reduce redundancies and inconsistencies, improve accuracy, and enable the identification of patterns that are not apparent in isolated datasets. Data integration can be achieved by combining multiple data sets into a single data set, or by ensemble modeling, where the outputs of multiple models are combined to enhance analytical insights (Dietterich, 2000).

The available evidence suggests that data integration can increase forecast accuracy, especially when using diverse methods and data (Graefe et al., 2014). For instance, in the field of electricity price forecasting, researchers have used multiple data sources to improve accuracy (Neupane et al., 2017). Similarly, in research on the prevalence of chronic diseases, the combination of multiple data sources has been proven beneficial (Chini et al., 2011). Furthermore, the fusion of search engine query trends, social media data, and traditional disease prevalence metrics has been effective in improving influenza surveillance (Santillana et al., 2015).

The combination of multiple data sources has also been recognized as a promising approach in the domain of forecasting election results using social media trends (Brito et al., 2021). Different data sources can provide distinct perspectives due to variations in voter decision-making strategies (Lau and Redlawsk, 2006). For example, studies have shown that factors such as national income growth and personal characteristics of candidates can influence election results (Eisenberg and Ketcham, 2004; Put et al., 2019). By integrating different data sources with trends on social networks, researchers can gather complementary insights into voting behavior and election outcomes.

The domain of election forecasting has seen several intriguing examples in which the combination of multiple data sources has yielded more accurate prediction results (Chin and Wang, 2021). One such study conducted by Chin and Wang (2021) focused on predicting the outcomes of Taiwan's 2018 county and city elections. They combined data from Facebook, opinion polls, and prediction markets to enhance their forecast models. The results indicated that the integration of diverse data sources led to increased prediction accuracy. This study showcases the potential of integrating multiple data streams to improve the precision of election forecasts.

Using a similar approach, Liu, Yao, Guo, and Wei (2021) explored the integration of Twitter sentiment analysis with indicators such as per capita personal income growth rate and unemployment growth rate to predict the 2016 US presidential election. Using a lexicon-based machine learning approach, they analyzed sentiment from Twitter data. Interestingly, their classification results were consistent with the actual voting outcomes. However, the regression analysis exhibited relatively high error rates (Liu et al., 2021). These findings underscore the potential benefits of combining social media sentiment with other relevant indicators to produce more accurate election predictions.

3.5.2 Working with Data from Various Sources

Combining multiple data sources into a single data set is a typical task in data science (Han et al., 2023). Data integration is closely related to the concepts of exploratory data analysis (EDA), data pre-processing, data wrangling. To use data from multiple sources in modeling, it has to be transformed to suitable format.

Data pre-processing refers to a systematic process that involves the use of various techniques to prepare raw data for analysis (García et al., 2015). Pre-processing encompasses tasks such as data transformation, normalization, imputation, noise identification, as well as tasks related to data reduction such as feature selection and instance selection (García et al., 2015). Furthermore, data pre-processing involves analyzing measures of similarity and dissimilarity, as well as basic statistical analysis (Han et al., 2023).

Data wrangling is described as a process by which the data is identified, extracted, cleaned, and integrated. The aim is to make the data suitable for exploration and analysis (Furche et al., 2016). Compared to data pre-processing, data wrangling encompasses all activities that are not related to the actual analysis, such as workflow automation and documentation. The data wrangling phase has been identified as the most tedious part of data scientists' work (Kandel et al., 2011).

EDA is described as a detective -like exploration of the data. EDA uses numerical and graphical techniques to uncover patterns and insights in the data (Tukey, 1977). EDA and pre-processing, and data wrangling all share common objectives. In contrast to data wrangling and data pre-processing, EDA places greater emphasis on examining and studying the data itself to reveal notable patterns or characteristics, rather than exploring a predefined hypothesis (Morgenthaler, 2009).

Data integration often encounters challenges arising from variations in data size, format, dimensionality, complexity, noise levels, and information content (Gligorijević and Pržulj, 2015). Accomplishing functional data integration often requires identification and merging of unique entities, correlation tests to establish connections between numeric and nominal data, as well as techniques to handle tuple duplication and resolve conflicts in data values (Han et al., 2023).

Data integration is typically performed in the early stages of a data science project (Wirth and Hipp, 2000). The exact method depends on the case. In practice, the process often requires iterative adjustments as new insights emerge or new data become available. Therefore, efficient data integration requires a flexible and dynamic research process.

3.5.3 Ensemble Modeling

During the modeling stage, data can be included as a feature within the model or modeled individually and later combined into an ensemble model. In the first approach, the model uses features directly from the data. Alternatively, the data undergo separate ensemble modeling process and are subsequently merged to form an combined model.

Ensemble modeling involves combining multiple less accurate models to create a more accurate composite model (Wolpert, 1992; Dietterich, 2000). Ensemble modeling allows for the use of different algorithms for different types of data, suggested to improve predictive accuracy (Wolpert, 1992). Ensemble modeling is a versatile approach to improve predictive performance. The most popular ensemble methods are bagging, boosting, and stacking (Geron, 2019).

It has also been shown that there is no universally optimal combining rule. Even poor classifiers and suboptimal feature sets can contribute valuable information when combined effectively (Duin and Tax, 2000). The best performance is often achieved by combining both diverse feature sets and different methods, making ensemble modeling a robust approach to improve predictive performance.

An important aspect of ensemble modeling is the use of a variety of base learners that may include linear models, decision trees, neural networks, and other algorithms (Dietterich, 2000). For example, a data set might be best modeled using a linear approach, while data from another source might be better suited to a neural network. By employing multiple models, each optimized for different types of data, ensemble modeling leverages the unique strengths of each model type.

Some widely known ensemble methods are AdaBoost, which incrementally builds a strong classifier from several weak classifiers, and Random Forest, which builds multiple decision trees and merges their results to improve the accuracy and robustness of the model (Geron, 2019). Additionally, Gradient Boosting combines weak learners in a stage-wise manner to minimize loss, making it another highly effective ensemble technique. Especially XGBoost (Chen and Guestrin, 2016) gradient boosting algorithm is known for high performance (Geron, 2019). These methods work by aggregating the predictions of various base learners to create a final model that performs better than any individual model alone.

3.6 Ethical Considerations

Researchers should carefully consider ethical implications when working with publicly available online data. This involves clear communication of goals, methods, and results and taking responsibility for the impacts of their research (Boyd and Crawford, 2012; Salganik, 2019). Given that the technological landscape is constantly changing, ethical considerations should be an ongoing process, requiring continual re-evaluation and adaptation to ensure the validity of initial assessments throughout the research project (Salganik, 2019; Kitchin, 2014).

Informational and privacy-related concerns are among the most widely discussed topics when reusing online data for research. These issues are also cornerstones of global data protection policies. Substantial progress has been made in creating frameworks to preserve privacy in data mining, particularly through principles such as data minimization mandated by the GDPR in Europe. This regulation requires generating only the necessary data for a specific task, retaining them only as long as required, and using it solely for that task (Kitchin, 2014). Compliance with these legal frameworks involves anonymizing the data to prevent identification unless explicit consent is obtained (Han et al., 2023; Boyd and Crawford, 2012; Salganik, 2019).

Anonymization, the process of removing personally identifiable information (PII) from data, has raised concerns about its effectiveness and potential limitations (Salganik, 2019). Techniques like k-anonymity ensure that each record is indistinguishable from at least k-1 other records on identifying attributes, making it difficult to re-identify individuals (Han et al., 2023). Despite concerns about traditional anonymization methods, modern techniques appear promising in preserving privacy (Han et al., 2023).

Data ownership is another notable issue when dealing with online data and warrants considering all stakeholders (Salganik, 2019). Researchers should strive to identify and comply with contract and terms of service in their best effort. If they deviate from these agreements, they should explain their decision openly. Researchers should embrace transparency, clear communication of goals, methods, and results at all stages of their research, and the willingness to assume responsibility for their actions (Salganik, 2019).

Especially web-scraping raises intricate questions about ownership and the ethical implications of using the data. Although data may be publicly accessible, their use for secondary purposes may be unwanted. Regardless of ethical questions, web scraping is common practice when APIs are not available or when the use of existing APIs is limited (Broucke and Baesens, 2018).

In general, researchers should distinguish between data being publicly accessible and having permission for use. Public accessibility does not equate to permission for use. There is a difference between data being visible in public and actively seeking

public attention (Boyd and Crawford, 2012). This leads to a nuanced discussion about user consent. Researchers should understand that the public availability of data does not relieve them of their ethical responsibility to obtain informed consent whenever possible.

The ethical landscape of internet research and data mining is complex, but, fortunately, the topic has garnered a lot of academic attention (Van Wel and Royakkers, 2004). Central to this discussion is balancing potential benefits, such as advances in knowledge and policy, against risks like privacy violations.

Several researchers have suggested ethical guidelines that offer a valuable starting point for ethical evaluation. Ethical guidelines are often formulated as checklists or linear models, addressing questions related to legal, privacy, risk, and

Typical ethical guidelines provide checklists designed to help researchers navigate specific research contexts (Townsend and Wallace, 2016). These frameworks aim to tackle a range of considerations, such as legal compliance, privacy issues, risk management, and publication ethics.

Examples of these questions include:

- Does the research involve social media data? Have the terms and conditions of the specific platform been consulted?
- Can social media users reasonably expect to be observed by strangers?
- Are the research participants vulnerable, such as children or vulnerable adults?
- How sensitive is the subject matter?
- Will social media users be anonymized in published outputs?
- Is it feasible to publish or share the dataset?
- Are there relevant disciplinary, funding, legal, or institutional guidelines that must be followed?

Some authors advocate for a more principles-based approach, evaluating research through established rules and general ethical principles to guide researchers through ethical uncertainties (Salganik, 2019). This approach emphasizes four guiding principles for ethical research:

- **Respect for Persons:** Involves obtaining informed consent whenever possible (Salganik, 2019).
- **Beneficence:** Researchers should avoid harm, minimize risks, and maximize benefits (Salganik, 2019).
- **Justice:** Ensures the fair distribution of the risks and benefits of the investigation (Salganik, 2019).

- **Respect for Law and Public Interest:** Extends benefits beyond participants to all stakeholders, emphasizing compliance and transparency (Salganik, 2019).

In addition to ethical problems related to the data, this type of research is also prone to ethical issues when considering the validity and reliability of the research (Salganik, 2019). The researcher does not necessarily know in advance what he is looking for. There is a risk of overfitting and finding correlations where there are none by using black box models. Ethical researcher needs to consider validity of his claims, because it is unethical to present findings that are created by the model, but not observable otherwise.

4 Forecasting Elections

This chapter explores how historical data, political variables, and societal trends are leveraged to predict electoral outcomes. Understanding existing theories and known predictors of electoral behavior is vital for identifying variables that can be used to predict election outcomes (Cao et al., 2010). In addition, this chapter discusses the Finnish political landscape, including its electoral system and the specific characteristics that need to be considered when building a forecasting model.

4.1 The Role of Electoral Forecasting

Election forecasting is a specialized area within political science that focuses on predicting electoral outcomes. The field of election forecasting has a long tradition and captures significant interest because elections have profound implications for society (Ford et al., 2016). It employs a variety of methodologies, including statistical models, opinion polls, and demographic analyzes (Fisher, 2018). Data generated through election forecasts help to understand the current political landscape, evaluate the performance of political actors, and analyze how various events impact political dynamics (Fisher, 2018).

The literature on forecasting elections predominantly originates from the US and the UK, as these countries have been centers of significant methodological development documented in scholarly journals (Fisher, 2018). In other parts of the world, the academic literature on forecasting focuses primarily on long-term determinants and structural factors (Lewis-Beck and Dassonneville, 2015).

The practice of election forecasting is under continuous evolution, driven by the dynamic nature of democracies and rapid technological advancements. Although traditional long-term factors, such as sociopolitical and economic conditions, have historically played a pivotal role, their predictive precision appears to be shifting in the modern digital environment. Recent evidence indicates that contemporary campaigning practices and digital strategies should also be taken into account, as they appear to increasingly influence electoral outcomes (Nadeau et al., 2020).

Election forecasting is closely tied to theoretical frameworks. It is not just about predicting outcomes; it also involves testing the theoretical models developed within political science (Fisher, 2018). Political science, as a broad discipline, examines power structures, governance systems, political behavior, and related topics such as

public policies and the roles of governments and international organizations. Among the various sub-fields within political science, the study of elections, voting behavior, and public opinion are some of the most extensively researched areas (Fisher et al., 2018). Theories of voting behavior and public opinion are fundamental in the context of forecasting elections. They provide the basis for understanding how and why individuals make their electoral choices, as well as how public sentiment can shift over time.

4.2 Relevant Political Theories

4.2.1 How Voters Decide

Forecasting election results is about predicting which candidates the citizens will vote for. Therefore, the nuances of why certain candidates are preferred over other candidates is central to the concept of forecasting elections.

Democratic theory posits that citizens' policy preferences should influence public policy through mechanisms such as elections and opinion polls. This premise hinges on the assumption that citizens are both informed and engaged. When citizens are uninformed or apathetic, the very foundation of democracy is compromised. The extent to which citizens are uninformed or apathetic has been the subject of scholarly debate, which has led to ongoing discussions in democratic theory regarding the true capacity of citizens to effectively guide policy (Bølstad, 2018).

Doubt of the ability and interest of citizens in rational decision making has a long history (Lippmann, 1922; Redlawsk and Pierce, 2016). Evidence suggests that despite improvements in education and information availability, the average voter remains generally uninformed (Bølstad, 2018). Therefore, expecting citizens to behave in a wholly rational and predictable manner seems unreasonable (Simon, 1955; Kahneman, 2003). Decision making is often constrained by cognitive limitations, imperfect information, and time constraints, leading to suboptimal choices (Simon, 1955).

Accepting that citizens are constrained does not necessarily mean that the democratic process does not work (Bølstad, 2018). Many scholars have turned away from the expectations of citizens as rational actors (Redlawsk and Pierce, 2016). It has been suggested that citizens can use heuristics and information cues that in aggregate generate decisions that are satisfactory for the democratic process (Bølstad, 2018). However, while heuristics can help voters make decisions, there is no guarantee that the choices are satisfactory (Bølstad, 2018). Assessing a candidate based on heuristics works only if voters have different policy preferences. Unfortunately, citizens can be collectively biased and make superficial choices (Bølstad, 2018).

The concept of bounded rationality extends beyond information limitations. Emotions also influence individuals, often unrelated to the elections themselves. Public

mood can significantly impact voting decisions, with incumbents being rewarded or punished for unrelated events (Bølstad, 2018). Similarly, researchers have recognized that the position of candidates on the ballot influences election results, following a reversed J-shaped curve where candidates listed early or late have an advantage over those in the middle (Söderlund et al., 2021), indicating that citizens are influenced by factors completely unrelated to the salience of the candidates.

Lau and Redlawsk (2006) propose four models of political decision-making. The four models are rational, confirmatory, fast and frugal, and intuitive. The rational and confirmatory models emphasize cognitive processes, whereas the fast and frugal and intuitive models acknowledge the role of heuristics and limited information search in decision-making (Lau and Redlawsk, 2006).

Additionally, voters can be influenced by predicted election outcomes, leading to strategic voting. A strategic voter maximizes the expected utility of their vote by deviating from their most preferred option due to expectations about the election outcome. Under these circumstances, a less preferred option with a higher chance of success may be chosen (Gschwend and Meffert, 2017). An example of this phenomenon occurred in the Finnish parliamentary elections of 2023, where strategic voting appeared to influence voter behavior, with many opting to shift their support towards major parties in a closely contested race (Kestilä-Kekkonen et al., 2024; Grönlund and Strandberg, 2023).

Understanding the complexity of voter behavior and the theories behind it has several implications for election forecasting. First, given the intricacies of voter decision-making, no model can capture it entirely accurately, and recognizing this limitation is important. However, identifying relevant variables based on heuristics can enhance forecasting models. For instance, if voters partially base their decisions on economic conditions, including such indicators in the model should improve its accuracy—whether voters respond emotionally by punishing incumbents in bad economic times or evaluate the economy rationally is less important than recognizing the variable as is. Additionally, since the importance of different factors in decision-making can shift as public focus changes, forecasting models need to be flexible and adaptive.

4.2.2 Sociopolitical Determinants of Voting Behavior

Sociopolitical determinants of voting behavior have been the subject of extensive study and have a long history in political science (Åsa von Schoultz, 2016). Probably the most noteworthy theory is the theory of party identification, also known as the Michigan model (Hutchings and Jefferson, 2018; Dinas, 2016).

The Michigan Model (Campbell et al., 1960) posits that party identification is a stable psychological attachment to a political party, which significantly shapes electoral decisions. This model highlights how long-term loyalty to a party can influ-

ence voter behavior more consistently than short-term issues or candidate attributes (Hutchings and Jefferson, 2018). As such, party identification acts as a heuristic, simplifying the decision-making process for voters by providing a consistent framework within which they evaluate political choices. Interestingly, party identification appears to explain vote choice far better than policy preferences or ideological positions (Bølstad, 2018).

In addition to party identification, social cleavages are crucial sociopolitical determinants that impact voting behavior. In political science and sociology, cleavages refer to historically established social or cultural divisions within a society that separate citizens into distinct groups with different political interests (Åsa von Schoultz, 2016). The concepts of cleavages are central to understanding how, for example, class and religion can influence political choice (Evans and Northmore-Ball, 2018). The rationale for the powerful behavioral force of social characteristics lies in the enduring and profound influence that these characteristics have in shaping group identities and interests (Åsa von Schoultz, 2016).

Although sociopolitical determinants, such as social cleavages and party identification, are integral to understanding voting behavior, using these concepts in election forecasting can be challenging (Le et al., 2017). The complexity and multifaceted nature of these determinants make accurate predictions difficult. Nevertheless, they contribute to the understanding that voting behavior tends to be relatively stable over time, since individuals' long-term loyalties and identities significantly shape their electoral choices.

4.2.3 The Personal Vote and Candidate

The concept of personal vote centers on how a candidate's personal qualities, qualifications, activities, and historical record influence their electoral support (Zittel, 2016). The traits influencing voting decisions are discussed under the term personal vote-earning attributes (PVEAs) (Isotalo et al., 2020; Von Schoultz and Papageorgiou, 2021; Put et al., 2019). The discussion extends beyond the strategies and activities employed by the candidate to include inherent qualities such as attractiveness and gender.

Personal vote-earning attributes (PVEAs) act as information cues, helping voters make more informed decisions by signaling the credibility and aptitude of candidates (Put et al., 2019). PVEAs are observed to become a more relevant driver of voting behavior in situations where citizens are faced with a long list of candidates to choose from. In those scenarios, electorate will resort more frequently to heuristics or mental shortcuts to make their decisions. Attributes, such as the candidate's birthplace or electoral experience, provide easily recognizable cues that simplify the decision-making process in complex electoral environments.

The importance of personal vote-earning attributes varies across different elec-

toral systems. In systems where more candidates compete for voter attention, these attributes become more significant due to the increased choice available to voters (Shugart et al., 2005). Flexible lists and preference rules enhance competition in proportional representation systems by allowing voters to choose between different candidates from the same party (Zittel, 2016). Specifically, open-list proportional representation systems, such as Finland, emphasize personal vote-earning attributes, as candidates compete within large districts (Shugart et al., 2005). In contrast, the effects of personal vote seeking are typically weaker in plurality systems, such as in the UK (Zittel, 2016). In general, voters who remember a specific candidate are more inclined to vote for them, regardless of their political affiliation and level of political knowledge (Zittel, 2016).

There is also a global trend termed as 'personalization of politics' (McAllister, 2007; Garzia, 2016). The observed trend has shifted the voter focus from party to individual candidates' personalities and attributes. It is argued that a candidate's personal appeal has in many cases become even more significant than the political party they represent (McAllister, 2007). Proponents of the view highlight the increasing media coverage and public attention on candidates' charisma, personal life, and individual competencies. However, evidence from US presidential elections indicates that while personal characteristics are frequently highlighted and emphasized by the media, their actual influence on voter behavior may be relatively minor (Holian and Prysby, 2014, 2020; Wattenberg, 2016). Critics argue that the focus of the media on personality traits can distort public perception, making these attributes appear more relevant than they truly are. Thus, while personalization of politics is a noticeable trend, its real effect on electoral outcomes may be less significant than suggested. Furthermore, evidence suggests that the trend is a geographically limited phenomenon rather than a universal one (Garzia, 2016).

What comes to candidates and their agency, candidates have several strategies to attract more votes and distinguish themselves from other candidates representing the same party in the same district. Three main approaches were identified from the literature: position-taking, credit-claiming, and advertising (Zittel, 2016).

Position-taking involves clearly stating stances on various issues to resonate with voters who share similar views, thereby building a distinct political identity. Evidence suggests that aligning policy positions closely with the party's stance often leads to electoral success (Von Schoultz and Papageorgiou, 2021). Credit-claiming highlights a candidate's achievements and contributions, such as passing legislation or securing local project funding, to establish a reputation for effectiveness and reliability. Advertising uses media channels, from traditional campaign ads to digital marketing on social media, to promote the candidate's image and message, increasing visibility and reinforcing their personal brand. High campaign spending is also positively correlated with electoral success (Maddens et al., 2006).

There are also other influential factors that may affect voting behavior. One

example is candidate attractiveness, which has been found to increase competitive advantage (Klein and Rosar, 2016; Berggren et al., 2010). Other relevant factors include political experience, involvement in local politics, occupational background, sex, age, immigration background, aristocratic title, academic title, membership in political bodies, and celebrity status (Isotalo et al., 2020; Von Schoultz and Papageorgiou, 2021; Put et al., 2019; Klein and Rosar, 2016). The influence on voting behavior is not always linear, as, for example, some researchers suggest that attractive women may experience disadvantages in male-dominated spheres (Klein and Rosar, 2016), although research on Finnish political candidates does not observe such effects (Berggren et al., 2010). Another crucial factor is the familiarity of the candidate with the electorate; less familiar or first-time parliamentary candidates tend to benefit more from attractiveness than politicians with a long track record (Klein and Rosar, 2016).

The concept of incumbency advantage is closely related to the personal vote, referring to the edge that incumbents typically enjoy due to their greater access to resources and higher name recognition garnered through their work. These advantages can significantly boost an incumbent's re-election campaign. This advantage is especially pronounced in plurality systems, such as those in the UK and US, and is less prominent in proportional election systems (Redmond and Regan, 2015). The significance of incumbency-related characteristics varies between different types of elections (Christensen et al., 2020).

The key takeaway from this section is that the personal characteristics of the candidates influence voting behavior. Given this, it should be possible to leverage data on personal characteristics to forecast election results. Demographic information related to candidates, their occupational and political history, and social media data on their popularity can all provide valuable insights for building predictive models.

4.2.4 Economic Voting

Economic voting is a prominent framework in political science that posits voter behavior is substantially influenced by the prevailing economic conditions at the time of the election. According to this perspective, voters are inclined to reward the incumbent party when the economy is performing well, and are likely to punish them during economic downturns. Economic voting is often modeled with vote-popularity (VP) functions, which describe the relationship between economic conditions and the popularity or electoral success of incumbents. This view is supported by a wealth of empirical evidence demonstrating a consistent link between economic conditions and electoral outcomes across numerous democracies (Lewis-Beck and Stegmaier, 2000).

The fundamental idea behind economic voting is that voters' perceptions of economic performance, such as employment rates, inflation, and overall economic

growth, play a meaningful role in their choices during elections (Stegmaier et al., 2017). A large area of research has focused on whether voters base their decisions on past economic performance (retrospective voting) or on their expectations for future economic conditions (prospective voting). Surveys typically address the retrospective angle by asking voters about economic conditions over the past year, while questions about economic expectations for the coming year explore the prospective perspective.

While economic voting is a robust concept, its relevance can be limited in situations where voters find it difficult to attribute economic outcomes to specific parties or candidates (Stegmaier et al., 2017). Furthermore, the effect of economic issues is not always clear-cut. Especially, prospective discussions often involve conflicting views and disagreements over economic policy, as voters assess the potential impact of different proposals. For example, debates over taxation, government spending, and monetary policy can generate significant division among voters. Furthermore, there are situations where issues larger than the economy, such as national security, can influence the impact of economic change (Stegmaier et al., 2017).

An important use case for VP functions in recent years has been the application to election forecasting (Stegmaier et al., 2017). Utilizing economic indicators to predict electoral outcomes has proven to be a valuable tool for understanding voter behavior. VP functions have also been applied to predicting parliamentary elections. The typical variables are unemployment, economic growth, inflation, and change in cross domestic product. Subjective Measures of the Economy have also been included successfully.

One key advantage of the economic voting framework is that there is relatively wide availability of data to test the hypothesis. Economic statistics are widely released and readily available, and election results are public records, making the framework open to empirical testing and validation.

4.2.5 Influence of Mass Media

Mass media is often viewed as a powerful force in contemporary discussion (Newton, 2006). While it is true that the media plays an important role in elections, the extent of its influence remains a subject of debate (Banducci, 2018; Newton, 2006). As a primary source of information, the mass media ensures that citizens stay informed about current events and policy issues (Banducci, 2018).

The theory of agenda setting suggests that the media can shape public discourse and influence which topics are focused (Banducci, 2018). Historically, researchers have viewed the media's effects as mostly reinforcing existing predispositions. Only a few decades ago was the impact of the media on elections begun to be rigorously studied (Banducci, 2018). Current understanding suggests that the role of the media in determining election outcomes is modest and empirical results are mixed (Newton,

2006; Banducci, 2018).

The complexity of voting behavior is further compounded in the modern digital media environment, where algorithms and personalized content intensify selective exposure to media sources and stories that align with existing narratives (Banducci, 2018). There is an ongoing discussion related to filter bubbles and echo chambers, which are described as the phenomenon where individuals are exposed only to information that reinforces their preexisting beliefs, leading to reduced exposure to contrasting viewpoints (Zuiderveen Borgesius et al., 2016). Whether this phenomenon is more pronounced today than in the past is unclear, but what is evident is the much higher number of media channels available today, enabling new types of influence.

Beyond exposure to information, modern digital media can also be used as a tool for persuasion and manipulation (Isaak and Hanna, 2018; Bond et al., 2012). Current research indicates that social network effects and peer pressure on social media can influence voting behavior, introducing new dynamics in the media landscape that were not present in traditional mass media (Bond et al., 2012). The Cambridge Analytica scandal in 2018 brought to light how personal data harvested from social media platforms, browsers, online purchases, and more, could be leveraged to create highly targeted political advertisements (Isaak and Hanna, 2018). The highly targeted political ads were designed to influence voter behavior by exploiting individual psychological profiles and reinforcing specific narratives, and were deemed highly successful in the Trumps 2016 presidential campaign.

Digitalization has also enabled new approaches to studying the influence of the media. For instance, media influence has been studied using Google search trends (Whyte, 2016). Sensitivity tests demonstrate that web search information can serve as a proxy for public opinion. Additionally, empirical results suggest that while candidate visibility generally correlates with increased public interest in political issues, this interest is subject to other moderating effects.

Media content also holds potential for election forecasting. Researchers have used media content to predict brand importance and, subsequently, election outcomes (Colladon, 2020). By analyzing online news articles using a combination of social network analysis and text mining, researchers developed a semantic score to gauge popularity of various policy stances. The results were highly accurate, suggesting a link between the content of the media and the electoral results (Colladon, 2020).

Studying digital media content is feasible because media content is often publicly accessible, at least at the headline level. There are global news databases, such as GDELT (Project, nd), which provide extensive and accessible data for researchers. With the tools of text mining, media content can be systematically analyzed, making it a valuable resource for election forecasting. Furthermore, the availability of advanced computational methods allows researchers to uncover sentiment, patterns, and trends that were previously difficult to detect, enhancing the accuracy and reliability of election forecasts based on media analysis.

4.3 Electoral Systems

4.3.1 Types of Elections

The performance of forecasting models varies significantly between different regions, forms of democracy, types of elections, and electoral systems. The accuracy and efficiency of forecasting models are influenced by several factors, including the citizens' level of interest and engagement, which may change based on the specific election type, prevailing political environment, and salient regional issues.

The most studied elections are national legislative elections, such as parliamentary elections, and presidential elections (Golder, 2005). These elections often capture the highest level of public attention and participation, and consequently, substantial academic and media focus (Schmitt and Teperoglou, 2017). However, there are also other types of elections, such as referendums, general elections, and municipal elections, which receive varying levels of attention.

Legislative elections involve the selection of representatives to the legislative body, which may be a parliament, congress, or other forms of legislative assemblies (Golder, 2005). The structure and rules of legislative elections can vary significantly between countries. Traditionally, legislative elections have been categorized by whether they employ majoritarian or proportional formulas. However, this dichotomy is overly simplistic given the emergence of numerous electoral systems with varying complexity.

Majoritarian electoral systems typically award seats to individual candidates or parties that receive the most votes in a specific district (Golder, 2005). The most common majoritarian system is the plurality rule, where the candidate with the highest number of votes wins, even if they do not achieve an absolute majority.

Proportional representation (PR) systems are designed to allocate legislative seats in direct proportion to the votes each party receives (Golder, 2005). These systems aim to better reflect the electorate's political landscape and enhance representation opportunities for smaller parties. Among the various types of PR, list proportional representation (LPR) systems, including open-list proportional representation (OLPR), are the most prevalent. Although preferential voting systems, such as the Single Transferable Vote, are also used, they are less commonly adopted compared to LPR systems (Golder, 2005).

LPR systems offer several variants for seat allocation, primarily using the quota method or the highest average method (Golder, 2005). Quota systems distribute seats based on a predetermined vote threshold, utilizing variations such as the Hare quota, Droop quota, and Imperiali quota (Golder, 2005). Conversely, highest average systems use methods like the d'Hondt series and Sainte-Laguë series to calculate the highest average number of votes per seat (Golder, 2005).

In LPR, candidates may be presented on either an open-list or closed-list (Golder, 2005). Both approaches are widely adopted. The OLPR system stands out for the

large number of variations in how it is implemented, rather than a single set of common rules (Wall, 2021). A typical feature of OLPR is that it allows voters to influence the ordering of candidates on the party list, which introduces an element of voter preference into the proportional allocation.

Multi-tier systems are electoral structures where a single electoral formula is applied across multiple levels or tiers (Golder, 2005). These systems can be differentiated into those that are connected and those that are not. In connected multi-tier systems, linkage occurs when unused votes from one tier are transferred to another tier, or if the allocation of seats in one tier is conditional on the seats received in another tier. This linkage often aims to enhance proportionality and ensure that votes contribute to the election outcome at multiple levels. One common approach within these connected systems is quota-based proportional systems, where remaining seats after initial allocations are distributed in a higher tier using predefined quotas.

Mixed electoral systems combine elements of both majoritarian and proportional representation formulas within the same electoral framework, aiming to balance the advantages of each approach by promoting broader representation while maintaining a connection to geographic constituencies (Golder, 2005). These systems typically allocate a portion of the legislative seats through majoritarian rules, such as first-past-the-post, and the remaining seats through proportional representation methods. This dual approach seeks to capture the strengths of both local representation and proportional fairness. Mixed systems are becoming increasingly common, as they offer a pragmatic solution to the limitations of purely majoritarian or proportional systems.

Democratic presidents are elected using one of five methods: the plurality rule, absolute majority rule, qualified majority rule, single transferable vote (STV) or electoral college system (Golder, 2005). In the plurality rule, the candidate with the most votes wins, even without an absolute majority. The absolute majority rule requires a candidate to secure 50% of the vote, often leading to a runoff if no candidate initially meets this threshold. The runoff can complicate forecasting due to strategic voting and coalition building. The qualified majority rule demands an even higher vote threshold and is less common.

The STV is a preferential system in which voters rank candidates and votes are transferred according to preferences until a candidate achieves the required majority (Golder, 2005). The electoral college system, most notably used in the US, involves electors chosen by voters who then elect the president, adding another layer of complexity to forecasting efforts. Each of these methods influences electoral dynamics and presents unique challenges for accurate prediction models.

4.3.2 Election Salience

Not all elections carry the same weight for citizens. Some elections, such as the U.S. midterm elections, are often deemed less important compared to major electoral events (Schmitt and Teperoglou, 2017). The character of the electoral contest is significantly influenced by the perceived political importance, or salience, of the office to be filled, which varies across different types of elections.

Elections for high-profile positions, such as members of national parliaments or presidents with executive powers, are considered high-salience elections (Schmitt and Teperoglou, 2017). These contests attract substantial voter interest and media coverage, partly due to their direct impact on national governance and policy direction. Consequently, forecasting models for these elections benefit from a wealth of data and heightened voter engagement, which can enhance their accuracy.

Low-salience elections often suffer from a lack of comprehensive polling data and reduced media scrutiny, which can lead to less reliable predictive models(Schmitt and Teperoglou, 2017). Voter behavior in these elections can also be more volatile or influenced by local issues that are less predictable compared to the broader national trends observed in high-salience elections. Additionally, the lower political stakes might result in strategic voting or protest votes, further complicating forecasting efforts.

Less important elections are often influenced by the dynamics of more significant electoral contests. For example, US midterm elections are affected by the preceding presidential elections, with voter preferences and party strategies often reflecting the outcomes and issues from the higher-stakes contests. However, this relationship varies between different democracies (Schmitt and Teperoglou, 2017). Typically, when studying electoral decisions, such as whether to participate in an election and which party to support, researchers do not analyze these choices in isolation. Instead, they usually consider these decisions in relation to more important elections. Voter behavior and attitudes in low-salience elections are often shaped by the political environment and party performance in high-salience elections. This influence means that low-salience elections are part of a broader pattern of electoral behavior.

4.3.3 Implications for Forecasting

High-salience elections, such as presidential and national parliamentary elections, typically benefit from extensive polling, media coverage, and voter engagement. These factors increase the amount of data that forecasting models can leverage to improve accuracy. In contrast, low-salience elections often suffer from limited data availability and poorer quality polling information (Schmitt and Teperoglou, 2017).

The number of candidates running in an election significantly influences voter behavior and the complexity of forecasting models. In plurality first-past-the-post

(FPTP) systems, elections often involve a small number of candidates, typically two main contenders. Although this might seem straightforward, the low number of candidates can be challenging for statistical methods due to the limited variability and smaller sample size for modeling. Every vote can significantly impact the outcome, making the predictions sensitive to minor shifts in voter preferences.

In multiparty systems, there can be a large number of candidates. The high number of candidates introduces significant complexity into voter decision-making processes difficult to predict. Voters not only choose their preferred candidate, but also consider how their vote will influence the overall party list and the resulting seat allocation (Gschwend and Meffert, 2017).

Elections are often influenced by regional issues and local contexts, which can create variability that national-level models may fail to capture. Factors such as local economic conditions, regional party dynamics, and specific regional concerns can significantly impact voting behavior. Incorporating regional data and understanding local contexts are essential for improving the accuracy of forecasts, especially in decentralized democracies with significant regional variance.

Finally, lower-salience elections are also greatly influenced by broader electoral cycles (Schmitt and Teperoglou, 2017). For instance, US midterm elections may be impacted by the president's approval ratings and recent legislative performance, while local elections might be more insulated from national trends (Tufte, 1975). Forecasting models must adapt to the specific timing and context within the broader electoral cycle to make accurate predictions across different types of elections.

4.4 Methodological Approaches

4.4.1 Summary of Key Methodologies

The methodological approaches most widely used in election forecasting are opinion polling, structural modeling, and expert opinions. Opinion polling is one of the most widely used approaches. In this method, pollsters conduct vote-intention polls by asking citizens about their voting preferences, and these data are used to predict voter behavior (Williams and Reade, 2016; Fisher, 2018). Structural models, also known as statistical models, utilize regression techniques, and other statistical methods to process historical data and focus on causal relationships using predictors such as economic and political factors (Williams and Reade, 2016; Lewis-Beck and Stegmaier, 2014a; Fisher, 2018). Expert opinion, or judges, involves assessments from political experts who use their knowledge and experience to provide qualitative forecasts based on various inputs, including polls, models, and market data (Williams and Reade, 2016; Lewis-Beck and Stegmaier, 2014a).

Other approaches to electoral forecasting also provide valuable information. Big data methods analyze content from social media platforms to predict election out-

comes (Fisher, 2018). One approach is to use the so-called wisdom of the crowd, which involves aggregating individual predictions to take advantage of the collective intelligence of a large group, thus improving the accuracy of the forecast (Fisher, 2018). Prediction markets offer another method, where participants trade contracts based on election outcomes, effectively consolidating dispersed information into a collective forecast (Williams and Reade, 2016; Fisher, 2018).

Multiple different methods can also be combined in election forecasting to take advantage of the strengths of multiple approaches (Fisher, 2018). Aggregators combine data from multiple opinion polls to create a more comprehensive forecast by incorporating diverse polling results. On the other hand, synthesizers combine multiple data sources, such as polling and structural models, along with historical evidence, to refine their predictions and improve accuracy (Lewis-Beck and Stegmaier, 2014a; Fisher, 2018).

Different methods for forecasting elections are better suited to specific situations, often dictated by the availability of data. For example, in newer democracies where economic data may be sparse or non-existent, a structuralist approach may not be feasible. Additionally, electoral behavior can vary significantly from one election to the next, with certain issues becoming more salient, and thus making some forecasting methods more accurate than others. Evidence suggests that methods that were highly accurate in earlier elections did not necessarily perform well in subsequent ones (Graefe, 2019).

4.4.2 Structuralist Methods

The structuralist approach to predicting elections is based mainly on various political and economic factors to predict electoral outcomes (Kennedy et al., 2017; Lewis-Beck and Stegmaier, 2000, 2014b). Central to structuralist approach is the theory of Economic Voting, which posits that voters' decisions are heavily influenced by their perceptions of the economic performance under the incumbent government (Lewis-Beck and Stegmaier, 2000; Nadeau and Lewis-Beck, 2020).

A large body of literature highlights the effectiveness of these fundamentals in forecasting legislative and presidential elections, focusing on variables with low variability, such as economic trends and unemployment rates (Kang and Oh, 2024; Kennedy et al., 2017). Typically, these models employ regression techniques to develop explanatory models at the national level, focusing on factors such as presidential popularity and economic growth (Lewis-Beck and Stegmaier, 2014b).

Economic growth and unemployment are often highlighted as key indicators and are widely used (Kennedy et al., 2017; Lewis-Beck and Dassonneville, 2015). Economic factors have been found to significantly impact government vote shares, particularly in countries where there is a clear attribution of policy responsibility (Lewis-Beck and Stegmaier, 2000). In addition, inflation rates are another useful variable,

with empirical evidence suggesting that voters tend to punish governments for increases in inflation (Lewis-Beck and Stegmaier, 2000).

The structuralist approaches to election forecasting are particularly advantageous when it comes to longer lead times, offering reliable predictions well ahead of the election date. Unlike other methods that may require the proximity of the election for accurate forecasts, structuralist models can provide robust predictions several months in advance. Research suggests that the optimal lead time for these models is approximately three months before the election (Jennings et al., 2020). This extended lead time allows for a thorough analysis of long-term macroeconomic and political indicators, thereby enhancing the models' predictive power and offering valuable insights into electoral outcomes well before the campaign period intensifies.

The effects of macroeconomic variables can be mixed and context-dependent. Some studies show strong correlations between economic performance and electoral outcomes, while others find that these relationships depend on the political context (Lewis-Beck and Stegmaier, 2000). For example, in countries with multiple ruling parties or coalitions, the diffusion of responsibility can make it challenging for voters to attribute economic conditions to a specific party. This variability in the impact of macro fundamentals underscores the complexity of using these indicators for election forecasting, necessitating nuanced models that take political context into account.

Although objective measures are typically discussed, it is important to note that voters' subjective perceptions of the national economy influence their voting behavior (Lewis-Beck and Stegmaier, 2000). Economic conditions can shape perceptions, but these perceptions can also be unrelated to the fundamentals in some situations. Especially in multiparty systems, the diffusion of government responsibility complicates voter decision-making. For instance, in the UK, with clearer lines of responsibility, voters may find it easier to hold the government accountable compared to Italy, where coalition governments spread responsibility among multiple parties (Lewis-Beck and Stegmaier, 2000). Considering how important subjective perception is, there seems to be a research gap in understanding whether economic stimulus or social welfare mitigates the impact of negative economic conditions.

Structural models have generally performed well in predicting election outcomes before the commencement of election campaigns (Nadeau and Lewis-Beck, 2020). While there has been concern that the declining influence of long-term factors, coupled with the growing importance of campaign dynamics, could increase error rates in models based purely on fundamentals, there is currently no evidence to support this worry (Nadeau and Lewis-Beck, 2020).

However, structural models do face challenges, particularly in multiparty settings. Performing accurate forecasts of the performance of new and smaller parties remains difficult (Stoetzer et al., 2019; Walther, 2015). The complexity of assigning political and economic responsibility in environments with multiple parties complicates prediction accuracy, necessitating continual refinement of these models to

account for evolving political landscapes.

Most of the research on structuralist models has been conducted in the US (Lewis-Beck and Stegmaier, 2014b). However, these models have also been successfully applied in various European contexts, demonstrating their versatility and adaptability (Lewis-Beck and Dassonneville, 2015). For example, the structuralist approach has proven effective in Irish elections by focusing on incumbent coalitions, producing accurate predictions despite the complexities of a multiparty system (Quinlan and Lewis-Beck, 2021). In the UK, structural models have outperformed poll-based projections, especially in out-of-sample forecasts, showcasing their robustness and reliability (Mongrain, 2021).

Nevertheless, challenges arise in regions like Scandinavia, where vote share instability and intricate political landscapes pose significant difficulties (Nadeau and Lewis-Beck, 2020). Despite these challenges, there are successful cases, such as Denmark, where economic conditions have been shown to significantly impact election outcomes, confirming the relevance of structuralist models in capturing the relationship between economic performance and voter behavior (Nadeau and Lewis-Beck, 2020). This adaptability underscores the need for continual refinement and context-specific adjustments to enhance prediction accuracy across diverse political environments.

Despite advances in structuralist models, several open questions and challenges remain. One key question is whether voters genuinely understand the state of the economy and how this knowledge influences their voting behavior (Lewis-Beck and Stegmaier, 2000). The accuracy of economic perceptions among the electorate can significantly impact the reliability of forecasts based on economic indicators. Furthermore, the responsibility problem in multiparty systems poses an enduring challenge: assigning economic and political responsibility to individual parties becomes increasingly complex when multiple parties share governance (Quinlan and Lewis-Beck, 2021; Nadeau and Lewis-Beck, 2020).

There is also an ongoing need to continually refine forecasting models to capture the evolving dynamics of elections and voter behavior. As political landscapes and voter preferences shift, models must be adaptable to remain accurate and relevant. These issues can be addressed by incorporating contextual information, although this approach is still in its infancy (Kang and Oh, 2024).

Structuralist models for election forecasting rely on a variety of data sources to obtain the necessary political and economic indicators. Key data sources include national statistics agencies that provide critical metrics such as GDP growth, unemployment rates, and inflation rates. International organizations such as the International Monetary Fund (IMF), the World Bank and the Organization for Economic Cooperation and Development (OECD) also provide valuable economic data. Historical voting patterns are sourced from national electoral commissions.

4.4.3 Survey Methodology

Survey methodology, or polling, involves asking the public about their opinions. Both polling and structuralist approaches offer unique advantages. While the structuralist approach can fail to capture contextual issues that might influence the political landscape, opinion polls are adept at detecting these nuances. However, polling is also more sensitive to these contextual issues, which can often result in poor performance long ahead of elections. Multiple studies indicate that structuralist methods are more accurate with a longer lead time, whereas the accuracy of surveys improves as the elections approach (Lewis-Beck and Dassonneville, 2015; Williams and Reade, 2016).

The survey methodology is extensively used by commercial pollsters who use it beyond election forecasting to study public opinion. Polling enables informed citizens to anticipate, verify and comprehend political, social, and economic aspects of life (Bailey, 2024). Polling is a very popular endeavor globally, and there are numerous polling agencies (Victor, 2021). For example, Gallup, an US based polling agency operating since 1935, conducted around 500 polls only in 2015. The polls covered topics ranging from the economy and personal financial situations to the political climate and major news events (Newport, 2016).

Polling is widely used in election forecasting. Polls can be used individually, aggregated with other polls, or synthesized with other data types to enhance predictive accuracy. Polling functions as a predictive tool rather than an explanatory tool, and it is not based on election theory (Lewis-Beck and Dassonneville, 2015). Despite their extensive use in forecasts and predictive algorithms, polls carry substantial uncertainty about their efficiency. This uncertainty is compounded by the judgment of the surveyors, who often apply unpublished weighting schemes to ensure that the voters sampled are representative of the general population (Fritsch et al., 2024; Bailey, 2024).

Polling is currently facing a significant crisis. The industry has encountered notable difficulties in accurately predicting the outcomes of elections and referendums in recent years (Mongrain, 2021). For instance, polling efforts have recently failed in the US during the 2016 and 2020 presidential elections (Bailey, 2024), as well as in the UK during the Brexit referendum (Mongrain, 2021). Ensuring the accuracy of the surveys is crucial, but frequent inaccuracies have led to widespread distrust (Bailey, 2024). In addition, traditional opinion polling is struggling with rising costs and declining response rates in the developed world, even as demand for survey statistics continues to grow (Groves, 2011).

The literature suggests at least three different reasons why polls might struggle to accurately capture public opinion. First, polls cannot adequately address non-ignorable non-response rates, which skews the results (Bailey, 2024). Second, polls themselves can influence voting behavior, further complicating their reliability. This

explanation appears less significant according to current evidence, but should be studied in more detail (Roy et al., 2021). Third, some authors suggest that polls are intentionally designed to be misleading and are used to influence public opinion rather than measure it (Moore, 2008). There is limited support for the last claim, especially in the context of elections.

The non-response rates, particularly the non-ignorable non-response rates, significantly affect polling accuracy. A non-response rate measures how many people choose not to participate in a survey. Non-ignorable non-response occurs when those who do not respond have systematically different characteristics or opinions than those who do, leading to biased results and affecting the poll's accuracy and reliability (Bailey, 2024).

High non-response rates pose a substantial challenge, with some authors reporting rates exceeding 95 percent in the US (Bailey, 2023). Those who participate in surveys are willing to do what most people won't: engage with pollsters (Bailey, 2024). This introduces substantial bias since respondents likely differ from the general population. While weighting helps correct demographic imbalances, it cannot fully address the differences in opinions or characteristics between respondents and non-respondents. In essence, the reasons behind non-responses can significantly skew survey results (Bailey, 2024).

In addition to traditional survey methods, some researchers have explored forecasting elections using non-representative polls. This approach is intriguing, but it currently lacks conclusive evidence of its overall effectiveness. For example, Wang et al. (2015) conducted a study using data collected through Xbox, a platform that primarily attracts a younger, more tech-savvy demographic, leading to unrepresentative polling data. While the authors reported promising results, such data does not accurately represent the broader population, making it difficult to draw reliable conclusions about public opinion (Bailey, 2024).

Finally, it should be noted that while a well-conducted survey is usually a very efficient way to measure public opinion, it is often costly and not always available, particularly for low-salience elections. Although survey data is generally accessible for high-profile elections, it tends to be scarce for elections that attract less public interest. This scarcity can pose challenges for accurately gauging public sentiment in those less prominent contests.

4.4.4 Social Media

The use of data from social networking services to forecast election outcomes has attracted a great deal of academic interest. The central idea is that user-generated trails, often referred to as big data, incidentally collected data (Mellon, 2018), or digital traces (Jungherr et al., 2017), can provide valuable insight into predicting election results. However, the application of social media data for forecasting re-

mains controversial due to concerns about its reliability and accuracy (Gayo-Avello, 2013). Despite its promise, challenges such as demographic biases and the representativeness of the online population persist, making the effectiveness of social media data in election forecasting an ongoing area of investigation (Huberty, 2015).

Initial studies used rather simplistic methods. For instance, Tumasjan et al. (2010) found that the number of followers a political party had on Twitter closely correlated with the election results. Similarly, O'Connor et al. (2010) identified a strong correlation between Twitter sentiment and the results of the opinion polls. Franch (2013) claimed that platforms like Facebook, Twitter, Google, and YouTube could be used effectively to predict election results (Franch, 2013). DiGrazia et al. (2013) demonstrated that reliable data about political behaviors could be extracted from social media. Numerous other studies were also published, often claiming remarkable accuracy.

The initial studies received considerable critique. For example, Tumasjan et al. (2010) was directly contested by Jungherr et al. (2012), who showed that the forecast would have failed if the Pirate Party had been included in the analysis. Similarly, Chung and Mustafaraj (2011) demonstrated that the current simple methods for predicting election results based on sentiment analysis of tweet texts perform no better than random classifiers. Huberty (2013) echoed these findings, suggesting that any additional information captured by these simplistic forecasts, beyond incumbency, is unreliable for future election predictions. Moreover, many researchers have highlighted that sampling social media data poses a major challenge, as a random sample from social media users does not accurately represent the general population (Gayo-Avello, 2013; Schoen et al., 2013).

Over time, numerous recent studies have emerged. Currently, there are several literature reviews which provide valuable insights on the topic. For example, Chauhan et al. (2021) found that most studies on election predictions have used Twitter as their primary data source, with the majority achieving successful predictions. This underscores Twitter's value in understanding public opinion and accurately forecasting election outcomes.

However, Santos et al. (2021) argue that while social media analyses offer valuable insight into public opinion, they are not substitutes for traditional polls. They note that most published studies rely on a count-based approach using Twitter data, and the forecasting accuracy of these studies remains ambiguous. They emphasize the need for further research and data sharing among researchers to fully understand social media's potential in polling.

Brito et al. (2021) observed that most studies linking online sentiment with election outcomes focus on a single election and use Twitter data. They found that studies that are based solely on Twitter tend to have significantly lower success rates compared to those that use other social networks, such as Facebook.

Similarly, Gayo-Avello (2013) highlighted that the predictive power of Twitter

is often overestimated. He recommends using predicted vote rates or the number of votes rather than the winner, as the latter can be attributed to chance. Furthermore, he found that simple baselines can achieve better accuracy than Twitter predictions in many cases.

Phillips et al. (2017) noted that existing research often relies on relatively simple methods, ranging from linear regression to keyword matching. However, results successful in one context often fail in others. They suggested that incorporating domain-specific knowledge can guide statistical models to produce better results.

Rousidis et al. (2020) found that only about half of the studies managed to achieve a valid prediction. Furthermore, more than two-thirds of the research use data solely from Twitter, which has been demonstrated to be insufficient to generate highly accurate predictions.

Skoric et al. (2020) showed that combining multiple data sources can improve prediction performance. However, there are not enough studies that utilize this approach to draw definitive conclusions. They caution against replacing survey-based studies with social media analytics, suggesting that the latter should be used to gain additional information on public opinion and political behavior. They also call for more theoretically informed work that pays attention to the underlying mechanisms and processes.

As the literature shows, two methodological approaches emerged as the most common: counting metrics (such as mentions, likes, or followers) and sentiment analysis. Both methods are similar to traditional opinion polling in their foundational principles. Subsequently, both face the challenge of unrepresentative sampling. The demographic and behavioral biases of social media users mean that the data often do not accurately reflect the general population. Sentiment analysis, which interprets sentiments expressed in social media posts to gauge political attitudes and behaviors, also shares the limitations associated with its data sources (Le et al., 2017). Moreover, most studies focus on elections with low sample sizes and single-race events, limiting their explanatory power in the long term.

Twitter has historically been the most widely used data source for election predictions based on social media. However, its recent rebranding as X and the removal of the public API highlight how rapidly the landscape of social networks can change. Over the last decade, social media trends have changed and younger voters have increasingly moved away from platforms like Facebook to TikTok (Strandberg et al., 2024). This shift underscores the rise and potential decline of previously dominant platforms, including X, illustrating the dynamic nature of social media use and its implications for electoral predictions.

The evolution of the social media landscape has a negative impact on long-term studies. Changes in the popularity of the platform can alter the demographic composition and behavior of the user base, affecting the representativeness of the data. As new platforms emerge and old ones decline, researchers have to continuously adapt

their methodologies to ensure that they are capturing a broad and accurate sample of the electorate. Furthermore, different social media platforms have unique characteristics that influence user interactions and the nature of the data available for analysis. For example, the brevity of X contrasts with the multimedia content found on platforms such as Instagram or TikTok.

Despite known drawbacks, evidence shows that the number of followers on social media correlates with the number of votes a candidate receives. Similarly, sentiment analysis, which assesses positive or negative sentiments towards a candidate, often aligns with their electoral success. This correlation appears robust even after controlling for various variables, although some authors suggest that digital traces add little additional information beyond what conventional methods already provide (Cerina and Duch, 2020).

The traditional perspective of voting behavior does not really distinguish social media from other factors influencing how voters decide. Existing theories are primarily interested in causal mechanisms, whereas the relationship between social media indicators and election results is primarily predictive, rather than explanatory. Therefore, the basic premise of using social media metrics to forecast elections is that the act of becoming a candidate's social media follower is an indication of personal preference (Kosinski et al., 2013).

Meaningful use of social media data in public opinion research would require addressing sampling issues. Social media platforms do not provide demographic information about users, making it impossible to apply weighting techniques to adjust for sampling biases. Consequently, social media samples are not representative of the general population (Mellon and Prosser, 2017; Cerina and Duch, 2020).

Furthermore, even if weighting were possible, self-selection bias would still be difficult to address. Like in survey research, the reason people choose to use social networking services can differentiate users from the general population. Therefore, to extract meaningful information from social media content, it would be crucial to address the non-ignorability of the selection process (Bailey, 2024). Regardless of what big data enthusiasts might suggest, the volume of observations does not compensate for the lack of representativeness. A small, truly random sample will produce a error far smaller than that of a non-random sample several orders of magnitude larger (Bailey, 2024).

Some efforts have been made to address the challenges of non-representative data. For example, some methodological developments are taking place in the realm of survey research to improve representativeness and accuracy (Cerina and Duch, 2023; Wang et al., 2015). These advances are essential to improve the reliability of public opinion research based on social media data.

Social networking services can also be investigated through media effects research, which examines how social media services are used as tools to influence the electorate as an unique campaigning method. Political science recognizes that cam-

paigning methods matter (Iyengar and Simon, 2000). Evidence suggests that social media advertising had a significant influence on the 2016 US presidential election (Grinberg et al., 2019). Other significant cases include the Brazilian 2018 elections (do Nascimento Silva and Silva, 2019), the 2017 Italian parliament election (Giglietto et al., 2019), the 2017 Presidential Elections in Ecuador (Rofrío et al., 2019), and Malaysia's 14th General Election (Nizah and Bakar, 2019).

Some prior studies have investigated the impact of social networks on the political activities of citizens, revealing mixed findings. On the positive side, evidence suggests that social media can be a powerful tool for disseminating information, mobilizing voters, and shaping political opinions. It can influence individuals' decisions to participate in elections, especially for those seeking political information online or lacking other resources for political engagement. Social media serves as an enabler by lowering barriers to participation through information dissemination, discussion, and mobilization (Aldrich et al., 2016; Bimber et al., 2015; Bond et al., 2012; David et al., 2016; Carlisle and Patton, 2013; Gil de Zúñiga et al., 2014).

However, some studies indicate that social media can also negatively impact political activity. Exposure to conflicting political views on social media has been shown to decrease political participation among certain user groups (Lu et al., 2016; Theocharis et al., 2015). For users who encounter political perspectives that conflict with their own, social media may adversely affect their willingness to engage in political activities (Lu et al., 2016).

Building on the discussion of social media's influence, we can further consider its potential role in voter mobilization through social peer pressure. Numerous studies have explored the role of peer pressure in voter mobilization, showing that people are more likely to vote when they know their peers expect them to (Panagopoulos et al., 2014; Rogers et al., 2017). Researchers have even conducted experiments on social networking services like Facebook, finding that these platforms can be effectively used to increase voter turnout (Haenschen, 2016). Whether a large number of social media followers and therefore increased visibility on citizens' screens has similar effects on potential voters remains unexplored. However, existing evidence on voter mobilization suggests that this visibility could exert a form of peer pressure, thus affecting voting behavior.

In conclusion, while social media offers valuable data for election forecasting, significant obstacles related to sampling biases, theoretical foundations, and model interpretability must be addressed. Future research should focus on developing robust and transparent methodologies that balance predictive accuracy with explanatory power. This approach will help optimize the potential of social media data in understanding and forecasting electoral outcomes.

4.4.5 Other Approaches

In addition to structural models, opinion polls, and social media analysis, valuable insights can also be gained from approaches like prediction markets, expert opinions, and media analysis.

Prediction markets, where participants place bets on electoral outcomes, harness collective intelligence to predict election results. Research indicates that prediction markets tend to offer highly precise forecasts, often surpassing traditional opinion polls, particularly in the period leading up to an election. Prediction markets' accuracy and precision improve as Election Day approaches, contrasting with the declining performance of opinion polls during this same period (Williams and Reade, 2016). This suggests that prediction markets are particularly effective for short-term electoral predictions. Additionally, crowd-based predictions are significant in situations with limited historical data, highlighting their capacity to encapsulate real-time voter sentiment and aggregate diverse informational inputs (Atanasov et al., 2024).

Expert opinions continue to supplement data-driven approaches in the realm of election forecasting. Although they exhibit certain biases and their precision may lag behind other methods, they consistently provide valuable insights. The bias in expert opinion is evident, yet their qualitative assessments often add context to quantitative predictions (Williams and Reade, 2016). Expert opinions can be particularly useful in interpreting complex electoral dynamics and offering nuanced interpretations that purely data-driven methods might overlook.

Analyzing traditional media, such as news in print, online, or broadcasts, can also enrich election forecasting. For instance, the significance of a candidate's presence in online news can serve as an indicator of electoral success (Colladon, 2020). Integrating data from diverse platforms like GDELT, which tracks media coverage globally, may enhance the predictive power. Studies suggest that metrics such as the volume of articles or broadcasts mentioning a specific candidate could provide additional insights (Colladon, 2020).

Google Trends, which can be used to aggregate relevant search queries, can refine election predictions by providing real-time data on public interest. Although its predictive accuracy varies and may not generalize across different elections, it remains a valuable tool that requires more research to fully understand its utility (Behnert et al., 2024).

Various institutions also use aggregated vote intentions collected from opinion polls to improve the accuracy of forecasts (Lewis-Beck and Dassonneville, 2015). This approach takes advantage of the breadth of data from various sources, mitigating biases and errors that may be present in individual polls.

Finally, combining forecasts from multiple methodologies has often been shown to yield the most reliable results. A combined forecast includes comprehensive information and helps offset systematic and random errors that could be present in

individual methods (Graefe, 2015). For example, synthesizing structural models with opinion polls can provide nuanced insights. While a well-constructed structural model might already predict elections accurately months in advance, polls can add crucial, dynamic information as Election Day approaches, capturing campaign effects that structural models might miss. This multifaceted approach provides robust predictions, accommodating a range of variables and reducing the uncertainty inherent in electoral forecasts (Graefe, 2019) (Lewis-Beck and Dassonneville, 2015).

4.5 Finnish Democratic System

4.6 Overview of the Finnish Political Framework

Finland operates under a stable parliamentary democracy characterized by a multiparty system and a framework that emphasizes both representative governance and the rule of law (Jääskeläinen, 2023). The Finnish political system is rooted in democratic principles, ensuring that power resides with the people who exercise it through regular, free, and fair elections. The system fosters transparency, citizen participation, and accountability in governance.

Central to this system are the parliamentary elections, which are governed by the Finnish Constitution (Jääskeläinen, 2023). Sovereign power in Finland belongs to the people and is exercised through the unicameral Finnish Parliament (*Eduskunta* in Finnish), which comprises 200 representatives. These members are elected every four years on the third Sunday of April. All Finnish citizens 18 years or older are entitled to vote. Parliamentary elections are also central to this dissertation and will be discussed in more detail later.

In addition to parliamentary elections, Finnish citizens participate in presidential elections, municipal elections, regional elections, and elections for representatives of the European Parliament (Jääskeläinen, 2023). Presidential elections involve choosing the President of the Republic by direct vote for a six-year term, with the possibility of serving up to two consecutive terms.

Municipal elections occur every four years and elect representatives for municipal councils. In 2025, Finland will have 292 municipalities on the mainland and 16 in the Åland Islands (Ministry of Justice, Finland, 2024c). County elections involve electing councils responsible for overseeing social services, healthcare, and rescue services. Beginning in 2025, county and municipal elections will be held on the same day (Ministry of Justice, Finland, 2024a). Furthermore, Finnish citizens vote every five years to elect representatives to the European Parliament, determining Finland's participation in legislative processes at the European level (Ministry of Justice, Finland, 2024b).

Several major political parties dominate the political landscape in Finland (Karvonen, 2014). The Centre Party (*Suomen keskusta*) has traditionally held significant

influence in rural areas and parts of central and northern Finland. It advocates for both private enterprise and agricultural subsidies. The Conservative Party (*Kansallinen kokoomus*), serving the business elite and middle classes, particularly in urban areas, supports market liberalization while maintaining support for the welfare state. The Social Democratic Party (*Suomen Sosialidemokraattinen Puolue*) emerged from a split in the social democratic movement in 1922 and draws the core support of industrial workers and urban wage-earners, particularly in the southern part of Finland. It has strong ties with major labor unions and emphasizes the welfare state and full employment. The True Finns (*Perussuomalaiset*), known for their emphasis on national sovereignty and social conservatism, have become a significant player in recent years by addressing concerns about globalization and immigration.

In addition to these major parties, several smaller parties contribute to Finland's diverse political spectrum (Karvonen, 2014). The Left Alliance (*Vasemmistoliitto*), a reformed socialist party, appeals to parts of the working class and intellectual radicals with its emphasis on social justice and economic equality. The Greens (*Vihreä liitto*) attract young, urban, and well-educated voters with their focus on ecological issues, liberal social policies, and immigrant rights. The Swedish People's Party (*Svenska folkpartiet i Finland*) serves the Swedish-speaking minority with a moderate stance, leaning right economically but liberal on social issues like immigration and multiculturalism.

The Christian Democrats (*Kristillisdemokraatit*) represent traditional Christian values and focus on moral issues such as abortion and same-sex marriage, despite limited electoral success. Other smaller parties, while not always maintaining a strong presence in parliament, contribute to the vibrancy of Finland's multiparty system (Karvonen, 2014).

4.6.1 Parliamentary Election

This study presents several forecast models specifically developed for Finnish parliament elections, in order to capture the unique dynamics and complexities inherent to this electoral system. Understanding these particularities is important for accurate prediction and meaningful analysis in the context of Finnish politics.

The Finnish parliament consists of 200 seats, elected every four years from 14 multi-member constituencies and one single-member district, with a median district magnitude of approximately 13 seats (Jääskeläinen, 2023). Finnish parliamentary elections utilize a variant of the OLPR system, where voters are required to select an individual candidate they prefer. Parties and alliances nominate candidates for lists, and the ballot numbers are randomly assigned. Voters cast their votes by writing the number of their chosen candidate on the ballot, without the option to vote solely for a party or a list. The cast vote is called a preference vote.

The allocation of seats is determined using the d'Hondt divisor method (Jääskeläinen,

2023). Initially, the votes received by each candidate are summed to determine the total vote count for each party list at the district level. Seats are first allocated by giving the total list votes to the most popular candidate on the list. For subsequent candidates, the list's total votes are progressively divided by integers: by two for the second-most popular, by three for the third-most popular, and so forth. The candidate associated with the highest total wins the first seat in the electoral district. This process continues, with the vote totals being recalculated and seats being awarded in descending order of votes until all seats in the district are allocated.

Finnish parliamentarism shares characteristics common to parliamentary systems with proportional list systems (Karvonen, 2014). Cabinets rely on the support of the parliamentary majority, making party discipline crucial. Although the d'Hondt method marginally favors larger parties, the Finnish system remains highly proportional. The lack of a minimum vote threshold encourages smaller party representation. Mandatory candidate voting requires candidates to actively campaign for personal votes, adding a preferential element unique to Finland compared to many other parliamentary countries. During the analysis period, the largest party garners just over a fifth of the seats.

The parliamentary elections present voters with a large number of candidates, leading many citizens to rely on heuristics and voting advice applications to make informed decisions (Von Schoultz and Papageorgiou, 2021; Isotalo et al., 2023). Heuristics such as attractiveness have been found to correlate with electoral success (Berggren et al., 2010). Furthermore, the position of the ballot introduces some randomly assigned advantages, showing a J-shaped effect where the candidates listed first gain the most significant advantage, those in the middle receive the least, and those placed last also benefit more than those in the middle (Söderlund et al., 2021).

Digital technologies and social media significantly influence Finnish elections, highlighting the electorate's digital adeptness and connectivity (Kestilä-Kekkonen and von Schoultz, 2020; Strandberg, 2013; Strandberg et al., 2024). Furthermore, the lack of clear attribution of policy responsibility and economic accountability means that economic factors may exert an unclear influence on voting behavior (Lewis-Beck and Stegmaier, 2000). Lastly, the role of parties is declining, which reduces their importance as a heuristic for voters (Karvonen, 2014).

Periods of populist protest have recurrently characterized Finnish electoral and parliamentary politics, marking the current era as the third significant wave of populism. This pattern reflects a reaction against many features of Finnish party politics and illustrates ongoing volatility within the political landscape (Karvonen, 2014).

Overall, the Finnish parliamentary elections provide a unique and compelling environment for studying and forecasting election outcomes. The interplay of complex electoral rules, societal transformations, and modern campaign strategies presents both challenges and opportunities for accurate prediction. This specificity highlights the importance of tailored forecasting models that can account for the nuances of

different electoral systems.

4.6.2 Forecasting Elections in Finland

The most prominent source in political forecasting in Finland is opinion surveys. Opinion poll data, often a commercial product, is widely published and readily available online. For example, in the US, thousands of polls are conducted, and aggregators combine poll results to enhance accuracy. Polls are popular regardless of the challenges related to representativeness (Newport, 2016).

In Finland, polling is conducted by three main pollsters, with occasional local polls also performed. Finnish polls typically focus on party-level support rather than individual candidates. The limited number of polls in Finland reduces the potential for aggregating results, which is a common practice in regions with more prolific polling activities.

Another less widely used approach in multiparty election forecasting is the structuralist approach, which often incorporates economic indicators to predict election outcomes. Although this method is frequently used in two-party systems, it is less adapted to multiparty contexts (Nadeau and Lewis-Beck, 2020). Furthermore, there are no substantial publications on the use of economic models to predict Finnish elections. However, research suggests that economic models may be able to reveal whether the incumbent will remain in power (Nadeau and Lewis-Beck, 2020). This indicates that pursuing a structuralist approach could have potential applicability in the Finnish context, although it remains unexplored.

4.7 Validation and Evaluation

Forecasting elections always includes uncertainty (Lewis-Beck, 2005). Accurate prediction of election outcomes involves more than just synthesizing large amounts of data and also requires a rigorous evaluation of the principles that drive these forecasts. This section aims to distill the core principles that underlie effective evaluation of election forecasts. According to the literature, the key metrics to evaluate the quality of different models as forecasting instruments go beyond the assessment of accuracy; they also include lead time, parsimony, and reproducibility (Lewis-Beck, 2005).

Accuracy refers to the degree to which the forecast predicts the outcome. Accuracy is the primarily mean to asses the quality of the forecast (Lewis-Beck, 2005). It can be measured by employing classical regression analysis of past forecasts and elections to evaluate how well the model fits the data (Lewis-Beck and Stegmaier, 2014a).

A prevalent approach in electoral forecasting is the mean absolute error (MAE) (Lewis-Beck and Stegmaier, 2014a). MAE is the sum of absolute errors divided

by the sample size, and offers a straightforward measure of the prediction error (Shcherbakov et al., 2013). However, MAE is scale dependent and comparisons can only be made between forecasts if the scales are identical (Hyndman, 2006). While MAE can be useful for comparing different models with results presented in identical ways, comparing these metrics across different elections may be unfeasible due to variations in fundamental factors and available data.

Depending on the object of the forecast, different metrics may be more appropriate (Geron, 2019). For instance, in classification tasks, the F1 score is useful for evaluating performance, especially in cases with imbalanced datasets. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both false positives and false negatives.

For continuous outcomes, providing percentage error values, such as the root mean square error (RMSE), can improve understanding of the precision of the forecast (Lewis-Beck, 2005). RMSE is particularly valuable as it penalizes errors more than MAE, giving a alternative view of the accuracy of the prediction. Although MAE is frequently used, it is important that the results are clearly presented to validate the model's efficacy fully.

Accuracy is not a sufficient metric on its own (Lewis-Beck, 2005). Another essential metric is lead time, which refers to how much in advance a model can produce accurate forecasts (Fritsch et al., 2024). Longer lead times are typically considered to be better, provided that the accuracy does not suffer significantly.

Balancing lead time with accuracy can be challenging. Although shorter lead times tend to increase accuracy, they also reduce the relevance of the results, as stakeholders generally prefer to have accurate predictions well in advance (Lewis-Beck and Stegmaier, 2014a; Jennings et al., 2020).

Determining the optimal lead time for a given model is not straightforward, as the relationship between lead time and accuracy is often non-linear (Jennings et al., 2020). There is likely a point where the trade-off between lead time and accuracy is optimal (Jennings et al., 2020). For example, Jennings et al. (2020) observed that their model achieved optimal accuracy at a lead time of 48 days prior to the election. However, with a slight sacrifice in accuracy, they could obtain almost equally accurate results almost three months before the election (Jennings et al., 2020). Thus, the optimal lead time depends on the specific model and context and should be compared with other alternatives under the same conditions.

Beyond accuracy and lead time, parsimony is another essential principle in evaluating election forecasts (Lewis-Beck, 2005). Parsimony refers to the simplicity of a model, which means that it explains the data with the fewest possible parameters without sacrificing predictive power (Lewis-Beck, 2005). In other words, less is more.

As the number of parameters grows, the flexibility of the model increases, potentially leading to overfitting (Hastie et al., 2009). Overfitting can decrease the

model's ability to generalize to new data, thus making reliable forecasting difficult. This challenge is closely related to the bias-variance tradeoff, a fundamental concept in machine learning and statistical modeling that explains the balance between the simplicity of a model and its complexity (Hastie et al., 2009).

Parsimony also improves model usability. Simpler models are easier to understand, interpret, and replicate, fostering greater trust in the forecasts provided (Lewis-Beck, 2005). Focusing on a few relevant variables rather than many questionable ones ensures that the model remains robust, interpretable, and practically useful (Lewis-Beck and Stegmaier, 2014a).

Finally, for the sake of scientific integrity, the forecast should be reproducible. For a forecast to be trustworthy, it must be reproducible by both the original author and other analysts (Lewis-Beck, 2005). Reproducibility can be compromised if the measures used are obscure, costly, or difficult to obtain. Such measures pose a barrier to independent verification and validation. Models should employ data and variables that are readily accessible and available well before the election event. Ensuring reproducibility not only validates the model's predictions, but also fosters greater trust and credibility in the forecasting process.

Nonetheless, while parsimonious and easily reproducible models are ideal, there are instances where complex models may offer significant benefits or scientific insight (Shmueli, 2010). In such cases, it is important to employ tools that improve the interpretability and transparency of these models. For example, many machine learning algorithms, such as random forests, provide mechanisms to analyze the importance of characteristics, helping to identify the variables that influence the most the predictions (Geron, 2019).

Tools such as SHAP (Shapley Additive exPlanations) that can be used to interpret model outputs across different types of machine learning algorithms (Lundberg, 2017). SHAP values provide a unified measure of the importance of the features, making models more interpretable.

4.8 Ethical Considerations in Elections Forecasting

Predicting election outcomes has become an important aspect of democratic processes worldwide, influencing both voter behavior and the electoral process itself (Victor, 2021). For instance, anticipated election outcomes can shape voters' choices and influence the manner in which candidates conduct their campaigns. However, over-reliance on forecasts may result in misinformation or reduce voter turnout if the outcome seems predetermined. Consequently, it is essential to employ rigorous methodologies in election forecasting to ensure accuracy and reliability (Victor, 2021).

The mechanisms through which election forecasts influence voter behavior are complex. The literature identifies several key pathways, such as strategic voting and

contagion effects (Blais et al., 2006). Strategic voting occurs when voters choose a candidate they perceive as more viable over their preferred choice, often to prevent an undesirable outcome. On the other hand, contagion effects arise when forecasts generate momentum for certain candidates, prompting others to support them based on perceived popularity.

Researchers must be cautious about what they release, ensuring that forecasts are presented transparently and responsibly (Victor, 2021). Transparency involves clearly disclosing factors that may affect the reliability and accuracy of the method. Responsibility entails not overstating the certainty of the forecasts and considering the potential impact on public attitudes and behaviors.

In recent years, there have been many instances where traditional methods have failed due to factors that were not considered. Forecasting is inherently uncertain, and various factors can undermine the accuracy of forecasts, including sampling errors, biases in data collection, insufficient sample sizes, and flawed models (Victor, 2021). Furthermore, external elements, such as sudden political events or changes in public sentiment, can affect reliability.

Some studies have suggested that during the 2020 US presidential elections, the combination of polarization and the COVID-19 pandemic increased the visibility of the forecasts while decreasing their reliability (Victor, 2021). These factors introduced biases that were difficult to quantify and address, potentially resulting in misleading predictions that could shape public perceptions and expectations. Similarly, in the context of the 2016 US presidential elections, some researchers proposed that the heightened focus on forecasts could have increased voter certainty about the election outcome, possibly leading to reduced voter turnout and affecting democratic participation (Victor, 2021).

The influence on voters can also be indirect. For example, forecasts can contribute to a media environment that prioritizes electoral competition over substantive issues, thus reducing the quality of public discourse (Victor, 2021). This focus can lead to a narrow scope in media coverage, resulting in less informed voters and potentially detracting from important policy discussions that require public attention.

Forecasts can also influence the trust towards science. When forecasts or their reporting overstate confidence in an estimate, they mislead the public about the role of science and engender skepticism about scientific and media reliability. It has been suggested that public distrust in science and media contributes to democratic decline in the US (Victor, 2021).

Despite these criticisms, some argue that forecasting has its benefits, such as generating increased interest in elections and voting. Although it is often difficult to determine whether forecasts drive voter interest or vice versa, interest in forecasting and voter mobilization can covary, making it challenging to establish a direct causal relationship (Victor, 2021).

Although Finnish election forecasting may not be as monetized as it is in the

US, where it has become a commercial commodity, political communicators globally must contextualize forecasts responsibly. This includes emphasizing the precise parameters of the forecast estimates, favoring vote share predictions over win probabilities, using data visualizations to highlight uncertainty, and providing a clear and honest context to evaluate the forecast limitations (Victor, 2021).

5 Research Approach

This study investigates the use of online data to predict the results of the Finnish parliamentary elections in 2015, 2019, and 2023. The primary focus is on understanding the potential of online data to provide predictions for real-world scenarios, with elections serving as a specific application of this concept. Through this exploration, the research aims to contribute to the broader discussion of the applications of digital data in enhancing predictive capabilities across various domains, highlighting both the opportunities and challenges associated with such approaches.

5.1 Research Questions

The overarching research question guiding this study was presented in the Introduction Section as follows:

"How accurately can publicly available online data predict election outcomes?"

Addressing this question in a comprehensive way is a challenge. Therefore, the exploration was conducted through a series of original research publications included in this thesis. Each publication explores specific phenomena from different perspectives. The specific research questions are presented in Table 3

The first article focuses on a retrospective assessment of the feasibility of the forecast. The retrospective approach allowed for a straightforward evaluation of how the forecasts aligned with real outcomes. After assessing the accuracy of the forecast, the study focused on analyzing the relationship between the number of likes on the public campaign pages and electoral success.

The second article explores research questions aligned with traditional quantitative frameworks, specifically examining the role of social media platforms such as Twitter and Facebook in election predictions. The study also covered how the personal attributes of the candidates could moderate these effects. The study tested the underlying theories of political behavior in the context of using publicly available data in forecasting.

The third article represents the culmination of this research, combining the insights of previous studies into a more refined investigation. This study improved existing forecasting models by applying machine learning techniques to predict individual vote shares in the 2023 Finnish parliament elections.

Table 3. Research Questions

Article I:	Was it possible to forecast the results of Finland's 2015 parliamentary elections using candidates' Facebook page Likes?
Article II:	What is the role of social media platforms, such as Twitter and Facebook, in predicting election outcomes?
	How do candidates' personal attributes moderate the role of social media platforms, such as Twitter and Facebook, in predicting election outcomes?
Article III:	How accurately can a model, trained on publicly available online data, predict the number of votes received by candidates?

5.2 Philosophical Considerations

"All models are wrong, but some are useful."

– Attributed to George Box (1919-2013), a British statistician

Statistical modeling requires simplification, which results in the loss of details in order to gain insights. This is a common challenge in any forecasting effort, where simplicity is often preferred over complexity for practical reasons (Lewis-Beck, 2005). Philosophically, this reflects both the limitations and the utility of predictive modeling.

Despite their limitations, statistical models are crucial for understanding and progressing in science (Box, 1976). The iterative process of refining models through theory and practice allows them to remain relevant as new data emerge, fostering a critical approach that seeks perpetual improvement.

Recognizing the limitations, but focusing on the practical utility of models, aligns closely with a pragmatic view of scientific modeling. The pragmatic view of scientific modeling emphasizes the practical effectiveness and utility of models in decision-making and focuses on results and applications in the real world (Legg and Hookway, 2024). This approach encourages the selection of models based on their effectiveness in solving problems, balancing theoretical rigor with practical applicability. Pragmatism also aligns with the epistemic standards commonly applied in data science, where the utility of new knowledge for decision-making is judged by its predictive power rather than its ability to explain past events (Dhar, 2013).

Complementarily, scientific realism offers an additional layer of understanding by positing that models, while imperfect, are instrumental in uncovering insights

about underlying realities (Chakravartty, 2017). This perspective supports the use of quantitative methods such as statistical analysis and machine learning to model and predict phenomena, recognizing the iterative process as crucial to refining our understanding of the structures of the world (Chakravartty, 2017). Although aiming to describe an objective reality, scientific realism acknowledges inherent uncertainties and remains open to model revision.

Together, pragmatism and scientific realism provide complementary epistemological perspectives. Pragmatism values the practical utility of models, focusing on their effectiveness and applicability in decision-making and real-world applications (Legg and Hookway, 2024). It supports selecting models based on their accuracy, thus balancing theoretical rigor with practical needs. Scientific realism, on the other hand, acknowledges that although models might not perfectly represent reality, they are useful in uncovering insights about the world (Chakravartty, 2017).

Although pragmatism and scientific realism offer valuable frameworks to understand the utility and truth-seeking aspects of models, they generally do not account for the subjective role researchers play in modeling processes (Fuchs, 2017). Researchers make numerous decisions when defining project goals and selecting appropriate features, which greatly affects results (Barocas and Selbst, 2016). These human-induced choices can introduce biases, undermine objectivity, and lead to inconsistencies, potentially compromising the reliability and validity of the research outcomes (Barocas and Selbst, 2016; Lipton, 2018).

Recognizing these subjective influences highlights the importance of reflection and transparency in scientific inquiry (Mingers et al., 2013). By being critically aware of the assumptions and choices made during the research process, researchers can actively work to mitigate bias and enhance the robustness of their findings. This critical perspective complements the philosophical considerations of pragmatism and scientific realism by advocating for methodological rigor and openness.

These philosophical concepts fit well within the field of IS, as there is strong advocacy for a pluralistic approach to research methodologies, supporting different ontological and epistemological stances (Orlikowski and Baroudi, 1991; Gregor, 2006).

5.3 Research Approach

The research focuses on developing a forecasting model that uses publicly available online data to predict electoral outcomes. These data serve as the model's input, enabling it to learn patterns and relationships relevant to predicting election results. For each parliamentary election studied, a forecast was published on the University of Turku website and in local newspapers prior to the election. Publishing the results before elections allowed for the validation of the model's performance under actual conditions and increased the visibility of the research.

The selected research approach is predictive modeling, a type of quantitative research aimed at building models to predict future events (Kuhn and Johnson, 2013; Shmueli and Koppius, 2011; Shmueli, 2010). Predictive modeling is closely associated with data science and data mining (Dhar, 2013).

A quantitative research approach fits well within the scope of IS. Quantitative research is very popular within the discipline and has played an important role in empirical research over the years (Chen and Hirschheim, 2004; Orlikowski and Baroudi, 1991; Sarker et al., 2013). Traditionally, quantitative research has focused on causal-explanatory statistical modeling (Shmueli and Koppius, 2011; Creswell and Creswell, 2018). In a typical quantitative study, researchers test hypotheses, which are predictions about the expected outcomes and relationships among variables. These hypotheses provide numerical estimates of population values based on data collected from samples (Creswell and Creswell, 2018).

In recent years, technological advances have generated significant interest in data science (Abbasi et al., 2023). These advances have expanded the potential for data-driven research, enabling more sophisticated analyses and insights. Unlike traditional methods that often begin with a predefined hypothesis, data-driven approaches can be more exploratory or inductive, uncovering patterns and relationships directly from the data itself (Maass et al., 2018). Some authors describe data science as the fourth paradigm of scientific discovery, underscoring its transformative role in modern research (Abbasi et al., 2023).

Predictive modeling can be defined as the process of developing a mathematical tool or model that generates accurate predictions (Kuhn and Johnson, 2013). Methodological frameworks designed primarily for predictive analytics provide structured approaches to solving data-related problems (Shmueli and Koppius, 2011). Among these, industry standards such as the CRISP-DM framework are often used to guide data science projects (Wirth and Hipp, 2000). These frameworks offer relevant guidance because predictive analytics shares similarities with data science. However, since these frameworks are designed primarily for data science rather than predictive modeling, there may be some incompatibilities. In addition, they can be quite rigid when applied to academic research and are generally intended for projects that involve multiple stakeholders (Martínez-Plumed et al., 2019).

Therefore, this research does not follow any strict framework directly. Instead, it adopts a flexible approach, allowing the incorporation of relevant elements from existing methodological frameworks and industry standards, while tailoring them to better suit the specific requirements of predictive modeling. This adaptive methodology ensures that the research can address the unique challenges associated with forecasting electoral outcomes using publicly available data. By not adhering to a rigid framework, the research can remain agile and responsive to the dynamic nature of electoral data and the evolving landscape of data science.

The theoretical foundation of this research can be examined from two perspec-

tives. Primarily, it is grounded in a broad theory of forecasting, which posits that present and past information can be systematically analyzed to predict future events (Petropoulos et al., 2022). This outlook is further supported by concepts from economics and behavioral psychology, such as prospect theory (Kahneman and Tversky, 1979) and expected utility theory (Von Neumann and Morgenstern, 1944), which extend this idea to the predictability of human behavior.

On a practical level, concepts and theories related to voter decision-making are instrumental in guiding feature selection and understanding the relationships between past behaviors and future events. Creating a data science model for forecasting election outcomes using online data requires a thorough understanding of the domain, which includes theories of voting behavior and existing forecasting methodologies. These theories seek to explain how sociopolitical, personal, economic, and media-related factors influence voter choices, as detailed in Chapter 4. They shed light on aspects of voter preferences and behavior, which are crucial for developing an accurate forecasting model. Using these specific theories, this research adheres to the need for a well-defined theoretical framework, ensuring that the model achieves practical objectives and contributes to theoretical discussions in the field.

1. **Theories of Voter Decision-Making**: Theories related to voter behavior provide insight into how sociopolitical, personal, economic, and media-related factors influence electoral outcomes. These theories help guide the model's structure and the interpretation of its predictions.

2. **Statistical Methods**: A solid grasp of statistical principles is essential to discern what is statistically feasible and to ensure the integrity of the predictive capabilities of the model. This includes understanding the limitations and potential biases inherent in the data and the selected methodologies.

3. **Methodological Frameworks**: Although not adopted in detail, CRISP-DM offers a valuable overview of the phases involved in the handling of data science projects. CRISP-DM helps outline the necessary steps and considerations for this research, from data collection to the deployment of predictive models.

4. **Software Development**: A clear understanding of software engineering principles is required to effectively design and implement the model. This involves coding and algorithm design to ensure that the software developed meets the project requirements effectively.

5. **Data Collection Methods**: Proficient knowledge of how and from where online data can be collected is necessary. This includes understanding APIs, web scraping techniques, and data curation practices, fostering the data used is reliable, comprehensive, and ethically sourced.

5.4 Validation and Evaluation

Validation and evaluation play an important role in any scientific study (Creswell and Creswell, 2018). Quantitative research is typically evaluated through statistical inference and is closely related to understanding the methods (Creswell and Creswell, 2018). Different methods require different evaluation metrics (Kuhn and Johnson, 2013). The literature on predictive analytics and data science emphasizes the evaluation of the quality of predictions in practice (Shmueli and Koppius, 2011; Dhar, 2013).

Typical forecasting models are initially developed by comparing them with historical election data, allowing researchers to assess predictive precision and refine model parameters (Lewis-Beck, 2005). Effective evaluation integrates both quantitative accuracy and qualitative consistency with theoretical paradigms, ensuring that models provide robust and reliable predictions that inform practical electoral strategies (Lewis-Beck, 2005).

In the included studies, the accuracy of the election forecast models is assessed using various techniques, including classification metrics, regression metrics, and ranked correlations. All three forecasts present classification results to facilitate a high-level understanding of precision. Furthermore, all forecasts evaluate the mean absolute error (MAE) at the national level compared to the actual election results.

Article III enhances the evaluation metrics by expanding beyond general comparisons between election outcomes and predicted results. It includes additional classification metrics such as precision and confusion matrices to better analyze the predictions of elected candidates. Furthermore, Article III enriches the evaluation process by incorporating the root mean square error (RMSE) and the statistic R^2, which provide a deeper insight into the prediction precision and the fit of the model. Bias is also examined to detect any systematic prediction errors.

Although statistical measures of error such as RMSE or MAE are necessary, discussing the narrative and theoretical consistency of a model is equally important (Lewis-Beck, 2005). This ensures that the model is not merely a statistical artifact but also aligns with political theories and realities (Lewis-Beck, 2005). In Article III, Shapley additive explanations (SHAP) are utilized to evaluate how well the model aligns with theoretical constructs from political science, providing transparency by illustrating how features contribute to predictions. This approach enhances both global and local interpretability, highlights potential biases, and improves understanding of the model's internal processes.

The literature on election forecasting stresses the importance of testing models in real-world conditions to enhance their rigor and reduce the risk of overfitting (Lewis-Beck, 2005). This is achieved in the present research by consistently publishing findings before elections occur. This approach aligns with the broader call for reflection and transparency in scientific inquiry to recognize and account for sub-

jective influences (Mingers et al., 2013).

In the context of electoral prediction, the predictive performance is also evaluated by the robustness and reliability of the model in different electoral contexts and cycles (Lewis-Beck, 2005). This study presents three election cycles, each using a slightly varied version of the model. However, the presented approach has not been tested in other electoral contexts.

In line with the pragmatic perspective, many authors emphasize that models should be evaluated based on their practical utility (Lewis-Beck, 2005; Dhar, 2013; Shmueli and Koppius, 2011). This includes considering the model's interpretability and its communicability to stakeholders such as political analysts and campaign strategists. In the electoral context, practical utility is related to two key factors: lead time and accuracy (Jennings et al., 2020; Lewis-Beck, 2005). The study transparently reports the lead time and the accuracy is presented in all forecasts. The practical utility heavily depends on how the method compares with other state-of-the-art approaches. The forecast for the 2023 parliamentary elections includes a comparison with the results of the polls.

It should be noted that the model is not directly comparable to other state-of-the-art election forecasting frameworks *as is*. As such, identifying suitable metrics that accurately capture the performance and usefulness of the model on multiple levels is essential. The selected metrics reflect the unique attributes and capabilities of the model while allowing a meaningful evaluation of its forecast accuracy compared to actual election outcomes.

In addition to scientific evaluation metrics, ethical considerations are also important when assessing research. Forecasting models can influence democratic processes and might unexpectedly affect election outcomes (Victor, 2021). Therefore, if the concept presented is applied in other contexts, these ethical implications must be evaluated on a case-by-case basis to ensure responsible and fair use.

6 The Journey

6.1 Development Process Overview

This chapter synthesizes the findings from the three research articles which collectively form the iterative design process. While each article employs distinct methods suited to its specific research questions, the collective insights from these studies together enhance our understanding of predictive accuracy and methodological robustness in election forecasting. By engaging with diverse data sources and analytical techniques, these articles contribute to building a comprehensive understanding of election forecasting.

The project can be conceptualized as three distinctive development cycles, each related to an election corresponding to the Finnish Parliamentary Elections of 2015, 2019, and 2023. Each cycle consisted of two key steps:

1. **Forecast**: Collecting data, creating the forecast, and publishing the results.
2. **Analysis**: Analyze the data and the forecast to improve the model.

Each development cycle contributes towards the ability to address the challenge of electoral predictions by combining practical application with theoretical exploration. The three encompassed research articles *I, II, and III* provide more detailed information about the analysis phase.

The first article, *"Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections"*, examines the predictive capability of social media interactions, specifically Facebook likes, and their correlation with electoral success. It revealed a positive correlation but suggests that these metrics are less reliable than traditional baseline models, indicating the nuanced nature of social media data in forecasting contexts.

The second article, *"The Role of Social Media Platforms in Forecasting Elections: A comparison of Twitter and Facebook"*, highlights the relative strengths of different social media platforms in electoral prediction. It particularly emphasizes the higher forecasting power of Facebook likes over Twitter followers when predicting electoral outcomes for less experienced candidates, thus revealing the need for a multifaceted approach to social media data that includes candidate attributes.

The third article, *"Predicting Candidate Votes in Multiparty Elections"*, demonstrates the efficacy of integrating diverse data types beyond social media for more ac-

curate electoral forecasts. This study achieved the highest accuracy of the presented studies, which aligns closely with the results of independent polls. The results underscore the potential value of data integration in the refinement of prediction models.

The Table 4 summarizes the findings, illustrating the diverse methodologies and implications for election forecasting. By integrating insights from these studies, this thesis aims to contribute to the development of more robust and comprehensive electoral forecasting models, leveraging both social media interactions and other heterogeneous data types.

Table 4. Summary of Research Articles

Research Article	I	II	III
Title	Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections	The Role of Social Media Platforms in Forecasting Elections: A Comparison of Twitter and Facebook	Predicting Candidate Votes in Multiparty Elections
Findings	Facebook likes showed a positive correlation with electoral success, though less predictive than simple baseline models.	Facebook likes outperformed Twitter followers as election success indicators, especially for less experienced candidates. Highlights the importance of combining social media data with candidate attributes for better forecasts.	Achieved high accuracy and outperformed baseline forecasts. Had a closer alignment with independent polls.
Implications	Acts as a significant but weak predictor and highlights the need for using data from multiple platforms.	Shows potential when integrating platforms and underscores the limitations of relying on single-source data.	Suggests the benefit of using diverse data types, pointing towards enhanced precision with comprehensive data integration.

6.2 Studying Social Media in Electoral Forecasting

6.2.1 Facebook and Public Opinion

The first model was very simplistic. The model relied solely on Facebook likes for each candidate, operating under the assumption that one like equaled one vote.

Initial modeling efforts were conducted using the R programming language. For the 2015 election, these likes were converted into vote shares using the D'Hondt method, which is commonly used to allocate seats proportionally. However, this approach led to suboptimal performance.

The 2015 forecast indicated that relying solely on Facebook likes as a proxy for votes does not provide accurate predictions. However, there is a clear linear relationship, suggesting that while likes alone are insufficient for precise forecasts, they can still offer valuable insights when used alongside other data sources.

The findings of the first research articles suggest that by 2015, Facebook had become an integral component of electoral campaigning in Finland. Despite this integration, a forecast based solely on the number of likes on a candidate's campaign page was not very accurate, having worse accuracy even compared to a simple baseline model that assumed all incumbent candidates seeking re-election would be elected again.

Nevertheless, the research identified a positive correlation between the number of Facebook likes and individual success in elections, with candidates receiving more likes being significantly more likely to be elected. These insights lead to the conclusion that there is a notable relationship between Facebook likes and election outcomes, indicating that while Facebook metrics may not serve as reliable standalone predictors, social media engagement remains a valuable asset in a candidate's overall campaign strategy.

6.2.2 Data Collection and Preparation

The first iteration relied only on Facebook data. Collecting it involved a substantial amount of manual labor. Page identifiers for each candidate's campaign page were collected manually first and then used to pull the relevant data from the Facebook Graph API. This involved querying Facebook's social graph to collect data about public campaign pages and their likes. Data on candidates before elections was downloaded from the election website maintained by the Ministry of Justice. The files were downloaded as CSV files and processed using various software tools.

The data were collected approximately one month before each election, in alignment with the official release of the candidate lists which were not available earlier. The data was also collected multiple times during the last month of the race. There was a delay of a few days after the candidates were released as it took some time to collect the page ids. However, the data collection scripts were prepared in advance to allow for quick iterations.

The data collected underwent manual validation to ensure its accuracy and relevance. Although candidate identities were known, there were instances in which incorrect Facebook campaign pages were linked to candidates.

During this initial phase, ethical considerations surrounding the retrieval and use

of data were not addressed. However, after evaluating these ethical dimensions retrospectively, no significant issues were identified.

6.2.3 Finnish Parliamentary Elections 2015

The 2015 parliamentary election in Finland marked a triumph for the Centre Party, the Green League, and the Swedish People's Party (Nurmi and Nurmi, 2015). In contrast, left-leaning parties such as the Social Democratic Party and the Left Alliance, along with the National Coalition Party, the ruling party from the 2011 election, faced setbacks. The campaigns were notably cautious, reflecting the poor economic conditions at the end of 2014, characterized by stagnant growth and an unemployment rate that exceeded 9%. Of the 200 parliamentarians elected in 2015, 73 were newcomers, indicating a smaller turnover compared to the 84 new seats in the 2011 election. Among the 73 departing members of the parliament, 40 stood for re-election but were unsuccessful, while 33 opted not to contest.

6.2.4 Forecast Results

In anticipation of the elections, a forecast based on the number of Facebook likes on each candidate's page was published on the university website. Figure 4 is a copy of the article. It projected victories for the National Coalition, the Green League, and the Left Alliance. However, it overestimated support for these parties, while underestimating the popularity of the Finns Party and the Centre Party. The prediction aligned closely with the actual results for parties such as the Social Democratic Party, Christian Democrats, and the Swedish People's Party.

The forecast accurately predicted the election of 86 candidates. Of the total of 1,226 candidates assessed, there were 86 correct predictions, 114 false positives, 857 true negatives, and 69 false negatives. 45 elected candidates were not part of the prediction due to the absence of Facebook campaigns. When combined with the 69 false negatives, the figure reaches 114, as candidates not included in the forecast were automatically assumed not to be elected.

The actual and predicted totals are presented in Table 5. Further methodological details, data analysis, and insights can be found in the underlying research article I.

6.3 Building Theoretical Foundation

6.3.1 Adjusted Simple Model

The second cycle was characterized by diving deeper into political science, which was hoped to contribute towards a model which could deal with the biased nature of social media data. The results of the first iteration clearly indicated that the social

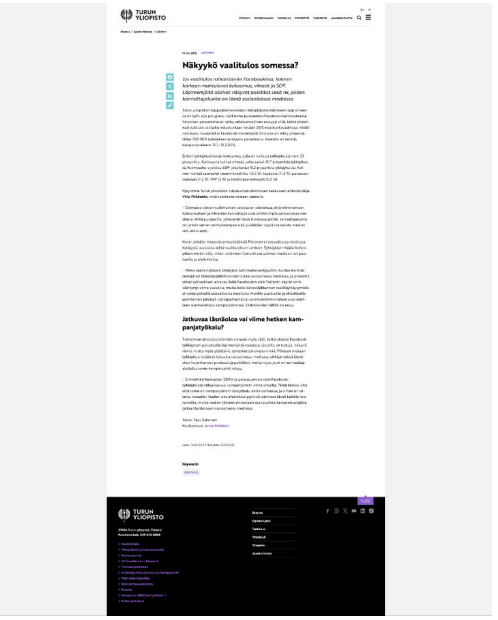

Figure 4. 2015 Election Forecast.

Table 5. Comparison of Actual and Predicted Vote Shares for 2015

Party	Election Result	Predicted Result
National Coalition Party	18.2%	23.1%
Green League	8.5%	19.5%
Social Democratic Party	16.5%	15.2%
Left Alliance	7.1%	13.3%
Centre Party	21.1%	11.1%
Finns Party	17.7%	7.1%
Swedish People's Party	4.9%	5.3%
Christian Democrats	3.5%	5.0%
Others	2.5%	0.3%
MAE		4.81

media data alone does not provide practically relevant results.

Due to time constraints, the second iteration of the forecast could only have minor adjustments. In the 2019 forecast, the model was adjusted to use the raw number of Facebook likes without converting them through the d'Hondt method. This change aimed to maintain simplicity while slightly improving performance, although it highlighted the need for more sophisticated modeling techniques that could integrate diverse data sources and account for complex relationships between variables. This learning paved the way for the adoption of more advanced machine learning models in subsequent forecasting efforts.

The findings of the second article explore the role of multiple social media platforms in forecasting electoral outcomes, specifically focusing on the Finnish parliamentary elections of 2019. The study underscores the potential of both Facebook likes and Twitter followers as predictors of election success, demonstrating that both metrics are positively associated with electoral outcomes.

Facebook emerged as a stronger indicator compared to Twitter. Upon examining the attributes of the candidates, the study found that political experience moderates the impact of social media engagement on election outcomes. It was revealed that social media metrics, particularly Facebook likes, are more influential for candidates with less political experience, suggesting that social media serves as a critical channel for novice politicians to reach a broader audience and boost their election chances. In contrast, seasoned politicians, with established media presence and supporter bases, may find social media having a relatively weaker impact on their electoral success.

The research highlights the importance of integrating social media data with personal attributes of the candidate, particularly political experience. The most effective predictive model combined social media engagement metrics and candidate political experience, capturing the interacting effects of these variables. This approach is in line with earlier studies that demonstrated a positive correlation between social media metrics and election success. The study's insights emphasize the need for election forecasting frameworks to incorporate both social media data and candidate attributes, offering a nuanced understanding of voter behavior in the digital age.

6.3.2 Data Collection and Preparation

During the second cycle, the data collection process was significantly expanded to incorporate additional data sources beyond Facebook. This included Twitter data from the 2015 and 2019 elections, as well as data from voting advice applications. These new sources were integrated to provide a detailed view of social media engagement and its potential impact on electoral outcomes.

In addition, a list of candidates and relevant details was compiled from Statistics Finland, which included a range of demographic and political information. This data set encompassed the sex, age, occupation, and party affiliation of candidates, as

well as their electoral district, incumbency status, past membership in the European Parliament, previous service in the Finnish parliament, participation in the last parliamentary elections, and any history of election to municipal government positions. By incorporating both social media and demographic data, the analysis was able to offer a more nuanced understanding of the electoral dynamics.

This iteration of data collection was partially automated, using the links to Twitter and Facebook pages provided by the voting advice application data. This automation reduced the need for manual data collection, improving both the efficiency and accuracy of the process.

The initial data collection was performed 10 days before the elections took place. The timing was justified because according to the 2015 study, the number of Facebook likes and Twitter followers is relatively static, not subject to rapid fluctuations.

Inconsistencies in the number of Facebook likes and Twitter followers, as well as votes received in previous elections, served as flags to identify outliers. These inconsistencies enabled the identification and correction of errors, thus ensuring that the data accurately corresponded to the correct individuals.

Details related to data collection and preparation are presented in Article II.

Twitter

Twitter data for the 2019 elections, specifically follower counts, was collected using the Twitter API. Twitter accounts were found for 1,466 candidates.

Collecting Twitter data involved acquiring the Twitter user ID, commonly known as the ”handle,” for each candidate. This information was primarily gathered manually, although a significant portion of candidates had included their Twitter handle in their responses to voting advice applications. Regardless of the source, manual verification of each entry was necessary to ensure accuracy.

Twitter follower counts for the 2015 elections were also collected during the second cycle. A challenge encountered here was the difficulty in obtaining historical Twitter data from 2015. Fortunately, existing data from another researcher partially addressed this issue, although it was limited to only the 50 candidates with the most followers.

Facebook

In 2019, the Facebook data collection method transitioned from using the API to web-scraping due to the new restrictions and policy changes imposed by Facebook (Bruns, 2019). The collection of campaign page identifiers for candidates was achieved through a combination of manual efforts and data gathered from candidates responses to voting advice application.

Once the identifiers were collected, web-scraping scripts were used to directly

extract public data from the candidates' campaign pages. This process involved accessing only publicly available data and was conducted without the need to log into Facebook accounts, ensuring compliance with data privacy regulations and terms of service. Facebook page was found for 1,502 candidates.

Voting Advice Applications

The data from voting advice applications were sourced from Yle, the national broadcasting company of Finland. The data were retrieved after the elections, as they were not used for the published forecast. The data were readily available as open data on open data platform of Yle. Other historical voting advice application data was retrieved from the same source.

6.3.3 Finnish Parliamentary Elections 2019

The 2019 general election in Finland was notable for its tightly contested race among the Social Democrats, the Finns Party, and the National Coalition, none of which exceeded 18 percent of the vote (Borg, 2019). The Social Democrats secured 40 seats, becoming the largest party and gaining six more seats than in 2015, while the Finns Party and National Coalition gained 39 and 38 seats, respectively. However, the Center Party experienced a significant decline, falling below 15 percent support for the first time since 2011. Climate change and immigration dominated the campaign, benefiting the Greens, who gained five additional seats. Of the 200 parliamentarians elected, 74 were newcomers, including 33 first-time candidates, while 48 incumbents were not re-elected.

The election campaign was dominated by issues of climate change and immigration, while discussions on the economy and employment were less prominent compared to previous elections (Borg, 2019). Climate change was a key focus in candidates' campaigns, contributing to the Greens' best-ever showing in Finnish parliamentary elections. Climate-related issues also benefited the Left Alliance, which campaigned on a similar 'green agenda' as the Greens. As in previous elections, support levels for the Swedish People's Party and the Christian Democrats remained relatively stable.

6.3.4 Forecast Results

The forecast model used was again based on Facebook likes, similar to the model used in 2015. It ranked candidates and predicted the 200 most likely to be elected. Out of the 2,468 candidates, the model correctly predicted the outcome for 105 candidates, surpassing the performance of the 2015 forecast, which accurately predicted 92 candidates. Nationally, MAE improved from 4.81 in 2015 to 3.21 in 2019, indi-

cating an increase in accuracy. Table 6 presents a summary of predicted versus actual vote shares.

Table 6. Comparison of Actual and Predicted Vote Shares for 2019

Party	Election Result	Predicted Result
National Coalition Party	17.0%	18.3%
Green League	11.5%	18.4%
Social Democratic Party	17.7%	13.3%
Left Alliance	8.2%	15.9%
Centre Party	13.8%	10.7%
Finns Party	17.5%	11.5%
Swedish People's Party	4.5%	4.0%
Christian Democrats	3.9%	3.2%
Others	5.0%	3.7%
Blue Reform	1.0%	1.1%
MAE		3.21

An article discussing the forecast was published on the University of Turku website. Figure 5 illustrates the published article.

These results suggest that the relevance of Facebook likes as a metric for electoral forecasting has improved since the 2015 elections. However, relying solely on Facebook likes is not sufficient to predict election outcomes. To figure the path forward, it was necessary to integrate social media data with other metrics. Several potential sources of investigation were identified, such as the use of economic indicators as predictors of electoral success (Lewis-Beck and Paldam, 2000). Combining insights from political science, forecasting theory, and experimental data science could yield better results.

During this period, Facebook also introduced changes to its data access policies. The platform's policy updates have imposed limitations on data retrieval, making it more challenging to gather similar data.

A research article detailing the forecast for the 2019 elections was not published, as there was little addition to article I. Article II depends on the 2019 election data, but focuses more on the interaction of different features.

6.3.5 Implications

The key takeaway from the findings was that there are several features that can be added to enhance the predictive power of election forecasting models. By including multiple factors, such as social media engagement metrics, campaign budgets, candidates' political experience, and traditional metrics like party affiliation, models can achieve more accurate predictions of electoral outcomes.

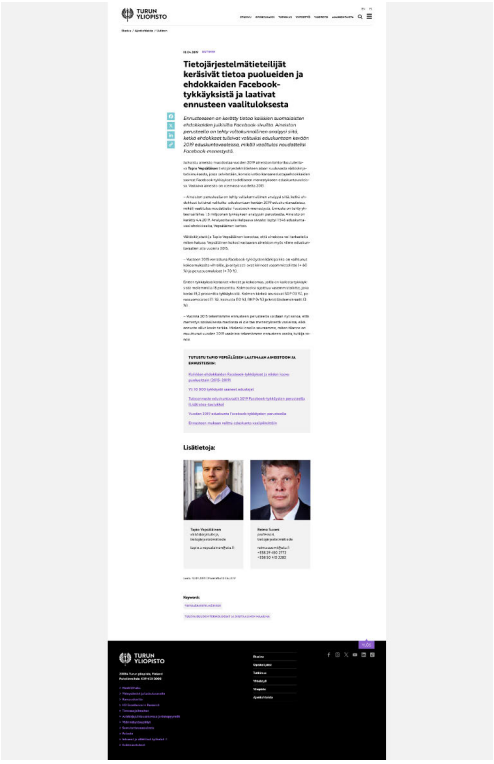

Figure 5. 2019 Election Forecast

This suggests that forecasting models should adopt an integrated approach, leveraging the diverse range of available data to capture the different dimensions of voter behavior and candidate appeal. The ability to include and assess the relative importance of various features highlights their combined potential in improving the precision of election forecasts.

6.4 Developing a Model to Predict Candidate Success

6.4.1 Predicting Candidate Votes in Multiparty Elections

The third iteration incorporated lessons from earlier models and addressed newly identified research gaps. With the introduction of machine learning techniques and data science, the project was better informed about standard workflows in data science, such as the CRISP-DM framework, which served as a reference point. Identifying research gaps and assessing the shortcomings of current methods was essentially serving as the business understanding phase.

Research gaps were primarily assessed from two streams of research: the study of using social media to predict real-world events and the field of electoral forecasting.

In studies focused on using social media to predict elections, one proposed strategy to deal with low data quality was the implementation of multi-source data integration, which involves combining information from various sources (Skoric et al., 2020). Another suggested method was the use of machine learning methods, which was proposed as a way to mitigate the noise and bias in the data from social media (Brito et al., 2021; Gayo-Avello, 2013). Furthermore, a need for more theoretically informed work paying attention to theories related to citizen decision making was also identified (Skoric et al., 2020). As the work progressed, the focus shifted away from social media, and more towards the domain theory related to elections.

A new research gap in the field of election forecasting was also identified. Most of the existing methods for forecasting elections depend on polling or economic data (Lewis-Beck, 2005). The typical approaches were designed to make predictions for races with only a few candidates, such as presidential races. The structural methods are used to predict the winning party, revealing little about the standings of other parties (Nadeau and Lewis-Beck, 2020).

Together, these details clarified the limitations of existing methods and highlighted the need for innovation in forecasting approaches. Subsequently, the candidate-based approach was proposed as a new methodological approach suited for multi-party elections.

Adopting a candidate-focused perspective in forecasting is also supported by theoretical foundations in the case of elections with a large number of candidates. As detailed earlier, voters often rely on candidate-specific heuristics, such as political experience, past success, and personal appeal (Lau and Redlawsk, 2006; Zittel, 2016).

6.4.2 Data Collection

Facebook and Twitter

As the 2023 elections approached, it became necessary to redesign the data collection algorithm due to disruptions in existing methods. The original web-scraping scripts for Facebook had stopped functioning, and the Twitter API was no longer accessible. Additionally, Twitter's acquisition introduced uncertainty about the future of the platform.

To address these challenges, the research explored using search engine indexing as an alternative source of public data. Both Google and Bing APIs were considered, and the Bing API was ultimately chosen because of its higher volume availability in the free tier.

The Bing API was used to retrieve data for both Facebook and Twitter. For Twitter, search queries incorporated candidate first names, last names, party affiliations, and Twitter user IDs to gather page details, specifically focusing on followers. Regular expressions were used to parse the Bing search response and extract follower

counts by matching the term ”followers.”

Similarly, for Facebook, the search queries included the candidates' first names, last names, party affiliations, and Facebook page IDs to identify relevant page details. The retrieval process involved parsing the search results with regular expressions to match the term ”likes” and extract the corresponding numbers.

This data collection was conducted about one month before each election to coincide with the official release of candidate lists, as they were unavailable beforehand.

Manual data validation was performed to improve accuracy. Although the candidates' identities were known, there were instances in which the wrong Twitter account was matched with a candidate. To identify disparities, significant inconsistencies in the number of followers on Facebook, Twitter, and the votes received in previous elections were used as flags. These inconsistencies helped pinpoint errors, allowing for re-evaluation and correction. This approach ensured that the data corresponded to the correct individuals, thus maintaining the integrity of the data set.

Official Government Data

The forecast model on the third cycle was designed to include a wider range of features, including data from official government sources.

Data on candidates before the elections was downloaded from the election website maintained by the Ministry of Justice. Historical election data was obtained from Statistics Finland, an agency under the Ministry of Finance, which provides statistical databases on their website. The files were downloaded as CSV files and processed using various software tools.

The model used demographic information on candidates, past electoral performance in different types of elections, and past positions in municipal, regional, European Parliament, and presidential offices to assess candidates' political experience. These data points were used as features in the model. The information considered includes:

- Demographic information such as age and sex for the 2015, 2019, and 2023 elections.
- Past votes in parliamentary elections for the years 2011, 2015, 2019, and 2023.
- Past votes in municipal elections from 2012, 2017, and 2021.
- Past votes in European Parliament elections for the years 2014 and 2019.
- Past votes in presidential Elections for 2012 and 2018.

To integrate these data points with each of the parliamentary candidates, the match was made based on name, party affiliation, and region. The matching process for the European Parliament and presidential elections was performed manually,

while the municipal elections required several iterations and automation to achieve alignment and ensure accuracy.

Voting Advice Applications

Voting Advice Application was accessible through a publicly exposed API before the elections on the Yle website. However, data being publicly available does not mean that they can be used, which requires considering ethical questions about privacy and data ownership (Salganik, 2019). The ethical considerations involved in retrieving and utilizing publicly available data should be addressed transparently. It's essential to distinguish between data that is available online and data that actively seeks public attention (Boyd and Crawford, 2012). The following points were taken into account:

1. The data is openly accessible and appear to invite public engagement.
2. The data is analyzed at an aggregate level, ensuring that no individual names are disclosed.
3. The use of the data does not appear to cause any apparent harm to any of the candidates.
4. The data is used solely for research purposes, with no commercial gains involved.

Regarding data ownership, although this information was previously made publicly available, it was not openly distributed in 2023. The suspected reason is that the data contain sensitive personally identifiable information. Therefore, given that these privacy concerns have been thoroughly evaluated, using the data for research seems justified.

6.4.3 Data Preparation

The CRISP-DM process describes five generic steps in data preparation: select data, clean data, construct data, integrate data, and format data (Wirth and Hipp, 2000). Selecting data identifies relevant sources and attributes, while cleaning data addresses errors and missing values. The construction of data generates features to improve the predictive capacity. Integrating data merges information from multiple sources, and formatting data ensures readiness for analysis. Data selection was covered in the last section.

Data cleaning addressed issues related to missing values among many candidates, notably in models reliant on Facebook data for the 2015 and 2019 elections. The final model also encountered missing data because many candidates lacked records of past election participation, campaign budgets, or public social media profiles. Removing

these rows was not feasible, as it would skew the representation of candidates' election probabilities.

The construction of data required various pre-processing actions to ensure compatibility with statistical models. The handling of missing values and the encoding of categorical variables were key focuses. Imputation techniques were utilized where appropriate to address missing data thoughtfully, in order to minimize bias. Categorical variables, including various candidate attributes, were transformed into the numerical formats necessary for statistical analysis.

The integration of data helped to consolidate the full spectrum of features for the final model. To address incomplete past election vote counts, missing values were replaced with zeros to maintain consistency. This was supported by a binary indicator column to specify whether candidates had participated in those elections, preserving relevant participation information.

Formatting data involved the conversion of the textual campaign budget data to an ordinal scale. Missing budget values were imputed based on averages for parties and districts, enhancing the completeness of the data set. Political experience was quantified using a composite score derived from multiple aspects of political involvement, facilitating comparisons among candidates. The effects of the ballot position were adjusted using min-max scaling to ensure comparability between districts.

These pre-processing techniques were applied to refine the dataset, thus informing the model's development process. Further details and results of these approaches are elaborated in the full articles.

6.4.4 Modeling

In the context of CRISP-DM, modeling is recognized as a fundamental phase within data science projects (Wirth and Hipp, 2000). CRISP-DM identifies four general tasks that are integral to the modeling phase (Wirth and Hipp, 2000). The first task involves selecting the appropriate modeling technique, which is essential for addressing the problem at hand. Following this, the process involves generating a test design that outlines how the model will be evaluated. The subsequent step is to build the model, during which the selected technique is applied to develop the structure of the model. Finally, the phase concludes with the assessment of the model to evaluate its performance and suitability in relation to the defined objectives.

The outline of the model developed in the third iteration is presented in Figure 6. The forecast is created by first predicting votes for each candidate using a regression model. This data is then converted to proportional votes according to the d'Hondt method. The predicted proportional vote shares are used to forecast election outcomes, showing which candidates are likely to win seats in their respective districts. A more detailed description of the model can be found in article III.

The modeling was mostly done with Python and Jupyter Notebooks. Both tools

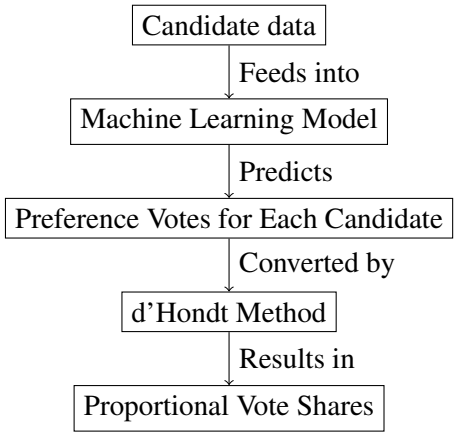

Figure 6. Forecast Structure

have a central role in modern data science practices. Jupyter Notebooks is an interactive platform for coding, visualizations, and annotations, allowing for integration of documentation and computational analysis (Geron, 2019).

The model design was an iterative process. Building on earlier understanding, several new features were incorporated into the model and a comparison of several machine learning algorithms was commenced. Initially, classification algorithms were considered to be an easy approach. However, as the comparison of modeling techniques was conducted, it was found that regression methods outperformed classification algorithms due to their ability to adjust outcomes based on the d'Hondt method.

After selecting the design, a further comparison of the regression algorithms was commenced. The predictive accuracy of the models was evaluated by training the model with data from the 2015 elections and testing against the results of the 2019 elections. The algorithms assessed included Random Forest regressor, Extreme Gradient Boosting regressor, Linear Regression, Ridge regression, Multilayer Perceptron regressor, Linear Support Vector regression, and K-Nearest-Neighbors regressor. A variation of linear regression with log transformation of skewed variables was also explored. This comparison is not a part of any of the published articles, but presented here for some additional rigor.

Ensemble models such as Random Forest and Extreme Gradient Boosting have been suggested to excel in supervised machine learning regression tasks due to their ability to capture nonlinear interactions (Chen and Guestrin, 2016). Our results were in line with the expectations, and subsequently the Extreme Gradient Boosting Regressor (XGBRegressor) was selected as the best model for the forecasting task. All algorithms, except XGBRegressor, were implemented using the Scikit-learn library (Pedregosa et al., 2011).

The initial selection process was completed prior to the elections. Table 7 provides details how each of these algorithms would have performed had they been selected. This validation was done after the 2023 elections using actual election results.

Performance was measured across multiple metrics. MAE reflects the error related to the predictions of the regression models. In terms of MAE, the K-Nearest-Neighbors regressor was the most accurate with a score of 590.67. For accuracy and precision, which involve converting predicted vote shares into a binary prediction of whether a candidate is elected or not, Extreme Gradient Boosting regressor and Random Forest regressor demonstrated the best results, with both achieving an accuracy of 0.96 and a precision of 0.76. The other models produced similar accuracy and precision results, indicating effective performance across different methodologies. Additionally, the Kendall Rank Correlation provides insight into the ordinal ranking performance of each model, with Linear Support Vector regression achieving the highest correlation at 0.69.

Table 7. Comparison of ML Models Following 2023 Elections

Model	MAE	Precision	Accuracy	Kendall Rank Correlation
XGBRegressor	645.23	0.76	0.96	0.68
RandomForestRegressor	630.60	0.76	0.96	0.68
LinearRegression	642.39	0.75	0.96	0.67
LinearRegression(LOG)	654.29	0.73	0.96	0.67
Ridge	642.11	0.75	0.96	0.67
MLPRegressor	675.74	0.74	0.96	0.67
LinearSVR	601.08	0.72	0.95	0.69
KNeighborsRegressor	590.67	0.74	0.96	0.66

The implications of these results indicate that the choice of algorithm has a limited impact on model performance for this task. This finding suggests that enhancing the features used in the models should be the main focus, rather than determining which model is the most suitable. By improving feature quality and relevance, prediction accuracy and reliability are likely to improve more effectively than switching between different algorithms. This insight shifts the emphasis of model optimization from algorithm selection to feature engineering, highlighting the role that data plays in achieving better model performance. As a result, prioritizing feature selection and data enrichment should be the main focus of future modeling efforts in this context.

6.4.5 Finnish Parliamentary Elections 2023

In the 2023 Finnish parliamentary elections, the main governing party Social Democrats and the major opposition parties, the National Coalition and the Finns Party, recorded

notable improvements compared to four years ago (Arter, 2024). In contrast, the junior coalition parties, the Centre Party and the Greens, faced substantial setbacks and emerged as the election's major losers. The Swedish People's Party managed to maintain its previous standing, while the Left Alliance recorded its poorest result since its establishment. Within the opposition, the Christian Democrats enjoyed a relatively successful performance. The Movement Now Party only succeeded in securing the re-election of their leader despite running 177 candidates. The Finns Party achieved its most successful result to date with 20.1% of the vote, surpassing its last surge.

The election witnessed a routine change in parliamentary representation with 60 complete newcomers elected and 22 candidates who had previously run for election gaining seats. Meanwhile, 31 candidates did not succeed in their re-election campaigns and 32 Members of Parliament chose not to run again.

The campaign was relatively subdued, with economic issues predominating discussion amid voter fatigue from recent municipal and county elections in 2021 and 2022 (Arter, 2024). The economic debates focused on Finland's modest GDP growth and the highest inflation in 40 years, influencing a voter turnout of 72.6%, which was marginally less than the previous election.

6.4.6 Forecast Results

The newly developed forecast model, trained with data from the 2015 and 2019 elections, was employed to predict the outcomes. The model used the Extreme Gradient Boosting machine learning framework and included a number of new features such as past electoral success, candidate's political experience, campaigning budget, and other features detailed in the articles. This model was much more successful than the earlier models and predicted 151 of the 200 elected candidates from a pool of 2,468. The prediction accuracy was 75.5%.

The predictive performance of the model was evaluated against the polling averages. The model recorded a MAE of 1.62 at the national level, whereas the polling averages deviated by 0.71 percentage points from the actual results. Typically, the margin of error for polling methods is around 2 percentage points. Although the average error of the model is within this margin, some predictions significantly deviated from the actual results. In particular, the greatest discrepancies were observed for the True Finns (-3.58%), the Social Democratic Party (-2.79%) and the Green Party (+3.56%). Detailed explanations for these errors are explored in Article III.

Another perspective is to look at the order of the parties. The prediction model successfully identified the National Coalition as the winning party, which is particularly important for considerations like government formation. However, the model reversed the actual positions of the Finns Party and the Social Democratic Party incorrectly anticipating the Social Democratic Party would receive more votes than the

Finns Party. Similarly, both the model and the polling averages misjudged the order of the Greens and the Left Alliance, predicting the Greens to outperform the Left Alliance, contrary to the actual results.

Table 8. Election Results and Predictions for Finnish Parliamentary Elections 2023

Party	Election Result	Predicted Result	Poll Average
National Coalition	20.82%	19.74%	19.80%
True Finns	20.06%	16.48%	19.35%
Social Democratic Party	19.95%	17.16%	18.95%
Centre Party	11.29%	11.78%	11.05%
Left Alliance	7.06%	8.48%	8.50%
The Green Party	7.04%	10.60%	8.70%
Swedish People's Party	4.31%	5.28%	4.30%
Christian Democrats	4.22%	3.56%	4.20%
The Åland representative	2.82%	4.12%	3.30%
Movement Now	2.42%	2.81%	1.85%
MAE		1.62	0.71

An article discussing the forecast was published on the university website. Figure 7 illustrates the published article.

Although the model does not yet match the precision of traditional polling, it offers unique insights by focusing on individual candidates rather than just party-level outcomes. The evaluation of results highlighted the strengths and limitations of the model. The effectiveness of the model was evidenced by its performance against a simple baseline model, although its accuracy was lower than that of pollsters.

The findings align well with the theoretical frameworks presented, suggesting that voters use heuristics and mental shortcuts when selecting a candidate. Although the practical applicability of the model is limited by its accuracy compared to polling, it provides a complementary perspective in election forecasting by concentrating on granular, candidate-specific data rather than broader party-level trends.

This approach holds promise as part of a comprehensive electoral prediction framework that combines traditional and innovative methodologies, especially as traditional polling faces challenges such as declining response rates. The study underscores the potential for candidate-specific and data-driven models to enrich election forecasting, serving as a foundation for future research and practical applications in electoral contexts. These results are discussed in detail in the research Article III.

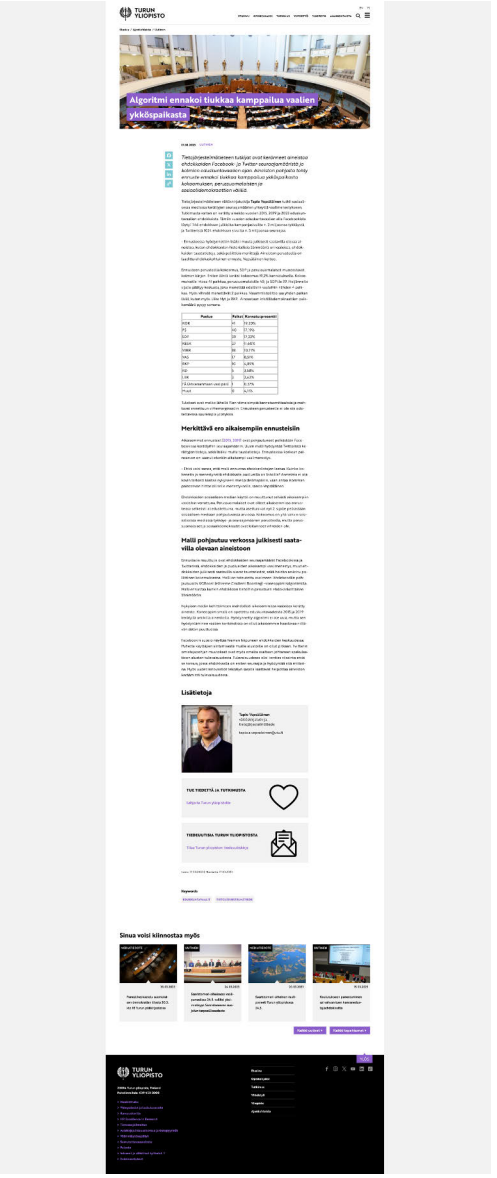

Figure 7. A prediction regarding the 2023 parliamentary elections was released on the University of Turku's website ahead of the elections.

7 Results

7.1 Model Overview

This research provides a complementary perspective on electoral forecasting. This viewpoint posits that publicly available candidate-related data can be leveraged to predict election outcomes, thus offering a fresh conceptual lens for political analysis.

This approach diverges from traditional forecasting methods by emphasizing the wealth of information accessible in modern digital environments. While this study has empirically tested the perspective within the context of three Finnish parliamentary elections, demonstrating its potential utility, it acknowledges the need for further adaptation and validation in diverse electoral contexts.

The forecast structure presented in Figure 8 serves as a concrete generalization of the concept. The conceptualization follows a structured four-step process:

1. **Candidate Data:** This step involves the collection and preprocessing of publicly available information related to candidates. These data form the foundation upon which predictions are made.

2. **Machine Learning Model:** The prepared candidate data feeds into a machine learning model designed to learn patterns and relationships that contribute to voter preferences.

3. **Expected Number of Votes:** Using the regression model, the number of votes expected by each candidate is predicted. This step transforms raw candidate data into actionable insights on candidate popularity and expected performance.

4. **Electoral Context:**Depending on the context, some adjustments may improve the accuracy of the model. In the example case of Finnish parliamentary elections, the predicted preference votes are then converted into proportional vote shares using the d'Hondt method, which simulates how votes translate into seats in a proportional representation system.

The model does not have strict requirements related to the use of the regression algorithm. Using a classification algorithm may be more suitable, depending on the learning material and prediction goals.

This model demonstrates an application of the viewpoint within the specific context of Finnish parliamentary elections. While the scope of this study is limited, the structured approach provides a framework for future adaptations and applications across different electoral systems. As such, the model serves as a practical element that complements the broader conceptual artifact, illustrating its potential utility and guiding future research efforts to refine and extend the approach to diverse electoral environments.

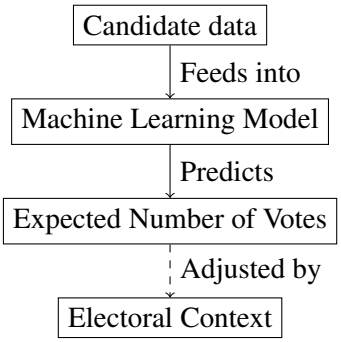

Figure 8. Forecast Structure

7.2 Summary of Key Findings

7.2.1 Evolution of Forecast Models

Between 2015 and 2023, three electoral forecast models were created, with their predicted outcomes being released before the elections. The forecasts and the data used to generate them were analyzed to assess the value of using publicly accessible information for electoral predictions. Each model underwent changes and improvements, showing a progression and refinement over time, which enhanced their performance across the three cycles of development. Each phase of development contributed to an understanding of voter behavior and the potential benefits of diverse data integration, as summarized in Table 9.

In 2015, the approach relied primarily on Facebook likes, which provided a starting point, but also revealed the limitations of relying on data collected from social media. The 2019 model continued to use Facebook likes, but allowed for the exploration and discovery of new perspectives that could be applied in future forecasts. For the 2023 elections, a major advancement occurred by incorporating multiple data sources and utilizing machine learning, resulting in higher accuracy and better alignment with theories related to political forecasting and political science in general.

Table 9. Forecast Summary: Finnish Parliamentary Elections 2015, 2019, and 2023

	2015	2019	2023
Prediction Method	Facebook Likes	Facebook Likes	XGBoost ML Model
Data Source	Facebook pages	Facebook pages	Combination of social media data, electoral history, and candidate attributes
Findings	Showed a positive correlation with vote shares. Noted instances of both overestimation and underestimation.	Demonstrated improved accuracy and better alignment with actual results, although not sufficient on its own.	Achieved high accuracy and outperformed baseline forecasts. Had a closer alignment with independent polls.
Accuracy Comparison	MAE was 4.81, which is less accurate than traditional polling.	MAE improved to 3.21, exceeding the performance of the 2015 model.	MAE was 1.62, superior to the baseline and approaching the accuracy of polling averages.
Implications	Acts as a significant but weak predictor and highlights the need for using data from multiple platforms.	Shows potential when integrating platforms and underscores the limitations of relying on single-source data.	Suggests the benefit of using diverse data types, pointing towards enhanced precision with comprehensive data integration.

7.2.2 Data Collection and Preparation

Online data collection requires a lot of effort due to the dynamic nature of the internet. Online content is continually updated, with frequent changes that affect the availability of information. Metrics such as likes, shares, or comments on social media exhibit constant fluctuations. Web pages are updated or removed, and new content is added consistently. This ever-changing environment poses challenges for consistent data collection, necessitating ongoing monitoring to capture accurate data snapshots over time. Although platforms like X and Facebook likely maintain extensive historical datasets, this information is not publicly accessible.

The data collection process used in this study was guided by feasibility and practicality, leveraging available data sources to the fullest extent possible. Due to the specific nature of the research and the data sources involved, an ad-hoc approach was necessary to address various practical constraints and issues related to data availability. In data science projects, such challenges are typical, requiring adaptations to

methodologies as needed (Salganik, 2019; Tufekci, 2014). The data collection phase was closely intertwined with data understanding, which required constant analysis and testing.

The data collection and preparation processes were partially guided by the CRISP-DM framework, focusing on the phases of data understanding and preparation (Wirth and Hipp, 2000). CRISP-DM framework provided a valuable benchmark for navigating typical challenges in data science projects. However, as the research advanced, it became evident that certain objectives, particularly those involving online data, required deviations from the standard approach.

An example of ad hoc adaptations in data collection was prioritizing candidates affiliated with established political parties due to the greater availability of pertinent data necessary for effective modeling. Although efforts were made to include a wide range of candidates, the focus on established parties facilitated a more streamlined data collection process. This selection introduces some subjective biases, which are disclosed here, particularly concerning social media data collection.

An outline of the data collection process is summarized in Table 10. More detailed information for each cycle can be found in the earlier chapters covering each election.

Table 10. Data Collection Overview

	2015	2019	2023
Facebook	Graph API	Web Scraping	Bing API
Twitter	Twitter API	Twitter API	Bing API
Voting Advice Applications	Post elections	Post elections	Pre elections, Using public API
Candidate Data	One month before elections, provided by Government	One month before elections, provided by Government	One month before elections, provided by Government
Historical Candidate Data	-	-	Collected before elections from Statistics of Finland

In 2015, Facebook data were collected using the Graph API. Voting Advice Applications were analyzed after the election. Candidate data was published by the government one month before the elections.

For the 2019 forecast, Facebook data collection method was changed to web scraping due to changes in API accessibility. Twitter data was collected using the Twitter API. Voting Advice Applications was published by Yle after the elections.

In 2023, significant adaptations were made, and the Bing API was used for data

collection on Facebook and Twitter, reflecting evolving data access methods. Voting Advice Applications data were collected before election using a public API, enhancing the timeliness of data acquisition. The collection of candidate data remained consistent with previous years, sourced from the government one month before the elections, while historical candidate data was collected in advance from Statistics of Finland.

7.2.3 Feature Selection

Feature selection and data collection went hand in hand, ensuring that the curated data was both relevant and useful in the context. During this process, the interplay between different features was investigated using typical measures such as correlation. Each published paper explores the features on a detailed level.

From the final evaluation, it was evident that feature selection is a key activity. Not all collected features contributed equally to the predictive accuracy. For instance, an analysis of the 2015 elections revealed that the number of shares and likes each publication received was strongly correlated with overall page likes. This redundancy indicated that incorporating these features did not add new information to the model's predictive capacity.

Certain features, such as occupation, were excluded due to feasibility concerns. Although the occupation should have predictive value, including it in the final model risked overfitting, particularly by disproportionately favoring candidates who are professional politicians. This is one example of subjective decisions that had to be made during feature selection.

7.3 Evaluation

The evaluation of our model demonstrates its predictive capability, which is evident from the results of our analysis. Even the most simplistic models seem to offer insight into trends and potential outcomes. Using historical data and trends, it successfully forecasts many important aspects of the 2023 elections, despite the inherent uncertainties and dynamics of political landscapes.

An in-depth assessment in article III reveals that all the features integrated into the model appear to have some predictive capability. Each variable contributes uniquely to the model's ability to anticipate election results, suggesting that the selection and incorporation of these features were pertinent to enhancing the model's accuracy. This aligns with the kernel theories presented earlier, which emphasize the importance of selecting relevant features to capture meaningful patterns. By analyzing the weight and influence of each feature, we aim to gain insight into whether the model captures relevant patterns and to improve our understanding of its outputs.

Assessing the results from a practical perspective indicates that the applicability

of the model is somewhat limited. The results of the 2023 electoral forecast show that while the model performs reasonably well overall, it currently has less practical applicability than opinion polls. However, its accurate prediction capability hints at the potential for future refinement and improvement.

8 Key Lessons

8.1 Forecasting Elections

The research presented herein makes contributions to the field of election predictions by introducing a method that utilize publicly accessible data concerning candidates' characteristics, in combination with information sourced from social media platforms. This approach represents a marginally studied dimension in election prediction research and addresses the growing demand for alternative forecasting methodologies.

Unlike many existing forecasting approaches that operate primarily at the macro level, this study provides a candidate-specific perspective. By directing attention to individual candidates, it refines the theoretical foundation of electoral studies and supports existing theories. This emphasis on granularity allows for a deep understanding of the dynamics surrounding electoral candidates and their potential outcomes.

Furthermore, the analysis advances the exploration of social media as a predictive tool for elections. Although the challenge of bias persists, the study identifies strategies to mitigate these biases, offering a pathway for future research and development in the area of election forecasting. The integration of social media insights into election predictions is a promising field that requires continued exploration to refine methods and understand dynamics.

From a practical viewpoint, the study offers contributions that are twofold. For commercial entities looking for novel prediction methods, the proposed approach provides an innovative pathway to refine electoral forecasts. A hybrid methodology, which combines traditional polling data with candidate-specific insights, has the potential to improve the precision of predictions.

For candidates themselves, access to candidate-level information can be practically beneficial. Although it does not address broader electoral challenges, it offers valuable self-assessment insights in relation to peer candidates. However, caution is advised in interpreting these insights, recognizing their limitations while making use of their practical utility.

8.2 The Data Driven Research Agenda

The study reinforces established concepts in the field of data-driven research, highlighting several important points. The principle of ”garbage in - garbage out” underscores the necessity of applying domain knowledge when dealing with data. Simple baseline models often outperform more complex predictions, highlighting the importance of choosing the right approach rather than focusing solely on statistical significance.

Frameworks such as the Cross Industry Standard Process for Data Mining (CRISP-DM) provide general guidance but may not align seamlessly with research conducted by single authors or small teams outside of the business context. Future research efforts could investigate how to refine the methodological approach for predictive analytics using online data.

A significant aspect of data-driven research is the requirement for cross-disciplinary understanding. In the realm of data science, having technical proficiency in data handling is important, but it alone does not suffice. Researchers must be aware of the contextual background of the data they work with. This involves grasping the nuances of domain-specific knowledge that inform the relevance and applicability of the data used. Whether it pertains to electoral studies, economics, or social media analysis, a understanding the field enhances the ability to interpret data meaningfully.

Furthermore, the study involves many subjective decisions that need to be presented transparently to maintain the rigor of the investigation. During research, subjective decisions may arise in the form of choices regarding data selection, methodological approaches, and interpretation of results. These decisions can significantly influence the results and conclusions of a study. In doing so, researchers enable others to critically evaluate the methodologies and findings presented in the study, fostering open dialogue and making it possible for the research to be replicated or expanded upon in future studies. Clear documentation of subjective decisions contributes to the overall robustness and credibility of the research process.

8.3 Utilization of Online Data

Using online data effectively requires a variety of skills, including proficiency in web-scraping and understanding the use of APIs. These technical skills are essential for accessing and gathering data from the vast array of online sources available today. Web-scraping involves extracting data from websites, which often requires knowledge of HTML and scripting languages such as Python. APIs, on the other hand, provide structured access to data made available by web services, usually through specified endpoints that require understanding of HTTP requests and responses.

Understanding the context and creation of the data is equally important. Online data can come from various sources, each with different purposes and potential

biases. For instance, recognizing the differences between self-reported and factual information in areas like budgets and education is important for accurate data interpretation. Self-reported data are often subject to individual perceptions and may not always align with objectively recorded facts. Researchers must be wary of these discrepancies and adjust their analyses accordingly to ensure data reliability.

Online data is sensitive to subjective biases, which requires careful interpretation and validation. The dynamic and unregulated nature of online data means that it can often reflect the biases and perspectives of those who create or contribute to it. For example, social media platforms are filled with user-generated content that can vary greatly in quality and reliability. Thus, when analyzing such data, researchers must incorporate strategies to assess the credibility of sources. This can involve cross-referencing data with verified information, applying statistical techniques to detect bias patterns, or using domain expertise to contextualize findings.

Ethical considerations are paramount when using online data. Issues related to privacy, data ownership, and consent must be addressed to protect individuals' rights. Researchers are required to adhere to ethical guidelines and legal frameworks, which may include anonymizing data to prevent identifying individuals or securing permission from data owners prior to use.

In summary, effective utilization of online data requires a holistic approach that includes technical skills, contextual awareness, bias assessment, and ethical sensitivity. By integrating these components, researchers can harness the potential of online data to generate meaningful insights while maintaining the integrity and credibility of their work.

List of References

Abbasi, A., Chiang, R. H. L., and Xu, J. (2023). Data science for social good. *Journal of the Association for Information Systems*, 24(6):1439 – 1458.

Agarwal, R. and Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for is research. *Information systems research*, 25(3):443–448.

Aichner, T., Grünfelder, M., Maurer, O., and Jegeni, D. (2021). Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Aldabbas, H., Bajahzar, A., Alruily, M., Qureshi, A. A., Amir Latif, R. M., and Farhan, M. (2020). Google play content scraping and knowledge engineering using natural language processing techniques with the analysis of user reviews. *Journal of Intelligent Systems*, 30(1):192–208.

Aldrich, J. H., Gibson, R. K., Cantijoch, M., and Konitzer, T. (2016). Getting out the vote in the social media era: Are digital tools changing the extent, nature and impact of party contacting in elections? *Party Politics*, 22(2):165–178.

Alhabash, S. and Ma, M. (2017). A tale of four platforms: Motivations and uses of facebook, twitter, instagram, and snapchat among college students? *Social media+ society*, 3(1):205630.

Arter, D. (2024). The making of an 'unhappy marriage'? the 2023 finnish general election. *West European Politics*, 47(2):426–438.

Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1, pages 492–499. IEEE.

Atanasov, P., Witkowski, J., Mellers, B., and Tetlock, P. (2024). Crowd prediction systems: Markets, polls, and elite forecasters. *International Journal of Forecasting*.

Azevedo, A. and Santos, M. F. (2008). Kdd, semma and crisp-dm: A parallel overview. In *Proceedings of Informatics and Data Mining*, pages 182–185. IADS-DM.

Badiee, S., Crowell, J., Noe, L., Pittman, A., Rudow, C., and Swanson, E. (2021). Open data for official statistics: History, principles, and implementation. *Statistical Journal of the IAOS*, 37(1):139–159.

Bailey, M. A. (2023). A new paradigm for polling. *Harvard Data Science Review*, 5(3).

Bailey, M. A. (2024). *Polling at a Crossroads: Rethinking Modern Survey Research*. Methodological Tools in the Social Sciences. Cambridge University Press.

Banducci, S. (2018). The role of mass media in shaping public opinion and vote behavior. In Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C., editors, *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*, pages 305–318. Routledge, London.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.

Baumeister, R. F., Vohs, K. D., and Oettingen, G. (2016). Pragmatic prospection: How and why people think about the future. *Review of general psychology*, 20(1):3–16.

Behnert, J., Lajic, D., and Bauer, P. C. (2024). Can we predict multi-party elections with google trends data? evidence across elections, data windows, and model classes. *Journal of Big Data*, 11(1):30.

Berggren, N., Jordahl, H., and Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of public economics*, 94(1-2):8–15.

Bimber, B., Cunill, M. C., Copeland, L., and Gibson, R. (2015). Digital media and political participation: The moderating role of political interest across acts and over time. *Social science computer review*, 33(1):21–42.

Blais, A., Gidengil, E., and Nevitte, N. (2006). Do polls influence the vote? In Brady, H. E. and Johnston, R. G. C., editors, *Capturing Campaign Effects*, pages 263–283. University of Michigan Press.

Bollen, J., Mao, H., and Pepe, A. (2021). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 450–453.

Bølstad, J. (2018). Is there a rational public? In Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C., editors, *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*, pages 383–393. Routledge, London.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.

Borg, S. (2019). The finnish parliamentary election of 2019: Results and voting patterns. *Scandinavian Political Studies*, 42(3-4):182–192.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679.

Brito, K., Filho, S., C., R. L., and Adeodato, P. J. L. (2021). A systematic review of predicting elections based on social media data: Research challenges and future directions. *IEEE Transactions on Computational Social Systems*, 8(4):819.

Broucke, S. V. and Baesens, B. (2018). *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. Apress, Berkeley, CA, 1 edition. Published: 19 April 2018 (Print), 18 April 2018 (eBook).

Bruns, A. (2019). After the 'apicalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11):1544–1566.

Buhl, H. U., Röglinger, M., Moser, F., and Heidemann, J. (2013). Big data: a fashionable topic with (out) sustainable relevance for research and practice? *Business & Information Systems Engineering*, 5:65–69.

Buyalskaya, A., Gallo, M., and Camerer, C. F. (2021). The golden age of social science. *Proceedings of the National Academy of Sciences*, 118(5):e2002923118.

Campbell, A., Converse, P. E., Miller, W. E., and Stokes, D. E. (1960). *The american voter*. University of Chicago Press.

Cao, L. (2010). Domain-driven data mining: Challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):755–769.

Cao, L. (2017). Data science: challenges and directions. *Communications of the ACM*, 60(8):59–68.

Cao, L., Yu, P. S., Zhang, C., and Zhao, Y. (2010). *Domain Driven Data Mining*. Computer Science, Computer Science (R0). Springer New York, NY, 1 edition. Published: 20 January 2010.

Carlisle, J. E. and Patton, R. C. (2013). Is social media changing how we understand political engagement? an analysis of facebook and the 2008 presidential election. *Political research quarterly*, 66(4):883–895.

Cavallo, A. and Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2):151–178.

Cerina, R. and Duch, R. (2020). Measuring public opinion via digital footprints. *International Journal of Forecasting*, 36(3):987–1002.

Cerina, R. and Duch, R. (2023). Artificially intelligent opinion polling. *arXiv preprint arXiv:2309.06029*.

Chae, I., Schweidel, D. A., Evgeniou, T., and Padmanabhan, V. (2024). Mixing user-and publisher-generated content: Quantifying ugc spillover effect in a hybrid content environment. *Journal of Interactive Marketing*. First published online April 15, 2024.

Chakravartty, A. (2017). Scientific realism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2017 edition.

Chauhan, P., Sharma, N., and Sikka, G. (2021). The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 12(2):2601–2627.

Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, pages 1165–1188.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chen, W. and Hirschheim, R. (2004). A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Information systems journal*, 14(3):197–235.

Chin, C. Y. and Wang, C. L. (2021). A new insight into combining forecasts for elections: The role of social media. *Journal of Forecasting*, 40(1):132–143.

Chini, F., Pezzotti, P., Orzella, L., Borgia, P., and Guasticchi, G. (2011). Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. *BMC public health*, 11:1–8.

Christensen, H. S., Rosa, M. S. L., and Groenlund, K. (2020). How candidate characteristics affect favorability in european parliament elections: Evidence from a conjoint experiment in finland. *European Union Politics*, 21(3):519–540.

Chung, J. and Mustafaraj, E. (2011). Can collective sentiment expressed on twitter predict political elections? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 1770–1771.

Colladon, A. F. (2020). Forecasting election results by studying brand importance in online news. *International Journal of Forecasting*, 36(2):414–427.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The public opinion quarterly*, 64(4):464–494.

Creswell, J. W. and Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Inc., Los Angeles, 5th edition.

David, E., Zhitomirsky-Geffet, M., Koppel, M., and Uzan, H. (2016). Utilizing facebook pages of the political parties to automatically predict the political orientation of facebook users. *Online Information Review*, 40(5):610–623.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

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.

Dinas, E. (2016). The evolving role of partisanship. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 265–286. SAGE Publications, United Kingdom, 1 edition.

do Nascimento Silva, V. and Silva, R. H. A. (2019). Are algorithms affecting the democracy in brazil? In *Proceedings of the International Symposium on Ethical Algorithms*, December 2018. Presented at the International Symposium on Ethical Algorithms.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Dongo, I., Cardinale, Y., Aguilera, A., Martinez, F., Quintero, Y., Robayo, G., and Cabeza, D. (2021). A qualitative and quantitative comparison between web scraping and api methods for twitter credibility analysis. *International Journal of Web Information Systems*, 17(6):580–606.

Duin, R. P. and Tax, D. M. (2000). Experiments with classifier combining rules. In *International Workshop on Multiple Classifier Systems*, pages 16–29. Springer.

Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. The National Academies Press, Washington, DC.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., and Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Eisenberg, D. and Ketcham, J. (2004). Economic voting in us presidential elections: Who blames whom for what. *Topics in Economic Analysis & Policy*, 4(1).

Evans, G. and Northmore-Ball, K. (2018). Long-term factors: Class and religious cleavages. In Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C., editors, *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*, pages 123–135. Routledge, London.

Fan, C., Esparza, M., Dargin, J., Wu, F., Oztekin, B., and Mostafavi, A. (2020). Spatial biases in crowdsourced data: Social media content attention concentrates on populous areas in disasters. *Computers, Environment and Urban Systems*, 83:101514.

Fayyad, U. M., Haussler, D., and Stolorz, P. E. (1996). Kdd for science data analysis: Issues and examples. In *KDD*, pages 50–56.

Finland, S. (2024). License applied to statistics finland's open data materials. Retrieved on August 24, 2024.

Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C. (2018). *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*. Routledge, London.

Fisher, S. (2018). Election forecasting. In Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C., editors, *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*, pages 496–508. Routledge, London.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., and Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of retailing*, 90(2):217–232.

Ford, R., Wlezien, C., Pickup, M., and Jennings, W. (2016). Polls and votes. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 787–812. SAGE Publications, United Kingdom, 1 edition.

Franch, F. (2013). (wisdom of the crowds) 2: 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10(1):57–71.

Fritsch, M., Haupt, H., and Schnurbus, J. (2024). Efficiency of poll-based multi-period forecasting systems for german state elections. *International Journal of Forecasting*.

Fuchs, C. (2017). From digital positivism and administrative big data analytics towards critical digital and social media research! *European Journal of Communication*, 32(1):37–49.

Fung, G. P. C., Yu, J. X., and Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. In *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.*, pages 395–402. IEEE.

Furche, T., Gottlob, G., Libkin, L., Orsi, G., and Paton, N. W. (2016). Data wrangling for big data: Challenges and opportunities. In *19th International Conference on Extending Database Technology*, pages 473–478.

García, S., Luengo, J., and Herrera, F. (2015). *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*. Springer Cham. Copyright Springer International Publishing Switzerland 2015.

Garzia, D. (2016). Voter evaluation of candidates and party leaders. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 633–653. SAGE Publications, United Kingdom, 1 edition.

Gayo-Avello, D. (2012). No, you cannot predict elections with twitter. *IEEE Internet Computing*, 16(6):91–94.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social science computer review*, 31(6):649–679.

Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems*. O'Reilly, second edition. edition.

Giglietto, F., Righetti, N., and Marino, G. (2019). Understanding coordinated and inauthentic link sharing behavior on facebook in the run-up to 2018 general election and 2019 european election in italy.

Gil de Zúñiga, H., Molyneux, L., and Zheng, P. (2014). Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. *Journal of communication*, 64(4):612–634.

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(7232):1012–1014.

Gligorijević, V. and Pržulj, N. (2015). Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National academy of sciences*, 107(41):17486–17490.

Golder, M. (2005). Democratic electoral systems around the world, 1946–2000. *Electoral Studies*, 24(1):103–121.

Golder, S. A. and Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40(1):129–152.

Graefe, A. (2015). German election forecasting: Comparing and combining methods for 2013. *German Politics*, 24(2):195–204.

Graefe, A. (2019). Accuracy of german federal election forecasts, 2013 & 2017. *International Journal of Forecasting*, 35(3):868–877.

Graefe, A., Armstrong, J. S., Jones Jr, R. J., and Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, 30(1):43–54.

Gregor, S. (2006). The nature of theory in information systems. *MIS quarterly*, pages 611–642.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Grönlund, K. and Strandberg, K., editors (2023). *Finland turned right: Voting and Public Opinion in the Parliamentary Election of 2023*. The Social Science Research Institute, Åbo Akademi University.

Groves, R. M. (2011). Three eras of survey research. *Public opinion quarterly*, 75(5):861–871.

Gschwend, T. and Meffert, M. F. (2017). Strategic voting. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 339–366. SAGE Publications.

Gu, Y., Qian, Z. S., and Chen, F. (2016). From twitter to detector: Real-time traffic incident detection using social media data. *Transportation research part C: emerging technologies*, 67:321–342.

Haenlein, M. and Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4):5–14.

Haenschen, K. (2016). Social pressure on social media: Using facebook status updates to increase voter turnout. *Journal of Communication*, 66(4):542–563.

Han, J., Pei, J., and Tong, H. (2023). *Data Mining: Concepts and Techniques*. Elsevier, 4th edition.

Hardt, D. (2012). The OAuth 2.0 Authorization Framework. RFC 6749.

Hargittai, E. and Sandvig, C. (2015). *Digital research confidential: The secrets of studying behavior online*. MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition. Accessed via ProQuest Ebook Central.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

Holian, D. B. and Prysby, C. (2014). Candidate character traits in the 2012 presidential election. *Presidential Studies Quarterly*, 44(3):484–505.

Holian, D. B. and Prysby, C. (2020). Polls and elections: Did character count? candidate traits and the 2016 presidential vote. *Presidential Studies Quarterly*, 50(3):666–689.

Huberty, M. (2015). Can we vote with our tweet? on the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, 31(3):992–1007.

Huberty, M. E. (2013). Multi-cycle forecasting of congressional elections with social media. In *Proceedings of the 2nd workshop on Politics, Elections and Data*, pages 23–30.

Hutchings, V. L. and Jefferson, H. J. (2018). The sociological and social-psychological approaches. In Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C., editors, *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*, pages 21–29. Routledge, London.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, 4(4):43–46.

Iivari, J. (2007). A paradigmatic analysis of information systems as a design science. *Scandinavian journal of information systems*, 19(2):5.

Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook. *Cambridge Analytica, and privacy protection. Computer*, 51(8):56–59.

Islam, R. (2006). Does more transparency go along with better governance? *Economics & Politics*, 18(2):121–167.

Isotalo, V. (2021). Improving candidate-based voting advice application design: The case of finland. *Informaatiotutkimus*, 40(3):85–109.

Isotalo, V., Helimäki, T., Mattila, M., and SCHOULTZ, Å. V. (2023). When does ideology matter? party lists, personal attributes and the effect of ideology on intra-party success. *European Journal of Political Research*, 62(4):1257–1279.

Isotalo, V., Mattila, M., and von Schoultz, Å. (2020). Ideological mavericks or party herd? the effect of candidates' ideological positions on intra-party success. *Electoral Studies*, 67:102187.

Iyengar, S. and Simon, A. F. (2000). New perspectives and evidence on political communication and campaign effects. *Annual review of psychology*, 51(1):149–169.

Jennings, W., Lewis-Beck, M., and Wlezien, C. (2020). Election forecasting: Too far out? *International Journal of Forecasting*, 36(3):949–962.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Jungherr, A., Jürgens, P., and Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social science computer review*, 30(2):229–234.

Jungherr, A., Schoen, H., Posegga, O., and Jürgens, P. (2017). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social science computer review*, 35(3):336–356.

Jääskeläinen, A. (2023). The finnish election system: Overview. Technical Report 2023:5, Ministry of Justice, Finland.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., and Damian, D. (2014). The promises and perils of mining github. In *Proceedings of the 11th working conference on mining software repositories*, pages 92–101.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., and Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288.

Kang, S. and Oh, H.-S. (2024). Forecasting south korea's presidential election via multiparty dynamic bayesian modeling. *International Journal of Forecasting*, 40(1):124–141.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331.

Karvonen, L. (2014). *Parties, Governments and Voters in Finland: Politics Under Fundamental Societal Transformation*. ECPR – Monographs. ECPR Press, Colchester, United Kingdom.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kennedy, R., Wojcik, S., and Lazer, D. (2017). Improving election prediction internationally. *Science*, 355:515–552.

Kestilä-Kekkonen, E., Sipinen, J., and Westinen, J. (2024). Harkintajoukot ja taktinen äänestäminen. In Kestilä-Kekkonen, E., Rapeli, L., and Söderlund, P., editors, *Pääministerivaalit polarisaation aikakaudella*, number 2024:10 in Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita, pages 198–223. Oikeusministeriö.

Kestilä-Kekkonen, E. and von Schoultz, Å. (2020). Ehdokkaat vaalikentillä: Eduskuntavaalit 2019. Technical Report 978-952-259-890-5, Oikeusministeriö.

Kilkenny, M. F. and Robinson, K. M. (2018). Data quality:"garbage in–garbage out". *Health Information Management Journal*, 47(3):103–105.

Kirilenko, A. P. and Stepchenkova, S. O. (2014). Public microblogging on climate change: One year of twitter worldwide. *Global environmental change*, 26:171–182.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE Publications Ltd.

Klein, M. and Rosar, U. (2016). Candidate attractiveness. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 688–708. SAGE Publications, United Kingdom, 1 edition.

Kondraganti, A., Narayanamurthy, G., and Sharifi, H. (2024). A systematic literature review on the use of big data analytics in humanitarian and disaster operations. *Annals of Operations Research*, 335(3):1015–1052.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Science*, 110(15):5802–5805.

Krumm, J., Davies, N., and Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4):10–11.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer Science+Business Media, New York, corrected at 5th printing 2016 edition.

Lampos, V., Majumder, M. S., Yom-Tov, E., Edelstein, M., Moura, S., Hamada, Y., Rangaka, M. X., McKendry, R. A., and Cox, I. J. (2021). Tracking covid-19 using online search. *NPJ digital medicine*, 4(1):17.

Lau, R. R. and Redlawsk, D. P. (2006). *How Voters Decide: Information Processing in Election Campaigns*. Cambridge Studies in Public Opinion and Political Psychology. Cambridge University Press, Cambridge.

Lawrence, J. A. and Ehle, K. (2019). Combatting unauthorized webscraping-the remaining options in the united states for owners of public websites despite the recent hiq labs v. linkedin decision. *Computer Law Review International*, 20(6):171–174.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., and Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866):189–196.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science*, 323(5915):721–723.

Le, H. T., Boynton, G., Mejova, Y., Shafiq, Z., and Srinivasan, P. (2017). Revisiting the american voter on twitter. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 4507–4519.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Legg, C. and Hookway, C. (2024). Pragmatism. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2024 edition.

Lewis-Beck, M. S. (2005). Election forecasting: Principles and practice. *The British Journal of Politics and International Relations*, 7(2):145–164.

Lewis-Beck, M. S. and Dassonneville, R. (2015). Forecasting elections in europe: Synthetic models. *Research & Politics*, 2(1):205316.

Lewis-Beck, M. S. and Paldam, M. (2000). Economic voting: an introduction. *Electoral Studies*, 19(2-3):113–121.

Lewis-Beck, M. S. and Stegmaier, M. (2000). Economic determinants of electoral outcomes. *Annual Review of Political Science*, 3:183–219.

Lewis-Beck, M. S. and Stegmaier, M. (2014a). Us presidential election forecasting: Introduction. *PS: Political Science & Politics*, 47(2):284–288.

Lewis-Beck, M. S. and Stegmaier, M. (2014b). Us presidential election forecasting: Introduction. *PS: Political Science and Politics*, 47:284–288.

Lippmann, W. (1922). *Public opinion: Harcourt*. Brace and Co.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Liu, R., Yao, X., Guo, C., and Wei, X. (2021). Can we forecast presidential election using twitter data? an integrative modelling approach. *Annals of GIS*, 27(1):43–56.

Lu, Y., Heatherly, K. A., and Lee, J. K. (2016). Cross-cutting exposure on social networking sites: The effects of sns discussion disagreement on political participation. *Computers in Human Behavior*, 59:74–81.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Luo, X. and Zhang, J. (2013). How do consumer buzz and traffic in social media marketing predict the value of the firm? *Journal of Management Information Systems*, 30(2):213–238.

Maass, W., Parsons, J., Purao, S., Storey, V. C., and Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, 19(12):1.

Maddens, B., Wauters, B., Noppe, J., and Fiers, S. (2006). Effects of campaign spending in an open list pr system: The 2003 legislative elections in flanders/belgium. *West European Politics*, 29(1):161–168.

Mao, Y., Wang, D., Muller, M., Varshney, K. R., Baldini, I., Dugan, C., and Mojsilović, A. (2019). How data scientist swork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction*, 3:1–23.

Markham, A., Buchanan, E., with feedback from the AOIR Ethics Working Committee, et al. (2012). Ethical decision-making and internet research: Recommendations from the aoir ethics working committee (version 2.0). *AoIR Ethics Working Committee*.

Martinez, I., Viles, E., and Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research*, 24:100183.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. (2019). Crisp-dm twenty years later: From data mining

processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, 33(8):3048–3061.

McAllister, I. (2007). 571 The Personalization of Politics. In *The Oxford Handbook of Political Behavior*. Oxford University Press.

Mellon, J. (2018). Making inferences about elections and public opinion using incidentally collected data. In Fisher, J., Fieldhouse, E., Franklin, M. N., Gibson, R. K., Cantijoch, M., and Wlezien, C., editors, *The Routledge Handbook of Elections, Voting Behavior and Public Opinion*, pages 522–533. Routledge, London.

Mellon, J. and Prosser, C. (2017). Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008.

Mingers, J., Mutch, A., and Willcocks, L. (2013). Critical realism in information systems research. *MIS quarterly*, 37(3):795–802.

Ministry of Justice, Finland (2024a). County elections and municipal elections. `https://vaalit.fi/en/county-elections-and-municipal-elections`.

Ministry of Justice, Finland (2024b). Elections to the european parliament 2024. `https://vaalit.fi/en/european-elections`.

Ministry of Justice, Finland (2024c). Municipal elections. `https://vaalit.fi/en/municipal-elections`.

Mongrain, P. (2021). 10 downing street: who's next? seemingly unrelated regressions to forecast uk election results. *Journal of Elections, Public Opinion and Parties*, 31(1):22–32.

Moore, D. W. (2008). *The opinion makers: an insider exposes the truth behind the polls*. Beacon Press.

Moradi, M. and Keyvanpour, M. (2015). Captcha and its alternatives: A review. *Security and Communication Networks*, 8(12):2135–2156.

Morgenthaler, S. (2009). Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44.

Muniesa, F. (2015). Actor-network theory. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 80–84. Elsevier, Oxford, second edition edition.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., others, and Harwood, P. (2014). Social media in public opinion research: Executive summary of the aapor task force on emerging technologies in public opinion research. *Public opinion quarterly*, 78(4):788–794.

Mustapha, S., Man, M., Bakar, W. A. W. A., Yusof, M. K., and Sabri, I. A. A. (2024). A demystified overview of data scraping. *International Journal of Data Science and Advanced Analytics*, 6(6):290–296.

Nadeau, R., Dassonneville, R., Lewis-Beck, M. S., and Mongrain, P. (2020). Are election results more unpredictable? a forecasting test. *Political Science Research and Methods*, 8(4):764–771.

Nadeau, R. and Lewis-Beck, M. S. (2020). Election forecasts: Cracking the danish case. *International Journal of Forecasting*, 36(3):892–898.

Neupane, B., Woon, W. L., and Aung, Z. (2017). Ensemble prediction model with expert selection for electricity price forecasting. *Energies*, 10(1):77.

Newport, F. (2016). *The Gallup Poll: Public Opinion 2015*. Rowman & Littlefield.

Newton, K. (2006). May the weak force be with you: The power of the mass media in modern politics. *European Journal of Political Research*, 45(2):209–234.

Nizah, M. A. M. and Bakar, A. R. A. (2019). Whatsapp election in malaysia: Assessing the impact of instant messaging on malaysia's 14th general election. *International Journal of Academic Research in Business and Social Sciences*, 9(3):132–146.

Nurmi, H. and Nurmi, L. (2015). The parliamentary election in finland april 19, 2015. *Electoral Studies*, 40(1):433–438.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. . (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Paper presented at the Fourth International AAAI Conference on Weblogs and Social Media*.

of Justice, M. (1999). Act on the openness of government activities. Section 2.

Orlikowski, W. J. and Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, 2(1):1–28.

Panagopoulos, C., Larimer, C. W., and Condon, M. (2014). Social pressure, descriptive norms, and voter mobilization. *Political Behavior*, 36:451–469.

Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert systems with applications*, 42(6):2928–2934.

Patel, A. S., Vyas, R., Vyas, O., and Ojha, M. (2022). A study on video semantics; overview, challenges, and applications. *Multimedia Tools and Applications*, 81(5):6849–6897.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., others, and Dubourg, V. (2011). Scikit-learn: Machine learning in python. the journal of machine learning research. *12*, pages 2825–2830.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P. H., Frazier, D. T., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Trapero Arenas, J. R., Wang, X., Winkler, R. L., Yusupova, A., and Ziel, F. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871.

Phillips, L., Dowling, C., Shaffer, K., Hodas, N., and Volkova, S. (2017). Using social media to predict the future: a systematic literature review. arXiv preprint.

Pinch, T. J. and Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science*, 14(3):399–441.

Pollock, R. (2015). Open definition. `https://opendefinition.org/od/2.1/en/`. Version 2.1, Released: November 10, 2015. Accessed: 2024-11-13.

Porter, A. and Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745.

Project, T. G. (n.d.). The gdelt project. `https://www.gdeltproject.org/`, accessed on 28 Aug 2024.

Put, G.-J., Smulders, J., and Maddens, B. (2019). How local personal vote-earning attributes affect the aggregate party vote share: Evidence from the belgian flexible-list pr system (2003–2014). *Politics*, 39(4):464–479.

Quinlan, S. and Lewis-Beck, M. S. (2021). Forecasting government support in irish general elections: Opinion polls and structural models. *International Journal of Forecasting*, 37(4):1654–1665.

Redlawsk, D. P. and Pierce, D. R. (2016). Emotions and voting. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, pages 406–432. SAGE Publications.

Redmond, P. and Regan, J. (2015). Incumbency advantage in a proportional electoral system: A regression discontinuity analysis of irish elections. *European Journal of Political Economy*, 38:244–256.

Rofrío, D., Ruiz, A., Sosebee, E., Raza, Q., Bashir, A., Crandall, J., and Sandoval, R. (2019). Presidential elections in ecuador: Bot presence in twitter. In *Paper presented at the 2019 Sixth International Conference on eDemocracy & eGovernment (ICEDEG*.

Rogers, T., Green, D. P., Ternovski, J., and Young, C. F. (2017). Social pressure and voting: A field experiment conducted in a high-salience election. *Electoral Studies*, 46:87–100.

Rosenbaum, P. (2021). *Design of observational studies*. Springer, 2nd edition.

Rousidis, D., Koukaras, P., and Tjortjis, C. (2020). Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9):6279–6311.

Roy, J., Singh, S. P., and Fournier, P. (2021). *The Power of Polls?: A Cross-national Experimental Analysis of the Effects of Campaign Polls*. Cambridge University Press.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., and Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513.

Santos, J. S., Bernardini, F., and Paes, A. (2021). A survey on the use of data and opinion mining in social media to political electoral outcomes prediction. *Social Network Analysis and Mining*, 11:1–39.

Sarker, S., Xiao, X., and Beaulieu, T. (2013). Guest editorial: Qualitative studies in information systems: A critical review and some guiding principles. *MIS quarterly*, 37(4):iii–xviii.

Savage, M. and Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5):885–899.

Sayago Heredia, J., Flores-García, E., and Solano, A. R. (2020). Comparative analysis between standards oriented to web services: Soap, rest and graphql. In *Applied Technologies: First International Conference, ICAT 2019, Quito, Ecuador, December 3–5, 2019, Proceedings, Part I 1*, pages 286–300. Springer.

Schmitt, H. and Teperoglou, E. (2017). The study of less important elections. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, pages 56–79. SAGE Publications.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., and Gloor, P. (2013). The power of prediction with social media. *Internet research*, 23(5):528–543.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., and Kamaev, V. A. e. (2013). A survey of forecast error measures. *World applied sciences journal*, 24(24):171–176.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Shmueli, G. and Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3):553–572.

Shugart, M. S., Valdini, M. E., and Suominen, K. (2005). Looking for locals: Voter information demands and personal vote-earning attributes of legislators under proportional representation. *American Journal of political science*, 49(2):437–449.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118.

Skoric, M. M., Liu, J., and Jaidka, K. (2020). Electoral and public opinion forecasts with social media data: A meta-analysis. *Information*, 11(4):187.

Söderlund, P., von Schoultz, Å., and Papageorgiou, A. (2021). Coping with complexity: Ballot position effects in the finnish open-list proportional representation system. *Electoral Studies*, 71:102330.

Statista (2022). Average time spent per day with digital media in the united states from 2011 to 2022. Accessed: 2022-07-15.

Stegmaier, M., Lewis-Beck, M. S., and Park, B. (2017). The vp-function: a review. *M. StegmaierM. Lewis-Beck, & B. Park The VP-Function: A Review*, 2:584–605.

Stoetzer, L. F., Neunhoeffer, M., Gschwend, T., Munzert, S., and Sternberg, S. (2019). Forecasting elections in multiparty systems: a bayesian approach combining polls and fundamentals. *Political Analysis*, 27(2):255–262.

Strandberg, K. (2013). A social media revolution or just a case of history repeating itself? *The use of social media in the*, 15(8):1329–1347.

Strandberg, K., Carlson, T., and Snickars, W. (2024). Nuoret ehdokkaat ja kansalaiset sosiaalisessa mediassa. In Kestilä-Kekkonen, E., Rapeli, L., and Söderlund, P., editors, *Pääministerivaalit polarisaation aikakaudella*, number 2024:10 in Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita, pages 66–87. Oikeusministeriö.

Strate, L. (2004). Media ecology. *Communication Research Trends*, 23(2):1–48.

Teixeira, J. (2018). *Coopetition in an Open-Source Way – Lessons from Mobile and Cloud Computing Infrastructures*. Phd thesis, University of Turku, Turku School of Economics, Department of Information Systems Science, Turku, Finland.

Thatcher, J., Pu, W., and Pienta, D. (2018). Is information systems a (social) science? *Communications of the Association for Information Systems*, 43(1):11.

Theocharis, Y., Lowe, W., Van Deth, J. W., and García-Albacete, G. (2015). Using twitter to mobilize protest action: online mobilization patterns and action repertoires in the occupy wall street, indignados, and aganaktismenoi movements. *Information, Communication & Society*, 18(2):202–220.

Townsend, L. and Wallace, C. (2016). Social media research: A guide to ethics. *University of Aberdeen*, 1:16.

Trezza, D. (2023). To scrape or not to scrape, this is dilemma. the post-api scenario and implications on digital research. *Frontiers in Sociology*, 8:1145038.

Trnka, M., Cerny, T., and Stickney, N. (2018). Survey of authentication and authorization for the internet of things. *Security and Communication Networks*, 2018(1):4351603.

Tsatsou, P. (2016). *Internet studies: Past, present and future directions*. Routledge.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 505–514.

Tufte, E. R. (1975). Determinants of the outcomes of midterm congressional elections. *American Political Science Review*, 69(3):812–826.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley series in behavioral science : quantitative methods. Addison-Wesley, Reading, Mass.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the international AAAI conference on web and social media*, volume 4, pages 178–185.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2011). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social science computer review*, 29:402–418.

Van Dijck, J. and Poell, T. (2013). Understanding social media logic. *Media and communication*, 1(1):2–14.

Van Wel, L. and Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, 6(2):129–140.

Victor, J. N. (2021). Let's be honest about election forecasting. *PS: Political Science & Politics*, 54(1):107–110.

Vijayakumar, V. and Nedunchezhian, R. (2012). A study on video data mining. *International journal of multimedia information retrieval*, 1:153–172.

Von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

Von Schoultz, Å. and Papageorgiou, A. (2021). Policy or person? the electoral value of policy positions and personal attributes in the finnish open-list system. *Party Politics*, 27(4):767–778.

Wall, A. (2021). Open list proportional representation: The good, the bad and the ugly. International IDEA & Friends' Asia & the Pacific Online Lecture No. 1. Accessed: 02 June 2021.

Walther, D. (2015). Picking the winner(s): Forecasting elections in multiparty systems. *Electoral Studies*, 40:1–13.

Wang, R., Ji, W., Liu, M., Wang, X., Weng, J., Deng, S., Gao, S., and Yuan, C.-a. (2018). Review on mining data from multiple data sources. *Pattern Recognition Letters*, 109:120–128.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Wattenberg, M. P. (2016). The declining relevance of candidate personal attributes in presidential elections. *Presidential Studies Quarterly*, 46(1):125–139.

Whyte, C. E. (2016). Thinking inside the (black) box: Agenda setting, information seeking, and the marketplace of ideas in the 2012 presidential election. *New Media & Society*, 18(8):1680–1697.

Williams, L. V. and Reade, J. J. (2016). Forecasting elections. *Journal of Forecasting*, 35(4):308–328.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

Wu, J., Gan, W., Chen, Z., Wan, S., and Philip, S. Y. (2023). Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Zhao, Y., Xu, X., and Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International journal of hospitality management*, 76:111–121.

Zittel, T. (2016). The personal vote. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 668–687. SAGE Publications, United Kingdom, 1 edition.

Zuiderveen Borgesius, F., Trilling, D., Möller, J., Bodó, B., De Vreese, C. H., and Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review. Journal on Internet Regulation*, 5(1).

Zuo, Z. and Zhao, K. (2018). The more multidisciplinary the better?–the prevalence and interdisciplinarity of research collaborations in multidisciplinary institutions. *Journal of Informetrics*, 12(3):736–756.

Åsa von Schoultz (2016). Party systems and voter alignments. In Arzheimer, K., Evans, J., and Lewis-Beck, M. S., editors, *The SAGE Handbook of Electoral Behaviour*, volume 2, pages 30–55. SAGE Publications, United Kingdom, 1 edition.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2025

ISBN 978-952-02-0188-3 (PRINT)
ISBN 978-952-02-0189-0 (PDF)
ISSN 2343-3159 (PRINT)
ISSN 2343-3167 (ONLINE)