UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

KELLYTON DOS SANTOS BRITO

MACHINE LEARNING FOR NOWCASTING ELECTIONS IN LATIN AMERICA BASED ON SOCIAL MEDIA ENGAGEMENT AND POLLS

Recife
2021

KELLYTON DOS SANTOS BRITO

MACHINE LEARNING FOR NOWCASTING ELECTIONS IN LATIN AMERICA BASED ON SOCIAL MEDIA ENGAGEMENT AND POLLS

Tese apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de concentração: Inteligência Computacional

Orientador (a): Prof. Paulo Jorge Leitão Adeodato

Recife
2021

Catalogação na fonte
Bibliotecário Cristiano Cosme S. dos Anjos, CRB4-2290

B862m Brito, Kellyton dos Santos
Machine learning for nowcasting elections in latin america based on social media engagement and polls / Kellyton dos Santos Brito. – 2021.
170 f.: il., fig., tab.

Orientador: Paulo Jorge Leitão Adeodato.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2021.
Inclui referências.

1. Inteligência Computacional. 2. Redes sociais. 3. Eleições. 4. Aprendizado de máquinas I. Adeodato, Paulo Jorge Leitão (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2021 – 97

Kellyton dos Santos Brito

MACHINE LEARNING FOR NOWCASTING ELECTIONS IN LATIN AMERICA BASED ON SOCIAL MEDIA ENGAGEMENT AND POLLS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de concentração: Inteligência Computacional.

Aprovado em: 18/03/2021.

BANCA EXAMINADORA

Profa. Dra. Patricia Cabral de Azevedo Restelli Tedesco
Centro de Informática / UFPE

Prof. Dr. Adriano Lorena Inacio de Oliveira
Centro de Informática / UFPE

Prof. Dr. Marijn Janssen
Faculty of Technology, Policy and Management
Delft University of Technology

Prof. Dr. Rafael Ferreira Leite de Mello
Departamento de Computação / UFRPE

Prof. Dr. Jarley Palmeira Nóbrega
Centro de Tecnologias Estratégicas do Nordeste / MCTI

Dedico este trabalho a meus pais Josélia e Brito, que se dedicaram a me dar a melhor educação possível, a minha amada esposa Débora, e a meu avô Domingos, nascido em 1894, quando a escravidão no Brasil era uma dura realidade.

ABSTRACT

Contemporary social media (SM) represents a new communication paradigm and has impacted politics and electoral campaigns. The mobilization of the Arab Spring social movements was attributed to SM platforms, as well as successful electoral campaigns such as those of Obama and Trump in the U.S. (2008, 2012, and 2016), the Brexit campaign in 2016, and the Bolsonaro campaign for the Brazilian presidency in 2018. Within this new scenario, the advantages of collecting SM data over traditional polling methods include the huge volume of available data, the high speed, and low costs. Hence, researchers are endeavoring to discover how to use SM for nowcasting election results. However, despite the alleged success, the most-common approach, based on counting the volume of mentions on Twitter and conducting a sentiment analysis, has been frequently criticized and challenged. On the other hand, recent approaches based on other SM platforms and on the advances of machine learning (ML) may be promising alternatives. In this context, this thesis aims to advance the state-of-the-art on predicting elections based on SM data. It proposes a new set of SM performance metrics to be input features for the ML techniques by changing the focus onto the number of people paying attention to the candidates. The defined metrics may be used not only with the most commonly-used current SM platforms (i.e., Facebook, Instagram, and Twitter) but even with future platforms which have not yet gained popularity. In addition, this thesis defines SoMEN, the **So**cial **M**edia framework for **E**lection **N**owcasting, a framework composed of a process and model for nowcasting election results based on the SM performance features and using ML approaches. It proposes well-defined steps, ranging from election understanding to prediction evaluation, and an ML model for predicting the final election results based on an ensemble of artificial neural networks (ANN) trained with SM metrics as features and offline polls as labeled data. It also defines SoMEN-DC, an execution strategy for SoMEN that enables continuous prediction during the campaign (DC). The proposed metrics and framework were applied on the most recent main presidential elections in Latin America: Argentina (2019), Brazil (2018), Colombia (2018), and Mexico (2018). More than 65,000 posts were collected from the SM profiles on Facebook, Twitter, and Instagram of the candidates, as well as data from 195 presidential polls. Results demonstrated that

the defined metrics presented a high correlation with the final share of votes in all the studied countries. Moreover, it was also possible to achieve a high level of accuracy in predicting the final vote share of the candidates, with competitive or better results than traditional polls. Lastly, despite the difficulty in measuring the quality of predictions during the campaign, results are promising and also competitive to polls. The strategies put forward in this thesis have attempted to handle several among the current challenges in this research area and indicate a new manner on how to face the problems. Furthermore, they may be directly used for nowcasting future elections in similar scenarios.

Keywords: social media; elections; machine learning; artificial neural networks; facebook; twitter; instagram.

RESUMO

As redes sociais contemporâneas representam um novo paradigma de comunicação e têm impactado a política e as campanhas eleitorais. A mobilização dos movimentos sociais da Primavera Árabe foi atribuída às redes sociais, assim como o sucesso de campanhas eleitorais como as de Obama e Trump nos Estados Unidos (2008, 2012 e 2016), o Brexit em 2016, e a campanha de Bolsonaro no Brasil em 2018. Neste novo cenário, as vantagens de coletar os dados das redes sociais sobre os métodos de pesquisa eleitoral tradicionais incluem a grande quantidade de dados disponíveis, a alta velocidade e baixo custo de coleta. Consequentemente, pesquisas estão sendo realizadas para usar as redes para prever os resultados eleitorais. Apesar do suposto sucesso da abordagem mais comum, baseada na contagem do volume de menções no Twitter combinada com análise de sentimento, esta tem sido frequentemente criticada e contestada. Por outro lado, novas abordagens baseadas em outras redes e nos avanços do aprendizado de máquina podem ser alternativas promissoras. Nesse contexto, esta tese objetiva avançar o estado da arte na previsão de eleições baseada em dados das redes sociais. Ela propõe um novo conjunto de métricas de desempenho nas redes, mudando o foco para o número de pessoas prestando atenção aos candidatos. As métricas definidas podem ser usadas tanto com as redes sociais mais populares atualmente (Facebook, Instagram e Twitter), quanto com plataformas futuras que ainda não ganharam popularidade. Esta tese também define o SoMEN (**So**cial **M**edia framework for **E**lection **N**owcasting), um framework composto por um processo e modelo para previsão das eleições baseado no desempenho nas redes sociais e usando abordagens de aprendizado de máquina. Ele propõe etapas bem definidas, que vão desde o entendimento da eleição até a avaliação das previsões, e um modelo para prever os resultados finais da eleição com base em um conjunto (*ensemble*) de redes neurais artificiais treinadas com as novas métricas de performance como variáveis e as pesquisas tradicionais como dados rotulados. Também definimos a SoMEN-DC, uma estratégia de execução para o SoMEN que permite a previsão contínua durante a campanha (DC). As métricas e o framework proposto foram aplicados nas principais eleições presidenciais mais recentes na América Latina: Argentina (2019), Brasil (2018), Colômbia (2018) e México (2018). Mais de 65.000 posts

foram coletados dos perfis dos candidatos no Facebook, Twitter e Instagram, bem como dados de 195 pesquisas eleitorais. Os resultados demonstraram que as métricas definidas apresentaram alta correlação com o percentual de votos obtido pelos candidatos em todos os países estudados. Além disso, foi obtido um alto nível de precisão na previsão do percentual final de votos dos candidatos, com resultados competitivos ou melhores do que as pesquisas tradicionais. Por fim, apesar da dificuldade em medir a qualidade das previsões durante a campanha, os resultados são promissores e competitivos com as pesquisas. As estratégias propostas nesta tese levaram em consideração os principais desafios desta área de pesquisa e apresentam uma nova maneira de enfrentá-los. Além disso, elas podem ser usadas diretamente para prever eleições futuras em cenários semelhantes.

Keywords: redes sociais; eleições; aprendizado de máquinas; redes neurais; facebook; twitter; instagram.

LIST OF FIGURES

Figure 2.1 – Projection of the world's most used SM platforms 34
Figure 2.2 – Example of a linear regression plotted function 43
Figure 2.3 – The McCulloch-Pitts (MCP) neuron model 45
Figure 2.4 – MLP model with one hidden layer 46
Figure 2.5 – The GRNN architecture and calculations 51
Figure 3.1 – Characteristics of studied elections: (a) coverage, (b) role, (c) type of vote, and (d) number of candidates. 62
Figure 5.1 – SoMEN Process 95
Figure 5.2 – SoMEN instantiation 103
Figure 5.3 – The SoMEN-DC execution 103
Figure 6.1 – Data modeling for the statistical tests 131
Figure 6.2 – Polls and predictions in Argentina: Fernandez and Macri, predictions using the MLP-BP PCA model and the final vote share 134
Figure 6.3 – Polls and predictions in Brazil: Bolsonaro and Haddad, predictions using the MLP-BP PCA model and the final vote share 135
Figure 6.4 – Polls and predictions in Colombia: Márquez and Petro, predictions using the MLP-BP PCA model and the final vote share 135
Figure 6.5 – Polls and predictions in Mexico: Obrador and Cortés, predictions using the MLP-BP PCA model and the final vote share 136

LIST OF TABLES

Table 2.1 – Most used SM platforms in the most populous Latin American countries 36
Table 3.1 – Main Approaches Identified 63
Table 3.2 – Average Success Rate 66
Table 3.3 – Summary of Main Challenges on this research area 72
Table 5.1 – Description of the metrics related to Facebook, Twitter and Instagram 94
Table 6.1 – Election Timelines 108
Table 6.2 – List of Candidates 108
Table 6.3 – Most used SM platforms in the analyzed countries 109
Table 6.4 – Candidate profiles on Facebook, Twitter, and Instagram 110
Table 6.5 – Overview of the collected data from the posts of Argentinian candidates 111
Table 6.6 – Overview of the collected data from the posts of Brazilian candidates 112
Table 6.7 – Overview of the collected data from the posts of Colombian candidates 112
Table 6.8 – Overview of the collected data from posts of the Mexican candidates 113
Table 6.9 – Comparison of engagement on SM and the total population 114
Table 6.10 – The pollsters and number of polls collected for each country 115
Table 6.11 – Summary of the final week of polls 116
Table 6.12 – Summary of all polls 116
Table 6.13 – An example of the generated dataset 118
Table 6.14 – Values for the ANN grid search parameters 119
Table 6.15 – Pearson correlation results for the SM performance and the Argentinian election results 122
Table 6.16 – Pearson correlation results for the SM performance and the Brazilian election results 123
Table 6.17 – Pearson correlation results for the SM performance and the Colombian election results 123
Table 6.18 – Pearson correlation results for the SM performance and the Mexican election results 124
Table 6.19 – Predictions for the Argentinian elections 125
Table 6.20 – Error metrics for the Argentinian predictions 126
Table 6.21 – Predictions for the Brazilian elections 126
Table 6.22 – Error metrics from the Brazilian predictions 127
Table 6.23 – Predictions for the Colombian elections 127
Table 6.24 – Error metrics from the Colombian predictions 128
Table 6.25 – Predictions for the Mexican elections 128
Table 6.26 – Error metrics from the Mexican predictions 129
Table 6.27 – MAE errors of all countries, compared with the respective poll average 130
Table 6.28 – Wilcoxon signed rank test between the election results, and the predictions and polls 131

Table 6.29 – Wilcoxon signed rank test between the errors obtained with the prediction and polls 132
Table 6.30 – MAE comparing predictions and polls. This shows how close the predictions are to the polls 137

SUMMARY

1 INTRODUCTION 15
1.1 MOTIVATION 17
1.2 OBJECTIVES 18
1.3 THESIS OVERVIEW 20
1.4 OUT OF SCOPE 22
2 RESEARCH BACKGROUND 24
2.1 THE EVOLUTION OF ELECTION POLLING AND PREDICTIONS 24
2.1.1 Approaches to Election Predictions 26
2.1.2 Measuring Prediction Accuracy 27
2.1.3 Poll Data Collection 30
2.2 THE EMERGENCE OF SOCIAL MEDIA 33
2.2.1 Social Media Newsfeed Algorithms and the Bubble Effect 36
2.2.2 Social Media Data Publishing and Gathering 38
2.3 MACHINE LEARNING REGRESSION 41
2.3.1 Linear Regression 41
2.3.2 Artificial Neural Networks – ANNs 44
2.3.2.1 Multilayer Perceptrons – MLPs 44
2.3.2.2 General Regression Neural Networks – GRNN 49
2.3.3 Committee Machines 52
3 PREDICTING ELECTIONS WITH SOCIAL MEDIA DATA 54
3.1 RESEARCH BACKGROUND 55
3.1.1 The Rise of Election Prediction with SM Data 55
3.1.2 Analysis of Previous Reviews 57
3.2 METHODOLOGY 58
3.2.1 Research Questions 58
3.2.2 Search Process 59
3.2.3 Quality Assessment 60
3.3 REVIEW RESULTS 60
3.3.1 Quality Assessment 61
3.3.2 Electoral Contexts 61
3.3.3 Main Approaches 61
3.3.3.1 Volume or Sentiment 62

3.3.3.2 Regression or Time Series 63
3.3.3.3 Profile or Post Interactions 64
3.3.3.4 Topic or Event Detection 65
3.3.3.5 Other Approaches 65
3.3.4 Main Characteristics of Successful Studies 65
3.4 DISCUSSION OF MAIN STRENGTHS, CHALLENGES AND FUTURE DIRECTION 67
3.4.1 Main Strengths of Predicting Elections with Social Media Data 68
3.4.2 Main Challenges of Predicting Elections with Social Media Data 68
3.4.3 Future Directions 73
3.4.3.1 Future Directions in Process Definitions 73
3.4.3.2 Future Directions in Model Definitions and Sampling 73
3.4.3.3 Future Directions in Evaluation 74
3.5 CONCLUDING REMARKS 75
4 PROBLEM DEFINITION AND METHODOLOGY 78
4.1 RESEARCH QUESTIONS AND HYPOTHESES 78
4.2 RESEARCH METHODOLOGY 81
4.2.1 Rejecting H_1' 81
4.2.2 Rejecting H_2' 82
4.2.3 Rejecting H_3' 83
4.3 CONCLUDING REMARKS 85
5 PROPOSALS 87
5.1 DOMAIN CHALLENGES 87
5.1.1 Social Media Challenges 88
5.1.2 Political Scenario Challenges 88
5.2 ENGAGEMENT METRICS FOR MEASURING SOCIAL MEDIA PERFORMANCE 90
5.3 SOMEN: A SOCIAL MEDIA FRAMEWORK FOR ELECTION NOWCASTING 94
5.3.1 The SoMEN Process 95
5.3.1.1 Election Understanding 95
5.3.1.2 Data Collection and Understanding 97
5.3.1.3 Data Preparation 97
5.3.1.4 Modeling 99

5.3.1.5 Evaluation 100
5.3.2 The SoMEN Model 101
5.4 SOMEN-DC: A SOCIAL MEDIA FRAMEWORK FOR ELECTION NOWCASTING DURING THE CAMPAIGN 103
5.5 CONCLUDING REMARKS 104
6 EXPERIMENTS – LATIN AMERICAN PRESIDENTIAL ELECTIONS 106
6.1 THE SOMEN EXECUTION 107
6.1.1 Election Understanding 107
6.1.2 Data Collection and Understanding 109
6.1.2.1 Data from SM Platforms 109
6.1.2.2 Data from Polls 114
6.1.3 Data Preparation 117
6.1.4 Modeling 119
6.1.5 Evaluation 120
6.2 THE SOMEN-DC EXECUTION 121
6.3 EXPERIMENT RESULTS 121
6.3.1 Research Question 1 121
6.3.2 Research Question 2 125
6.3.3 Research Question 3 133
6.4 RESULTS DISCUSSION AND COMPARISON 137
6.4.1 Research Area Challenges Addressed 137
6.4.2 Comparison with Related Works 139
6.5 LIMITATIONS AND VALIDITY DISCUSSION 142
6.6 CONCLUDING REMARKS 144
7 CONCLUSION 146
7.1 CONTRIBUTIONS 147
7.2 PUBLICATIONS ARISING FROM THIS WORK 149
7.3 FUTURE RESEARCH 150
REFERENCES 153

1 INTRODUCTION

“The secret of getting ahead is getting started. The secret of getting started is breaking your complex overwhelming tasks into small manageable tasks, and starting on the first one.” (Mark Twain)

This thesis sets out to investigates and defines a set of metrics for measuring performances on social media (SM), to discover correlations between these performance metrics and the electoral performance of presidential candidates, and to define a process and model for predicting elections by using SM data as input features and polls data as labeled data.

Contemporary social media platforms are new: Facebook was launched for public access in 2006, Twitter debuted in 2006, and Instagram emerged in 2010. Despite being a novelty, SM denotes a new communication paradigm, representing one of the greatest social innovation/revolution in communication history, fundamentally altering the way humans communicate and the practice of public relations, journalism, advertising, marketing, and business (KENT; LI, 2020). Indeed, communication has changed from an era in which citizens were viewed as consumers of information from large-scale media publishers, such as newspapers, magazines, TV, and radio. Nowadays, through the use of SM platforms, ordinary people with very few resources may be information producers, thereby reaching a larger audience and sometimes becoming, to use a contemporary term, digital influencers (KHAMIS; ANG; WELLING, 2017).

This new communication context has also impacted politics. Initially, by enabling a greater degree of political discussions, SM has helped to foster democratic processes and civic and political participatory behaviors, both online and offline (GIL DE ZÚÑIGA; JUNG; VALENZUELA, 2012). Moreover, political discussion has advanced towards the organization of collective action, thus attributed a critical role to SM connectivity in the Arab Spring social movements and anti-government protests (KHONDKER, 2011) that led, for example, to the resignation of the Egyptian leader (ELTANTAWY; WIEST, 2011).

As a natural evolution, politicians began to use the capabilities of SM to change political campaigns, moving from a scenario in which politicians were chiefly heard speaking at campaign rallies, and on TV or radio, and extra information regarding them was mainly obtained through the press. Hence, citizens had very few opportunities to actually confront politicians. Currently, mediated by SM, politicians no longer have geographic or time constraints since they may use their SM profiles to post content at anytime, anywhere, and to everyone. Thus, any additional information on them may be obtained not only through the press, but directly from their profiles and through other people sharing them on SM. Moreover, ordinary people may use SM platforms to obtain direct contact with politicians, amplify their voice by sharing content, ask questions, confront them and obtain direct responses.

The first presidential campaign identified as being firmly based on SM platforms was the 2008 Barack Obama campaign in the U.S. (BIMBER, 2014; COGBURN; ESPINOZA-VASQUEZ, 2011), through Facebook, Twitter, MySpace, e-mails, an iPhone application, and two websites. His campaign capability to translate online activity to on-the-ground activism (COGBURN; ESPINOZA-VASQUEZ, 2011) inspired the adoption of SM as a permanent campaign platform across the entire world.

After the Obama campaign, the success of many others has been attributed to their online campaigns, such as those for Brexit (HALL, W.; TINATI; JENNINGS, 2018) and Trump in the 2016 U.S. presidential elections (FRANCIA, 2018; HALL, W.; TINATI; JENNINGS, 2018). In the 2018 Brazilian presidential election, Bolsonaro, the candidate with the most followers on SM, but practically no time on TV, ran his campaign almost entirely online and was elected, while Alckmin, a candidate with more TV time and fewer SM followers, ended in fourth place.

Within this context, researchers have not only begun to investigate how candidates have used SM platforms to support their campaigns but also to predict election results based on SM data. Thus, a new research subject has been initiated.

In parallel, over the past two decades the ongoing exponential increase in the availability of online data and low-cost computation, together with the development of new learning algorithms and theory, has led to the wide adoption of machine learning (ML) methods and techniques (JORDAN; MITCHELL, 2015). ML addresses the question of how to build computers that automatically improve through experience (JORDAN; MITCHELL, 2015). This has been used in a number of huge, complex data-intensive fields such as medicine, astronomy, biology, hydrology, finance, and

economics (ARDABILI; MOSAVI; VÁRKONYI-KÓCZY, 2020; QIU *et al.*, 2016), since these techniques provide possible solutions to mine information hidden in data, which has impacted greatly on science and society (RUDIN; WAGSTAFF, 2014). Thus, ML methods and techniques are natural candidates for dealing with the challenge of predicting elections based on SM data.

1.1 MOTIVATION

Thus, the work described in this thesis began with a real-world application in mind: the nowcasting of elections based on SM data[1]. This application is not important only to satisfy human curiosity in knowing, in advance, who will be elected. If a prediction could be made during the actual campaign, it would be important for the candidates to be able to know the impact and response of their campaigns and daily actions, almost in real-time, and therefore adjust their campaigns accordingly. It is also important for enterprises, since the perspective of one candidate winning over another may directly impact their planning, activities, and perspectives for the future, as well as the impact this may have on the stock market.

Prediction through polling is an old discipline but has been contested since it began. Academics consider that the advent of modern scientific polling came in the U.S. presidential elections of 1936, after the first well-known polling crisis (CROSSLEY, 1937), when prestigious pollsters forecasted a victory for Landon, while others as yet unknown pollsters, such as George Gallup, correctly indicated a victory for Roosevelt. After this, there have been many other examples of “polling crisis”, such as that of 1948 (MOSTELLER *et al.*, 1949), which led to an in-depth analysis of polling methods and metrics, in 1996 (MITOFSKY, 1998) and later (ABRAMOWITZ, 2004).

The main theoretical advantages of using SM data for gathering information on citizens, over traditional polling methods, are related to its reach, speed, and low cost. Face-to-face and telephone-based surveys are only able to reach just a small fraction of the population, needs time to be performed and incur heavy costs. In practice,

[1] The term ‘nowcasting’ is used in this thesis along with prediction, since our predictions are an estimation of the present or a very near future, as many voters only decide who to vote for in the last few days leading up to elections, or even on the day itself. Thus, we consider that we are also “predicting the present”, and therefore use both terms as synonyms.

although in countries such as the U.S. traditional polls are conducted almost on a daily basis, in Latin American countries such polls are scarce and are mostly concentrated towards the end of campaigns. On the other hand, SM platforms were actively used by 51% of the world population in 2020 (WE ARE SOCIAL; HOOTSUITE, 2020), and data may be gathered and processed from these platforms for a fraction of the cost of traditional polls, and almost in real time.

This scenario reveals SM as a natural candidate for being used as input data, and ML approaches, capable of mining information hidden in data, may be a promising approach for performing predictions. However, this is not a traditional scenario of ML use. Despite the availability of SM data, its use presents a number of challenges, such as the possibility of being affected by volume manipulation and the rapidly changing SM landscape. The political scenario also presents new challenges for ML, such as the lack of available historical data, since elections usually take place every four or five years, and the existence of just one actual labeled data, the final vote share, which is the aim of prediction.

This subject area is still in its infancy, and the most common approach, based on counting the volume of mentions on Twitter and conducting a sentiment analysis, has been frequently criticized and challenged. Indeed, since its beginning, different researchers applying the same approach in the same context may have achieved opposite results (JUNGHERR; JÜRGENS; SCHOEN, 2012; TUMASJAN *et al.*, 2010), and the same researchers applying the same approach in different contexts may have achieved contradictory results (ANJARIA; GUDDETI, 2014; GOTO; GOTO, 2019).

1.2 OBJECTIVES AND METHODOLOGY

The main objective of this thesis is *to define a process and create an ML model based on the SM performance of candidates, which is capable of making daily nowcasting and final predictions of election results with competitive results to traditional polls.*

This objective was defined after an extensive study of the state-of-the-art of predicting elections based on SM data. In the study, we identified that a generalizable and repeatable process, as well as the use of modern nonlinear ML approaches, would address most of the current challenges.

Based on this main objective, three research questions were defined:

- *RQ1: Is there a correlation between the SM performance of candidates and their electoral performance?*
- *RQ2: Is it possible to define a process and create an ML model capable of predicting election results based on the SM performance of candidates?*
- *RQ3: Is it possible to define a process and create an ML model capable of performing daily nowcasting of election results based on the SM performance of candidates?*

These research questions will be answered sequentially. For RQ1, we identified that, as yet there is no clear definition of the SM performance, and therefore defined a new set of metrics to measure it. This new set of metrics is based on Zajonc's exposition theory (MURPHY; ZAJONC, 1993; ZAJONC, 2001, 1968), who hypothesized that "mere repeated exposure of the individual to a stimulus object enhances his attitude toward it". Thus, we focused on the number of people actively paying attention to the candidates by interacting with their profiles on SM platforms through likes, shares and comments on candidates' posts. It is worth noting that, due to the algorithms of SM platforms, more people interacting with candidates' profiles lead to the content being shown to even more people, in a snowball effect. Also, defined metrics may be used not only with the most commonly-used current SM platforms (i.e., Facebook, Instagram, and Twitter) but even with future platforms which have not yet gained popularity. Thus, we have attempted to discover correlations between the SM performance of candidates and their electoral performance.

For RQ2, we defined a process and model for prediction, called SoMEN – Social Media Framework for Election Nowcast, and attempted to nowcast only the final elections results. The process is based on the cross-industry standard process for data mining –CRISP-DM (SHEARER, 2000), and consists of the following phases: (i) election understanding, (ii) data collection and understanding, (iii) data preparation, (iv) modeling and execution, and (v) evaluation. For the ML approach, two models were chosen: MLP-BP and GRNN, and linear regression was used as a baseline model. The SM performance metrics were used as features, traditional poll data was considered "imprecise ground truth" and used as labeled data for training the models, and an ensemble of 10 predictors, whereby each one received different input data based on different observable windows of SM data, were used for predictions. Lastly, for evaluation, traditional metrics in the poll domain were used, as well as the statistical analysis of results.

For RQ3, we defined SoMEN-DC, Social Media Framework for Election Nowcast – During Campaign, this is an execution strategy for SoMEN to perform daily nowcasting. This execution strategy consists of design decisions regarding the minimal amount of data for starting predictions, and strategies for continuously updating the model as new data is released. Lastly, specific ways for measuring and comparing daily predictions were used.

This is a multidisciplinary thesis, and involves knowledge from the areas of polling and electoral predictions, social media studies, and machine learning. In addition, although not detailed in this text, there is also a need for knowledge on software engineering in order to develop information systems for data gathering on SM platforms, since public datasets with this data are unavailable. This study may be considered as an application study, and not a performance study, very common in artificial intelligence. This signifies that we studied the basis, proposed and analyzed a novel manner (using SM and ML) for dealing with an old problem (nowcasting elections) as opposed to increasing the performance of current methods and techniques, such as a new method for sentiment analysis.

In this context, we highlight contributions in three areas. For **machine learning**, this thesis study and apply SM data and ML approaches for a new, yet underexplored context, the prediction of electoral results. This context presents very specific challenges, both related to the SM and electoral scenarios. Thus, we design and apply a framework composed of a process and ML model for predictions. For the **social media** area, we study how to model SM performance and the existence of correlations between online behavior and offline outcomes. Lastly, for the subject of **electoral predictions**, we present a new approach to estimate citizen preferences, which is able to complement traditional polling or even to be incorporated into the poll methodology.

1.3 THESIS OVERVIEW

This thesis is organized into a total of seven chapters. In this chapter, the motivations for carrying out this research were presented together with a brief overview of the objectives and research questions.

Chapter 2 presents the background of the main areas essential for the development of this thesis: the evolution of electoral predictions, the emergence of social media, and machine learning regression techniques. In the background to the

electoral predictions, in addition to a brief historical overview, the main approaches to polling and predictions are presented together with the main metrics for measuring prediction accuracy. For the background to SM, we characterize the two types of platforms, newsfeed and conversation platforms, discuss the SM newsfeed algorithms and the bubble effect, and end with a presentation of the main ways of publishing and collecting SM data. As the main ML techniques, we present the basis of three ML regressors: linear regression, which is used as a baseline technique for comparisons, the multilayer perceptron (MLP) neural network, and the general regression neural network (GRNN). Lastly, we present the basis of committee machines.

Chapter 3 presents a shortened version of a systematic review on predicting elections based on social media data. In the review, 90 studies were analyzed: 83 were strictly focused on predicting elections based on SM, the focus of this thesis, and seven surveys. The analysis identified challenges in many areas, such as process, sampling, modeling, performance evaluation, and scientific rigor. The main findings included the low success rate of the most-commonly used approach, namely volume and sentiment analysis on Twitter, and the best results with new approaches, such as regression methods trained with traditional polls. Lastly, a vision of future research is also discussed regarding advances in process definitions, modeling, and evaluation, indicating, amongst other items, the need for better investigations into the application of more sophisticated machine learning approaches. The results of this review are the central point for the proposals of this thesis.

Chapter 4 presents the main goal of this thesis, followed by a presentation and discussion of the three defined research questions. Furthermore, three null hypotheses and their alternative hypotheses are also presented. Thus, the research methodology for rejecting the null hypotheses is also defined.

Chapter 5 presents the proposals of the thesis for achieving the main goals stated in the previous chapter. Following a discussion on the domain challenges, a new set of SM performance metrics is presented, based on the exposition theory. In addition, the **so**cial **m**edia framework for **e**lection **n**owcast (SoMEN) is defined, including the process and model, followed by a presentation of the SoMEN-DC framework, the **so**cial **m**edia framework for **e**lection **n**owcasting **d**uring the **c**ampaign.

Chapter 6 describes the experiments conducted with data from the most recent major Latin American presidential elections: Argentina (2019), Brazil (2018), Colombia (2018) and Mexico (2018). All phases of the defined processes are conducted, and the

main decisions are documented. The results are also presented and discussed, and the research questions are answered by the rejection of the null hypotheses. A discussion is presented regarding how the challenges identified in Chapter 3 were addressed, as well as direct comparisons with related works. Lastly, the limitations of the study and the validity are discussed.

Finally, Chapter 7 presents the concluding remarks and discusses the main contributions of this thesis, and directions are outlined for possible future research.

1.4 OUT OF SCOPE

The intersection between the domains of elections and social media may generate a variety of studies with many different focuses. These studies may range from technical issues, including the software engineering of collecting data from social systems and machine learning techniques for predicting elections, to social and philosophical issues regarding the use and impact of SM platforms. Thus, it is necessary to clarify certain subjects that are not addressed in this thesis.

- The popularization of fake news (MUSTAFARAJ; METAXAS, 2017), including its detection, spread, and impact on election results, is not addressed. It is assumed that the impact of these news items is captured by the variations in interactions within the candidates' profiles on SM platforms.
- Similarly, the occurrence of true campaign events and scandals are not directly addressed since their impact is also captured by the variations in interactions on the candidates’ profiles.
- The analysis of the content of posts on SM, such as the sentiment analysis of posts, is not addressed due to the definition and use of new metrics based on engagement.
- As an important task of this thesis, a fully functional information system was developed to enable data collection from SM platforms. Despite its importance, as this thesis is not focused on software engineering, the system is only briefly cited in Chapter 6, and its development is not detailed in this text.

- Essentially political, sociological, and psychological issues are not considered.

2 RESEARCH BACKGROUND

“Mountains of data are available for those who wish to examine the views of the American people on specific issues, or to judge the reliability of modern polling methods” (George Gallup, 1965)

This thesis is a multidisciplinary research, involving election polling and predictions, which emerged as a scientific method in 1936, contemporary social media (SM), with the appearance of Facebook and Twitter in 2006, and machine learning (ML), whereby the first “artificial neuron” was modeled using electrical circuits in 1943, but became popular and widely used in the 2010s with the increase of computing power. This chapter introduces these three areas, presents their fundamentals and brief evolution, and highlights the main concepts used in this thesis.

2.1 THE EVOLUTION OF ELECTION POLLING AND PREDICTIONS

As stated by Hillygus (HILLYGUS, 2011), “*Public opinion polls are now conducted on every topic under the sun–everything from presidential approval to celebrity outfits and sports predictions–but they remain specially fundamental to the conduct and study of elections.*” Although currently opinion polls are widely used, it is nonetheless an old discipline. Academics consider that the advent of modern scientific polling appeared in 1936, when prestigious pollsters, such as Literary Digest, conducted straw polls of millions of people to indicate a victory for Landon in the U.S. presidential elections, and others, such as George Gallup, who conducted a quota-controlled survey, correctly signposted Roosevelt’s victory (CROSSLEY, 1937).

Motivated by the level of criticism directed towards the polling errors of that particular election, Crossley (CROSSLEY, 1937) classified and grouped the various polling methods of the time and studied the reliability of each, and was considered a seminal study in this area. At that time, polls were conducted by mail, by personal interview, or a combination of both. Most were performed on huge randomly selected samples, but some were performed on small scientifically distributed samples. Some also used the cumulative method, adding its results together into a single report, and

others used interval sampling, sampling at several points during a campaign. The analysis found that polls with better results were conducted wholly or in part by personal interviews, with small samples, and repeated at frequent intervals. The polls with the worst results were those conducted through the postal service, a huge sample, and only provided one complete report accumulated over a period of several months.

The study highlighted what was considered at the time to be the ideal poll: (i) it needed to be flexible, not based on dated mailing lists, and designed so that it could be readily adjusted if new information became available during its course; (ii) a fairly small sample would work properly in all but close states; (iii) the distribution of the sample was of paramount importance; and (iv) it should not be cumulative, but repeated in similar cross-sections at intervals to show trends. Focus was then given to presenting practical guidance on the segmentation of population sampling. These principles are still valid today.

Only 12 years later, in 1948, another unexpected result in the U.S. challenged the polling industry yet again, when the candidate widely tipped as favorite obtained only 189 electoral votes and 45.1% of the popular vote, against 303 electoral votes and 49.6% of the popular votes of the elected candidate. After a strong adverse reaction, a distinguished group of social scientists and statisticians mounted an intensive review of the election polling procedures and results to evaluate "*the technical aspects of public opinion and the problem of throwing some light on the nature and magnitude of the discrepancies between poll predictions and election results. Until readers and users of poll results understand these errors and their role in the interpretation of results, there will continue to be adverse reaction every time a pre-election poll fails to pick the winner, no matter how small the error between a poll percentage and an election percentage.*" (MOSTELLER *et al.*, 1949) This initiative produced a report (MOSTELLER *et al.*, 1949), and the chapter “Measuring the Error” established the measures that have been used ever since to evaluate the accuracy of election polls. In reality, researchers such as Gallup himself, have advocated that errors are usually small, but much criticism has been levied due to misinformation regarding the methods and results of predictions (GALLUP, 1965).

Since then, election predictions have evolved into three approaches, presented as follows.

2.1.1 Approaches to Election Predictions

Researchers have recognized three main approaches for predicting elections: polling, statistical forecasting models, and political stock markets.

In **polling**, researchers directly ask a sample of people questions like “If the election were held tomorrow, which candidate would you vote for?” or “Regardless of your personal preferences, whom do you think will win the upcoming election?”. After this, the vote share percentages in the responses are taken to forecast the final vote shares. This approach began with the work of Gallup (CROSSLEY, 1937), and is the most well known and most commonly-used approach across the world. It is firmly based on selecting a representative sample of citizens and other design decisions—regarding the mode, timing, sampling method, question formulation, weighting, etc.—so as to enable extrapolation from a few answers to the entire population. Thus, each of these methodological decisions may potentially bias the results, thereby leading to many different results from many different pollsters for the same election.

By recognizing that individual poll results are subject to many potential biases, the aggregation of many different polls has become popular, and is referred to as polling aggregation or polling the polls (BLUMENTHAL, 2014; HILLYGUS, 2011; JACKMAN, 2005). Aggregating polls helps to reduce volatility in polling predictions and improves the precision of estimates by increasing the sample. However, the variety of methods for aggregation, ranging from simple averages to weighted averages based on polling samples or weighting by accuracy or detected bias on previous elections, are also seen as a challenge.

Statistical forecasting (LEWIS-BECK, 2005), also called macroeconomic models (HILLYGUS, 2011), began to appear around 1980 (FAIR, 1978; LEWIS-BECK; RICE, 1982; SIGELMAN, 1979). At the core of these models is the assumption that vote share is a function of other indicators, such as government performance, economic performance, economic growth, or incumbent popularity, to cite but a few. A typical equation of this model would be along the lines of equation 2.1.

$$\textit{Incumbent Vote} = \textit{Incumbent Popularity} + \textit{Economic Growth} + \textit{error} \quad \textit{(Eq. 2.1)}$$

Statistical models may also take into account polling results, such as incumbent popularity. An example from the U.S. presidential elections was offered by (LEWIS-BECK, 2005) and is presented in equation 2.2.

$$V = 37.31 + 0.28P + 1.32E + e$$

(Eq. 2.2)

Where V = presidential party share in the election; P = presidential popularity, in the July Gallup Poll of the election year; E = percentage growth in the real GNP over the first two quarters of the election year; and e = error. In the U.S., the most studied country, only the two most prominent candidates/parties are usually considered. Thus, calculating the incumbent (party) vote is sufficient to also have the percentage of the challenger as *challenger = 1 – incumbent*. A general theory is that voters reelect incumbents in good (economic) times and do the opposite in bad times (FIORINA, 1978). Thus, the debate largely regards which economic indicator to be used—either job growth, GDP growth, inflation rate, or perceptions of personal finances (ABRAMOWITZ, 2004; HOLBROOK, 2004; LEWIS-BECK, 2005)—and also regards the inclusion of other variables in the statistical model, such as polling numbers.

The third approach, the **political stock market**, considers how traders invest money in candidates running for office. The Iowa Electronic Market is the leading example of this type of forecasting. In this market, people buy and sell candidate futures based on whom they judge as being more likely to win, and a candidate's investment share provides a vote forecast for the candidate (RHODE; STRUMPF, 2004). This approach may be considered a specialized polling approach because it polls the opinion of traders. Although traders are not representative of likely voters and they do not even need to be eligible to vote, traders are just assumed to be making informed judgments, and have the confidence to wager their money on it. As a result, the benefits of this approach are under debate. Berg et al. (BERG *et al.*, 2008) argued that markets are more accurate than polls, but Erikson (ERIKSON; WLEZIEN, 2008) has challenged this conclusion.

2.1.2 Measuring Prediction Accuracy

After the 1948 polling crisis, the report produced by Mosteller et al. (MOSTELLER *et al.*, 1949) established eight different metrics for assessing polling

accuracy. Seven measures are related to the differences between election results and poll estimates, and the eighth is related to the difference between the poll estimations of participation and the real election participation. Mitofsky (MITOFSKY, 1998) summarized the measures, thus:

1. The difference in percentage points between the leading candidates' share of the *total vote* from a poll and from the actual vote.
2. The difference in percentage points between the leading candidate's share of the *major party* vote from a poll and from the actual vote. (The major parties are Democratic and Republican and are assumed to be the top two vote getters.)
3. The average (without considering the sign) of the percentage point deviation for each candidate between his/her estimate and the actual vote.
4. The average difference (without considering the sign) between a ratio for each candidate and the number one, where the ratio is defined as a candidate's estimate from a poll divided by the candidate's actual vote.
5. The difference between two differences, where the first difference is the estimate of the vote for the two leading candidates from a poll and the second difference is the election result for the same two candidates.
6. The maximum difference in percentage points between a party and the actual vote.
7. The chi-square to test the congruence of the estimated and actual vote distributions.
8. The difference between the predicted and actual electoral vote.

Scholars state that two of these measures, the “Mosteller Measure 3” and the “Mosteller Measure 5”, have been widely used ever since in order to evaluate the accuracy of elections polls (HILLYGUS, 2011; JENNINGS; WLEZIEN, 2018; MARTIN, 2005).

The “Mosteller Measure 3” is the average absolute error on all candidates between the prediction and the actual results. In statistics, it is widely known as the mean absolute error (MAE), and is mathematically defined as:

$$MAE = \frac{\sum_{c=1}^{n} \left(|v_c - p_c| \right)}{n}$$

(Eq. 2.3)

Where ***n*** is the number of candidates, ***c*** is the candidate from 1 to ***n***, v_c is the actual vote share result of the candidate ***c***, and p_c is the predicted result for the candidate ***c***.

The “Mosteller Measure 5”, which we call the absolute error on the margin (AEOM) is the absolute value of the difference between the margin separating the two leading candidates in the poll and in the actual vote. Mathematically it may be defined as:

$$AEOM = |(v_1 - v_2) - (p_1 - p_2)|$$

(Eq. 2.4)

Where v_1 is the actual vote share of the candidate with the most votes, p_1 is the predicted vote share for this candidate, v_2 is the actual vote share of the candidate with the second most votes, and p_2 is the prediction for this candidate.

These two metrics are very well suited for measuring the accuracy of the results of the U.S. elections because attention is mostly focused on the two main candidates. In an election with more candidates, although the main attention is also usually focused on the main candidates, there may be many candidates, and sometimes there is a close race for the second or third positions. Thus, Measure 4 is also relevant because it calculates the ratio of error regarding the candidate himself, and reveals that an error of 1 percentage point is high for a candidate who received just 2 percent of the vote, but fairly low for a candidate who received 50 percent of the vote. This metric is known in statistics as the mean absolute percentage error (MAPE) and is mathematically defined as:

$$MAPE = \frac{\sum_{c=1}^{n} \left(\frac{|v_c - p_c|}{v_c} \right)}{n}$$

(Eq. 2.5)

where ***n*** is the number of candidates, ***c*** is the candidate from 1 to ***n***, v_c is the actual vote share result of the candidate ***c***, and p_c is the predicted result for the candidate ***c***.

There has been little follow-up work by statisticians to improve Mosteller’s measures, but in 2005, Martin et al. (MARTIN, 2005) proposed a new measure of predictive accuracy (A) of election polls that enables both accuracy and bias to be examined. The metric is based on the odds ratio of a poll compared to the actual

outcome. For party candidate **c**, receiving ***p*** as a proportion of the poll share and ***v*** as a proportion of the vote share, this measure takes the form:

$$A'_i = \ln\left(\frac{p_i}{1 - p_i}\ x\ \frac{1 - v_i}{v_i}\right)$$

(Eq. 2.6)

The interpretation of results is simple for two-party elections: *A* is zero when there is a perfect agreement between a poll and an election result, and a significantly positive or negative value of *A* indicates that a poll is biased for the first or second candidate, respectively. As it fits perfectly for two-candidate elections, it is necessary to calculate *A* just for one candidate. Martin claims that the main advantage of this metric, in addition to calculating the bias of the poll, is that it is comparable across elections with different outcomes and amongst polls that vary in their treatment or numbers of undecided voters. Although Martin argued that the measure might be adapted for use in multi-party elections and the measure was extended by (ARZHEIMER; EVANS, 2014) with this aim, its definition, use, and properties are clear in two-party elections, and the interpretation of results in multi-party elections is still challenging.

These metrics, with emphasis on the MAE, were used in a recent study (JENNINGS; WLEZIEN, 2018), which aimed to assess prediction errors in pre-election polls. The analysis drew on more than 30,000 national polls from 351 general elections in 45 countries between 1942 and 2017. The study observed that, contrary to conventional wisdom, the recent performance of polls has not been out of the ordinary. By calculating the estimated average poll during the week prior to the elections, they discovered that the MAE was 2.1% during the 1940s and 1950s (in the early days of polling), 2.1% during the 1960s and 1970s, and since 2000, has been 2.0%. They also found that errors were higher in presidential elections (an average of 2.7 percentage points) than in legislative elections (1.8 percentage points). Lastly, as expected, they reported that errors decrease as the elections approach.

2.1.3 Poll Data Collection

From the seminal work of (CROSSLEY, 1937) until more recent research (BRICK, 2011), there has been a common agreement that the method and treatments

of data collection directly influences poll results. Historically, the survey research method has called the types of poll data collection “modes.” In early times, modes were only conducted by mail and by personal interviews. Today, while personal interviews are still common in certain places, the postal service is no longer used, and new modes, such as phone calls, internet polling, and collecting information on social media have been introduced.

In 2011 Couper (COUPER, 2011) reviewed the history and recent trends in modes of survey data collection, with a view to speculating on the future. The remainder of this section is strongly based on this review. After describing that the main modes of data collection from the 1940s to the 1970s were mail and face-to-face surveys, it goes on reveal how after 1970 telephone surveys were widely adopted in the U.S., and then later in Europe. This mode of collection was popularized due to the increasing spread of telephone coverage at that time, lower costs, the fact that is was quicker than personal interviews, and existing research, which demonstrated that the quality of data obtained was comparable to face-to-face surveys. Later, in the 1990s, internet surveys began to threaten the dominance of telephone surveys, mainly because of the advantages in terms of speed and costs. In parallel, modes evolved from paper questionnaires to computer-assisted interviewing and from interviewer-administered surveys to self-interviewing, which demonstrated advantages for sensitive questions or those subject to the effects of social desirability effects. Computer-assisted self-interviewing (CASI) then became popular, with, for example, the interactive voice response (IVR), administered by telephone or modern versions administered by the internet.

Whereas Web surveys may be viewed as a single-mode, there are many manners in which it may be implemented. It may be performed as interviewer-administered surveys over a specific population sample or viewed as replicating mail surveys, by being sent to many people without segmentation and gathering a fraction of responses sent back. Finally, Couper argued that all approaches may be used in a mixed-mode: either face-to-face or mail, due to the feasibility of address-based sampling; by telephone, which may use address-based probability methods with landline-based phone calls or nonprobability methods with mobile phone calls; and by the internet, using nonprobability samples. Hence, since each of the modes has its own challenges and source of bias, the strengths and weaknesses of different modes may compensate one another.

In the early days of internet, it was claimed that web surveys would replace telephone surveys and possibly all interviewer-administered surveys, as discussed by (LEEUW, 2005). However, internet modes have their own challenges. For example, the internet may not be considered a reliable sample of the whole population, and also requires some level of literacy, thereby limiting the generalizability for certain kinds of studies. It is also hard to implement a sample design, and internet polling requires the initiative of respondents, which leads to selective samples. Moreover, due to the commoditization of polls and the variety of pollsters implementing automated methods, there is greater competition for the attention of respondents, who are becoming a scarce resource, raising concerns about nonresponse and bias and the appearance of professional respondents (SINGER; YE, 2013).

In addition to reviewing academic research on the modes of survey data collection, we also performed a practical exercise of investigating methodologies of pollsters in recent elections, both in the U.S. and in Latin America.

We analyzed data from the 2016 U.S. elections published by the Huffington Post poll aggregator site[2]. The site shows that polls were conducted by live phone calls, automated phone calls, by the internet, and in a mixed-mode. It was not possible to obtain more details regarding the internet mode of the majority of polls, except for NBC News[3], which were performed in partnership with SurveyMonkey, a well-known online survey company. Respondents of the survey were self-selected from nearly three million people who take part in surveys on the SurveyMonkey platform each day. Thus, they considered that due to the bias of self-selection for participation, "*no estimates of sampling error could be calculated, and the survey was subject to multiple sources of errors, including but not limited to sampling error, coverage error, and measurement error.*"

In Latin America, the scenario is different, with face-to-face interviews predominating. We collected methodological data from polls relating to the most recent presidential elections in the four most populous countries (Argentina, Brazil, Colombia, and Mexico). The description of procedures is described in Chapter 6. Brazil and

[2] Available at: https://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton. Viewed on November 17, 2020.

[3] Available at: https://www.scribd.com/upload-document?archive_doc=330243723. – Viewed on November 17, 2020.

Colombia only used face-to-face and phone polls, live or automated. The methods were usually not mixed. In Argentina, two pollsters were web-based, and participants were self-selected. One of their websites[4] advertises "*answer our online surveys, earn points and exchange them for prizes and discounts on the best brands.*" More modes however were observed in Mexico, where the most popular approaches were the use of the secret ballot, simulating an election vote, and automated phone calls. Nonetheless, the internet was also used: one pollster performed interviews on Facebook, and another applied surveys on Facebook by advertising the poll with paid propaganda, and respondents self-selected in order to participate. It should be noted that one pollster in Colombia also measured mentions and number of followers on SM but did not use this data in predictions of vote share.

As conclusions, and pointing to the future, the rise of SM sites such as Facebook and Twitter present huge changes regarding how the internet may be used, and the way in which data is collected. This new media may be used not only as a new mode for traditional data collection, such as electronic questionaries, but also as a rich source of information. For example, it is possible to gather comments, or the repercussion of comments, from people concerning a particular subject or to attempt to discover the political inclination of people through their posts. The following section focuses on the emergence of social media and its new possibilities.

2.2 THE EMERGENCE OF SOCIAL MEDIA

Contemporary SM systems are new. Facebook was first launched in 2004[5], made accessible to Harvard students, who were able to post photographs of themselves and personal information about their lives, such as their class schedules and the clubs they belonged to. Its popularity and features increased and it was launched to the public in 2006, the same year as Twitter[6]. Twitter was initially defined as a "microblogging" service, where users could post content limited to 140 characters, to allow posts through Short Message Service (SMS), although since then its

[4] Available at https://www.ohpanel.com. Viewed on November 17, 2020.

[5] More information regarding the Facebook history may be found at https://www.britannica.com/topic/Facebook and https://www.brandwatch.com/blog/history-of-facebook /

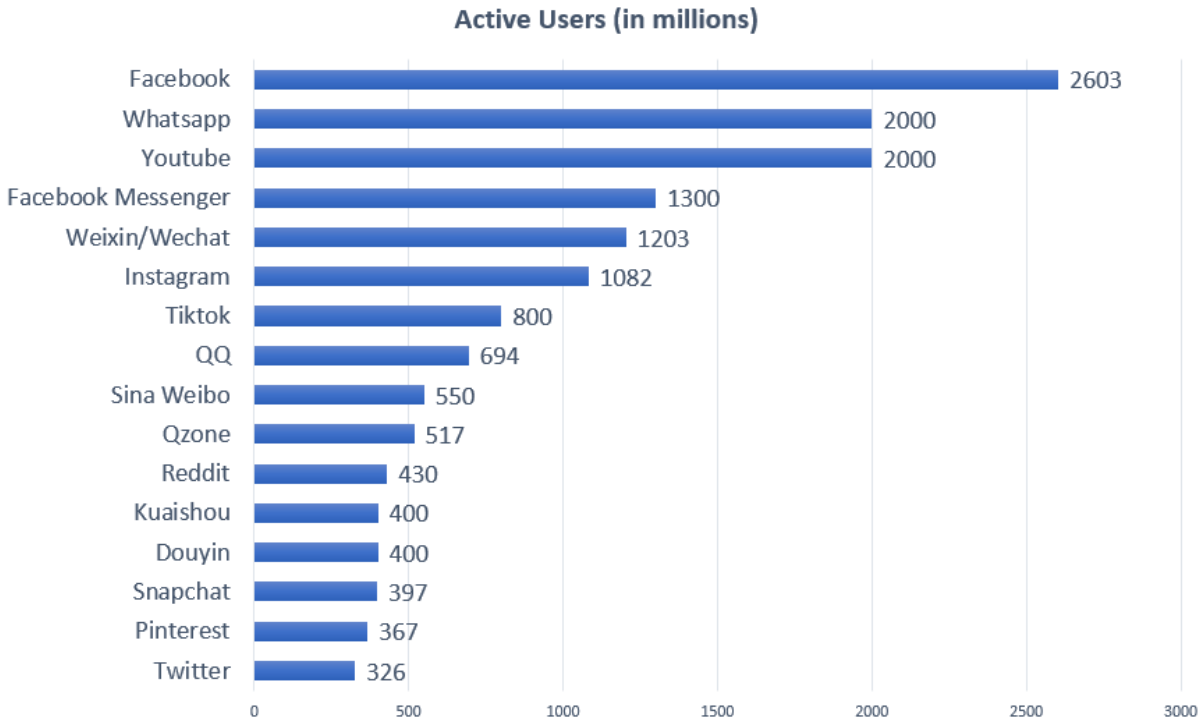
[6] More information regarding the Twitter history may be found at https://www.britannica.com/topic/Twitter

functionalities have evolved.

A recent report from July 2020 mentioned that more than half, roughly 51 percent, of the global population uses social media platforms (WE ARE SOCIAL; HOOTSUITE, 2020). The report estimated that 4.57 billion people (59% of the world population) are internet users, 3.91 billion (51% of the population) are active users of social media, and 99% of them access SM via mobile phones, but not exclusively. Whilst the growth of the global population in one year was 1.1%, the growth of internet users was 8.2%, and active social media users increased by +10.5%, demonstrating an increase in the penetration of SM. Users spend on average 2h 22m per day using SM, have on average 8 social media accounts, and 40% of them use SM for work purposes.

According to the report, the most used social platforms are Facebook, Youtube, Whatsapp, Facebook Messenger, Weixin/WeChat, and Instagram. Figure 2.1 presents a projection of the world's most-used SM platforms, based on monthly active users, active user accounts, or addressable advertising audiences (in millions).

Figure 2.1 – Projection of the world's most used SM platforms



Source: adapted from (WE ARE SOCIAL; HOOTSUITE, 2020).

Despite the high levels of usage, the concept of SM has not been well defined. The main feature shared across all platforms is to enable people to connect with one another and to send or receive content to and from each other. However, the

abovementioned SM platforms may be placed into two groups: newsfeed platforms and direct conversation platforms.

On **newsfeed platforms**, also considered wall-based platforms, one of the main features consists of the pair posts/newsfeed. People are encouraged to post contents, which may contain text, photos, audio, video, links, or a combination of them all. People are also encouraged to make connections with other people, such as friends or colleagues, and to subscribe to follow updates of accounts which interest them. Thus, items posted by users are shown to his/her connections (friends or followers) in a newsfeed format, one after the other. Users are able to interact with the content, usually either by clicking like (signaling that he/she liked the content), by commenting on the post, or by sharing it with his/her own connections, thereby amplifying the reach of the message.

As a monetization strategy, platforms usually add paid advertisements between the posts. Platforms also use optimization algorithms in order to choose which content would be more relevant to show to the user and in which order (these algorithms are discussed in more detail in Section 2.2.1) in an infinite scrolling newsfeed: as the user scrolls the screen, more posts are loaded. The main platforms that adopt this approach are Facebook, Instagram, Twitter, and Youtube each with its own slight differences.

On **conversation platforms**, the main feature consists of allowing direct conversation between two (or a group of) people. People may also send content, such as text, photos, audio, video, and links, but posts are directed towards specific people or groups of people, and they all receive the content without moderation by the platform. Generally, these platforms also allow audio or video calls. The main platforms of this category are Whatsapp and Facebook messenger.

The usage of SM platforms varies across the world. Many of them are extremely popular in Asia, such as WeChat, QQ, and Sina Weibo, but are almost unknown in Latin America. Table 2.1 shows the most used platforms in the four most populous Latin American countries, Argentina, Brazil, Colombia, and Mexico, according to reports produced by the same company (KEMP; WE ARE SOCIAL; HOOTSUITE, 2020b, 2020d, 2020c, 2020a).

According to this data, the most common newsfeed-based platforms in Latin America are Youtube, Facebook, Instagram, and Twitter, and the most popular conversation platforms are Whatsapp and Facebook Messenger. In this study we focus on newsfeed platforms. Thus, in the next subsection, we briefly discuss the wall

algorithm and the bubble effect intrinsic to these platforms, followed by a presentation of the official ways that data may be collected from them.

Table 2.1 – Most used SM platforms in the most populous Latin American countries

	Brazil	Mexico	Colombia	Argentina
1st	Youtube	Facebook	Youtube	Youtube
2nd	Facebook	Youtube	Facebook	Whatsapp
3rd	Whatsapp	Whatsapp	Whatsapp	Facebook
4th	Instagram	Facebook Messenger	Instagram	Instagram
5th	Facebook Messenger	Weixin / Wechat	Facebook Messenger	Facebook Messenger
6th	Twitter	Instagram	Twitter	Twitter

Source: self-provided.

2.2.1 Social Media Newsfeed Algorithms and the Bubble Effect

On SM platforms based on a newsfeed, people may connect to, or just follow, a few to thousands of other accounts. Thus, since it is impossible to show users all the posts of all their connections at once, platforms use optimization algorithms to choose which content to show (and when) on their news feed, in order to maximize showing them what is probably of most interest to them.

Youtube explains that the home screen is the platform's "*personalized best guess at what each viewer may want to watch … including videos that are new, watched by similar viewers, or from user subscriptions.*" (YOUTUBE INC., 2017) Not all videos from user subscriptions are shown at their home feed, but in a specific tab. The selection is based on two items: performance, or how well a video has engaged and satisfied similar viewers, and personalization, based on a viewer's watch and search history.

Other newsfeed-based SM platforms work in a similar manner. Facebook states that "*The goal of News Feed is to deliver the right content to the right people at the right time so they don't miss the stories that are important to them. Ideally, we want News Feed to show all the posts people want to see in the order they want to read them.*" (LARS BACKSTROM; THE FACEBOOK, 2013) Thus, the news feed algorithm takes into account the following measures:

- How often the user interacts with the friend, page, or public figure (such as an actor or journalist) who posted;

- The number of likes, shares, and comments a post receives from the world at large and particularly from user friends;
- How much the user has interacted with this type of post in the past;
- Whether or not the user and other people across Facebook are hiding or reporting a given post;

Thus, to summarize, user feed prioritizes in order to show popular content to users, as well as content that corresponds to what they have previously liked.

Initially, these algorithms help users by showing them what they probably want to see. However, this approach has received much criticism because it leads to filter bubbles. As users are only presented to content that they already like and are familiar with, possible new viewpoints and opposite ideas are filtered out, leading to a false sense of unanimity. Extensive research into the downside of this scenario is currently being performed by social scientists (FLAXMAN; GOEL; RAO, 2016; GESCHKE; LORENZ; HOLTZ, 2019; SPOHR, 2017).

Because the main feature of newsfeed-based SM is to post about some topic, SM platforms may be viewed as places where it is possible to gather public opinion. This argument becomes stronger when it is considered that half the world's population is currently using SM platforms (WE ARE SOCIAL; HOOTSUITE, 2020). Thus, SM allows researchers to gather a huge amount of data from an unpredictable number of people at a low cost. Amongst other comments, Jungherr et al. (JUNGHERR *et al.*, 2017) has argued extensively that "digital trace data"—data produced by people while interacting with digital services—may have a high potential for studying public opinion, while Beauchamp has also argued that social media data could track representative measures of public opinion (BEAUCHAMP, 2017).

A huge amount of data is produced by SM platforms, and collecting this data may also be a challenge. As a matter of privacy, direct conversation platforms do not usually allow any kind of data collection other than the direct conversation of the author (who is the only individual capable of collecting the data) with other people, strongly limiting its use in social behavior studies. However, this limitation does not occur on newsfeed-based platforms. Next, we overview the official methods of data collection from the newsfeed-based platforms mostly used in Latin America (Facebook, Instagram, Twitter, and Youtube).

2.2.2 Social Media Data Publishing and Gathering

The publishing of data by SM platforms is an important part of the entire ecosystem of SM. This publishing allows the development of third-party applications, partially or fully dependent on SM data, thereby extending their original features. Third-party applications range from social games, that allow friends to play together, to entire marketing platforms that define, execute, and track the performance of media campaigns by thousands of companies.

However, publishing this kind of data is a sensitive topic. SM platforms may collect data from more than half the entire population of the world, and leads to complex questions of privacy and data manipulation, such as guarantees on the potential release of data to unintended recipients and the use of user data by the service provider.

In 2013, we conducted a study (BRITO, K. S. *et al.*, 2013) to discover how people care about their personal data released on SM. We discovered that, amongst other results, people do not generally read the licensing terms and know very little about service policies. However, when presented with these policies people often disagree with them. We also discovered that older people are more concerned about their data than younger people. A combination of the increasing use of SM, people not knowing how their data is being used, the lack of concern by the young regarding the use of their data, the extensive publishing of data, and finally, the malicious use by certain companies, has led to problems such as the Cambridge Analytica scandal (ISAAK; HANNA, 2018)(BERGHEL, 2018). In this particular episode, the company used data analytics to sway the electorate, relying on the participation of SM platform users in their own psychological manipulation.

Before the scandal, the most used newsfeed-based SM platforms in Latin America (Facebook, Instagram, Twitter, and Youtube) already had well-defined methods for publishing data collected from their platforms, and used application programming interfaces (API's) for publishing. After the scandal, platforms toughened their policies, and their data publishing functionalities and processes in the years 2019/2020 are described below. It is important to highlight that these policies undergo constant changes, and by the time readers have access to this text it may already be out of date.

Facebook policies also apply to Instagram (FACEBOOK INC., 2020). Data access is based on the Graph API, which allows other systems, called APPs, to connect with it. It "*is the primary way to get data into and out of the Facebook platform. It's an HTTP-based API that apps can use to programmatically query data, post new stories, manage ads, upload photos, and perform a wide variety of other tasks.*" In order to have access to Graph API, developers must register an APP on the platform and provide information on the APP and their business, such as an URL to the APP, privacy and service terms policies, contact e-mail, and others. Thus, the APP receives access to the basic data, which in practice, is data from the developer's own accounts. If access to data from other accounts is required, it has to pass through a verification process: the developer must file a form containing information regarding why the APP needs that access and how it uses the data; a video must be recorded showing the working APP, and highlighting points of data gathering and use; and valid credentials must be created to allow Facebook representatives to log on the system and verify the information by themselves. If it passes, legal information concerning the company responsible for the app must be sent, including a legal document (for example, in Brazil it may be the company's "Social Contract" or tax documents). The entire process may take from between one week to several months. These exigencies make it difficult for the Facebook platform, both Facebook and Instagram data, to be accessed by researchers, since it is not allowed to obtain data for offline analysis, which is the main purpose of many researchers.

The Facebook API functionalities work with the concept of nodes, basically individual objects, such as a User, a Photo, a Page, or a Comment; edges, which are the connection between a collection of objects and a single object; and fields, data on an object. Thus, while it is possible to find Users or Pages, collect their posts, comments in their posts, or similar activities, it is not permitted to perform an open search by keywords. Also, the API limits the number of queries that an APP may perform by the number of users of the APP.

YouTube and Twitter allow access to their data in a simpler manner. The YouTube API (GOOGLE LLC, 2020) limits data access by volume. For example, APPs have almost no restrictions on accessing YouTube API if it performs less than 10,000 requests per day (it is slightly less, but the manner in which a request is counted was simplified for this text. More information on this may be found at (GOOGLE LLC, 2020)). If an app needs more requests, it has to undergo a verification process similar

to the Facebook process. The developer must also create a video and provide valid credentials to allow YouTube representatives to log on to the system and verify the information by themselves. As an example, in our last verification process, YouTube staff asked us to change one YouTube icon inside the application because it had the wrong background color: we had changed the shade of red to be more harmonious with the system.

The YouTube features work in a similar manner to Facebook. While it allows APPs to search for videos, channels, and playlists, it does not permit an open search by keywords in comments. It also allows APPs to collect the data and comments of posts.

For Twitter API (TWITTER INC., 2020), users have almost no restrictions if their requests are below the rate limits. APPs may perform 180 query requests through windows of 15 minutes by APP users. Until 2018, it was unnecessary to provide Twitter with information regarding the APP context. After this date, developers have had to apply for a developer account. For this, they have to complete a form to have the APP accepted, although the form is sufficient to be able to access, since no videos or credentials for Twitter staff is needed.

Some of the Twitter features are different to other platforms. Twitter also allows the search for users, gathering all user posts and related information based on User and Tweet objects. It differs however from other platforms in a number of ways. First, it allows APPs to perform an open search on the platform. For example, an APP may make a query for “Social Media” and receive the tweets that contains the term. However, due to the huge number of tweets and information that may be retrieved in this way, some peculiarities exist on this platform. First, it serves traditional query requests and also streams connections to serve APPs with real-time tweets. Thus, the query for “Social Media” may create a connection with Twitter servers and receive the tweets in real-time. However, for the basic, free access, the Twitter platform does not guarantee that all tweets will be retrieved. In fact, for the standard query, they limit results to the past 7 days and return just a sample of all the tweets, as the official documentation states “*The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.*”[7] At the time of completing this thesis in 2020,

[7] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview/standard

Twitter is launching a new version of its API and creating new access possibilities. Thus, this information may vary.

Chapter 3 describes how most of the current research on using SM for predicting elections results is based on the open search approach for Twitter posts containing the names of candidates or parties. By knowing the differences in the possibilities of data gathering on the most used SM platforms, we argue that most studies gather data on Twitter not because it is the most used or the most representative SM platform, but merely because it is the easiest or only platform where the open search approach is possible. Thus, only analyzing Twitter data may not represent a good sample of the SM population. Moreover, another bias in current studies becomes clear, since returned tweets on Twitter queries are not even representative of all tweets since queries return only a sample of recent tweets.

2.3 MACHINE LEARNING REGRESSION

There are dozens of regression techniques, and new techniques are constantly being created or refined. A recent extensive experimental survey of regression methods (FERNÁNDEZ-DELGADO *et al.*, 2019) evaluated the performance of 77 popular regression methods over 83 datasets and is a good reference regarding the comparison of methods. This thesis does not intend to create a new technique nor to find the best technique for the problem of predicting elections with SM data, but to study, select and apply a plausible, adequate technique for the problem and compare the results with traditional polls and previous studies. Thus, the question of discovering the best technique or model is for future studies.

This section reviews the regression techniques used in this thesis. The choice is discussed in Chapter 5. This study has used the traditional linear regression model as a baseline technique, and artificial neural networks (ANNs), multilayer perceptron (MLP) and general regression neural networks (GRNN), as suitable models. The three techniques are presented below.

2.3.1 Linear Regression

Linear regression is one of the oldest techniques (dating from 1805 according to (YAN; GANG SU, 2009)) and is the most studied linear approach for modeling the

relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands, usually denoted by *y*) and the predictors (also called independent variables, explanatory variables, control variables, or regressors, usually denoted by $x_1, x_2, \ldots, x_p$). This has been well documented in the statistics literature (SEAL, 1967; YAN; GANG SU, 2009) and may assume multiple forms. One explanatory variable is called a simple linear regression, or otherwise a multiple linear regression. It is called a multivariate linear regression when multiple correlated dependent variables are predicted, rather than a single scalar variable.

The simple regression model is often written in the following form

$$y = \beta_0 + \beta_1 x + \varepsilon \,,$$

(Eq. 2.7)

where y is the dependent variable, β_0 is the intercept, β_1 is the gradient or the slope of the regression line, *x* is the independent variable, and ε is the random error. The multiple linear regression model has one dependent variable and more than one independent variable, and its general form is as follows:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \,,$$

(Eq. 2.8)

where y is the dependent variable, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients, and $x_1, x_2, \ldots, x_n$ are independent variables in the model. In both cases, in the classical regression setting it is usually assumed that the error ε follows the normal distribution with $E(\varepsilon) = 0$ and a constant variance $\mathrm{Var}(\varepsilon) = \sigma^2$.

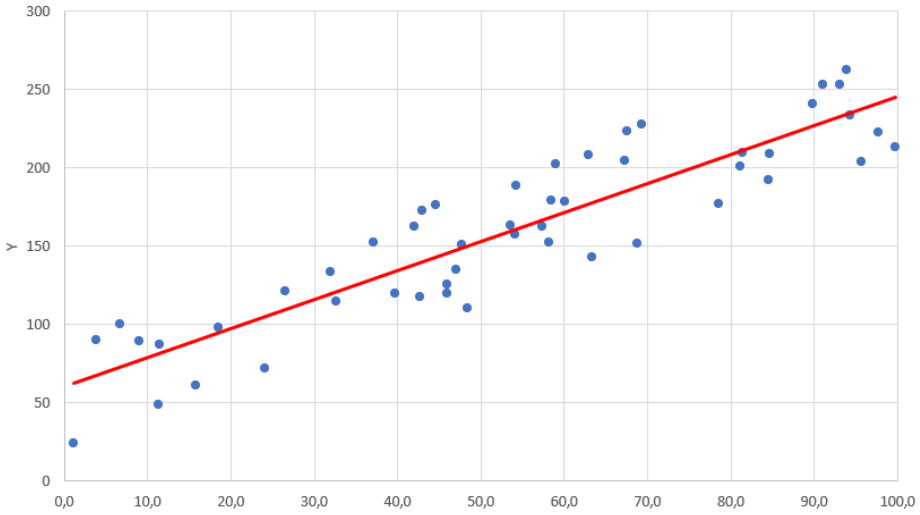
Despite also being considered as a machine learning technique, linear regression may be considered an optimization technique. The typical experiment of the linear regression is to observe *n* tuples of data $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ and use an optimization function to find the best values of β in order to minimize errors. The baseline method for this estimation is the least square estimation. The principle for the estimation is to find the estimates b_0 and b_1 such that the sum of the squared distance from actual response y_i and predicted response $y_i = \beta_0 + \beta_1 x_i$ reaches the minimum amongst all possible choices of regression coefficients β_0 and β_1, i.e.,

$$(b_0, b_1) = \arg\ min_{(\beta_0, \beta_1)} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

(Eq. 2.9)

For the multiple linear regression, the principle is similar. The motivation behind the least squares method is to find parameter estimates by choosing the regression line that is the “closest” line to all data points (x_i, y_i). A general example is shown in Figure 2.2, where the red line plots the best function $y = 60 + 1{,}85x$ found for generated values plotted in blue. More details about linear regression may be found in (YAN; GANG SU, 2009).

Figure 2.2 – Example of a linear regression plotted function



Source: self-provided.

It is important to highlight two essential characteristics of this method. First, it is a linear method, which signifies that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. To deal with this restriction, manipulation of input variables may be performed, such as the use of polynomial regression. This leads to new concerns, such as the overfitting of the data and the need for regularization terms. Second, the coefficient estimates for the ordinary least squares rely on the independence of the features, i.e., a linear correlation may not exist between two or more independent variables. When this occurs, the least-squares estimate becomes overly sensitive to random errors in the observed target, thereby producing a large variance. These kinds of restrictions lead to the creation of variations of the basic linear regression, such as ridge and lasso regression, and others, as well as nonlinear models. Further details on linear regression and these variations may be found at (EFRON *et al.*, 2004; HAWKINS, 1973; TIBSHIRANI, 1996; YAN; GANG SU, 2009).

2.3.2 Artificial Neural Networks – ANNs

There are many more sophisticated models for regression than linear regression, that provide better results in many different datasets (FERNÁNDEZ-DELGADO *et al.*, 2019), especially when there is a nonlinear correlation between dependent and independent variables. One of the prominent classes of these models is the artificial neural networks (ANNs). This class of computational model is inspired by the functioning of the human brain and designed to solve problems that traditional computing does not perform well, including perception problems such as face and speech recognition (HAYKIN, 1998)(BRAGA; CARVALHO; LUDERMIR, 2007). The main characteristics of ANNs include their highly distributed nature and the ability to "learn" from previous examples.

In this study, two ANN approaches are considered for regression problems: the multilayer perceptron (MLP) and the general regression neural networks (GRNN).

2.3.2.1 Multilayer Perceptrons – MLPs

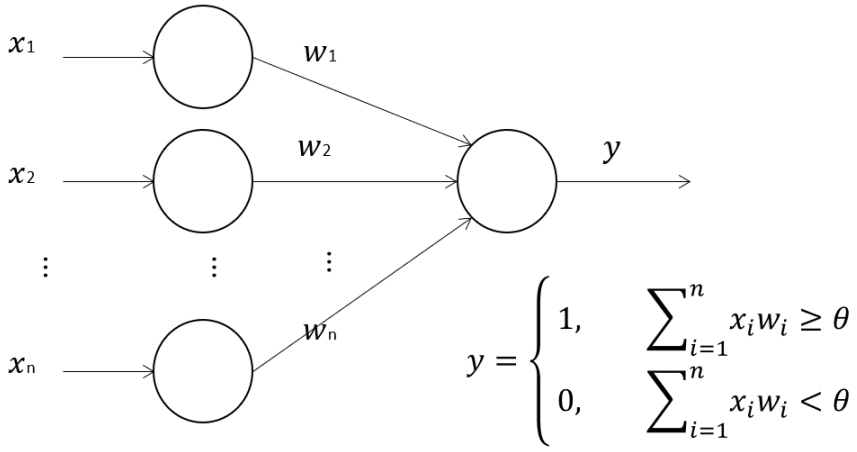
The multilayer perceptron (MLP) is one of the most popular neural network architectures and may be used both for classification and regression. It is based on the neuron model, defined in 1943 by McCulloch and Pitts (MCCULLOCH; PITTS, 1943), known as the MCP neuron model. The model was used in 1958 by Rosemblatt (ROSENBLATT, 1958), who defined the perceptron and for the first time, introduced the *learning* concept and the training algorithm.

McCulloch and Pitts assumed that "*because of the 'all-or-none' character of nervous activity, neural events and the relations among them can be treated by means of propositional logic.*" Thus, they defined the model by computing the weighted sum of its input and computing the output by using an *activation function*. Originally, the activation function produced only a binary output, 1 if the weighted sum of the inputs is greater than a threshold θ, or 0 otherwise.

Through the evolution to the perceptron, Rosemblatt (ROSENBLATT, 1958) proposed a learning theory by supervised learning. He concluded, amongst other things, that "*in an environment of random stimuli, a system consisting of randomly connected units, subject to the parametric constraints discussed above, can learn to associate specific responses to specific stimuli*"; and that "*the probability that a stimulus*

which has not been seen before will be correctly recognized and associated to its appropriate class (the probability of correct generalization) approaches the same asymptote as the probability of a correct response to a previously reinforced stimulus"; and that "*trial-and-error learning is possible in bivalent reinforcement system*." To summarize, it is possible to create an architecture of connected units with weights, that may be trained in a supervised training approach (by trial-and-error) and is capable of being applied in a new stimulus which has not been seen before. Figure 2.3 is a graphical representation of the model.

Figure 2.3 – The McCulloch-Pitts (MCP) neuron model



Source: adapted from (BRAGA; CARVALHO; LUDERMIR, 2007).

The weights *(ω)* of the MCP neuron are adjusted to adequately solve a given problem. Rosenblatt demonstrated that the neuron may be trained iteratively by adjusting the weights according to the following formula:

$$\vec{\omega}(t+1) = \vec{\omega}(t) + \eta \text{ x } e(t) \text{ x } \vec{x}\,(t)$$

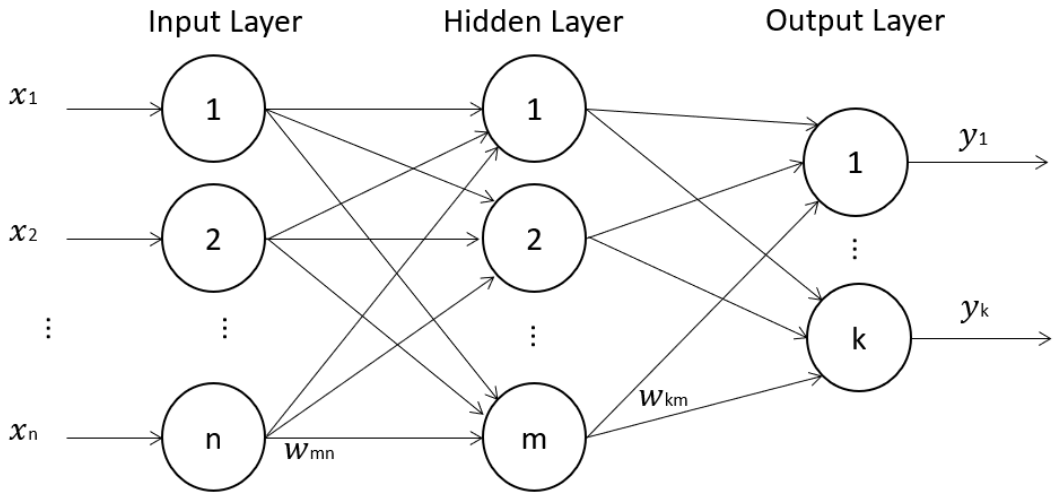
(Eq. 2.10)

Where *e(t)* is the output error, i.e., the difference between the output produced by the network *y(t)* and the actual output *a*; *η* is the learning rate, i.e., the rate at which the weights are adjusted, and $\vec{x}$ is the input vector. It has been proven that this training algorithm always converges to the solution if the problem is linearly separable (HAYKIN, 1998)(BRAGA; CARVALHO; LUDERMIR, 2007), making the approach similar to the linear regression technique.

Since many real-life practical problems are more complex and not linearly separable, this model has evolved to a multilayer perceptron (MLP). The idea behind the MLP is to add at least one hidden layer to the MCP neuron model to build a

multilayer network. Thus, many perceptrons are organized into these layers of nodes, of which there are at least three: an input layer, a hidden layer, and an output layer. The input nodes map the input data, and all other nodes are neurons that use a nonlinear activation function. The activation function is a linear function that maps the weighted inputs to the output of each neuron. The weights for each connection are independent and learned during the training phase. Thus, each neuron responds in a different way for the same input. This new kind of organization allows the neurons of the hidden layer to create hyperplanes, whereas the output neurons combine these hyperplanes in order to build more complex solutions (HAYKIN, 1998)(BRAGA; CARVALHO; LUDERMIR, 2007)(HASSOUN, 2010). Figure 2.4 is a graphical representation of the model with one hidden layer.

Figure 2.4 – MLP model with one hidden layer



Source: adapted from (BRAGA; CARVALHO; LUDERMIR, 2007).

It has been proven that the MLP with a single layer can approximate any continuous functions (CYBENKO, 1989). Indeed, the MLP is considered a universal function approximator (HORNIK; STINCHCOMBE; WHITE, 1989). In practice, MLP networks may be designed and implemented in many different forms, by many configurable parameters, leading to different outputs. The choices include, although not limited to, the architecture of the network (such as the number of hidden layers and the number of neurons in each layer), the activation function, the method for training, and the additional parameters for training, amongst others.

Considering the architecture of the MLP network, due to its proven characteristic, the most common architecture contains only one hidden layer. However, in some cases, the use of more hidden layers may facilitate network training and

provide additional advances, which has led to the development of deep neural networks or deep learning. These models have demonstrated practical results and won numerous contests, and have been applied to problems related to pattern recognition, image interpretation, and speech recognition, among others (SCHMIDHUBER, 2015).

In addition to the architecture, the MLPs are also dependent on other choices. Many activation functions, which map the weighted inputs to the output of each neuron, have already been proposed (KARLIK; OLGAC, 2011), the most common being the logistic sigmoid function, the hyperbolic tangent function, and the rectified linear unit (ReLU) function, presented in equations 2.11-2.13.

$$f(x) = \frac{1}{1 + e^{-x}}$$

(Eq. 2.11)

$$f(x) = \tanh(x)$$

(Eq. 2.12)

$$f(x) = \max(0, x)$$

(Eq. 2.13)

The training method is also very influential. The first successful and one of the most commonly used methods for training MLPs is the backpropagation (RUMELHART; HINTON; WILLIAMS, 1986). This is a supervised training method. The aim is to adjust the network weights in order to reduce the known errors in the final predictions. However, the errors of hidden layers are unknown and must be estimated.

The backpropagation algorithm consists of two phases, a forward and a backward phase. As well described by (OLIVEIRA, 2004):

"In the forward phase a pattern from the training set is presented at the network inputs. Next, the units of the first hidden layer compute their outputs. These outputs are used by the next hidden layer to compute the respective outputs. This procedure is carried out until the output layer units compute their outputs. Now the outputs produced by the network are compared to the desired outputs and the errors are calculated. In the backward phase, these errors are used to adjust the weights of the connections to the output layer. Subsequently, they are used to estimate the error on the hidden layer connected to the output layer, as described above. If the network has more than one hidden layer, this procedure is carried out until the first hidden layer (the one to which the input layer is connected)."

The weight is adjusted to minimize the sum of squared errors (SSE), given by:

$$SSE = \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{k}(d_i^j - y_i^j)^2$$

(Eq. 2.14)

where p is the number of training patterns, k is the number of output units, d_i is the desired value of the i-th output and y_i is the value produced by the network on the i-th output. Thus, the weight and bias are adjusted in each training epoch *t*, according to the equation:

$$w_{ji}(t+1) = w_{ji}(t) + \eta \text{ x } \delta_j(t) \text{ x } x_i(t)$$

(Eq. 2.15)

where η is the learning rate. If *j* is an output neuron, δ_j is given by $\delta_j = (d_i - y_j)f'(net_j)$, where net_j is the weighted sum of the inputs of the neuron and $f'(.)$ is the partial derivative of the activation function with respect to net_j. The complete derivation and explanation of the backpropagation algorithm may be found at (BRAGA; CARVALHO; LUDERMIR, 2007; HASSOUN, 2010; HAYKIN, 1998; RUMELHART; HINTON; WILLIAMS, 1986).

One of the challenges of the MLP trained with backpropagation (MLP-BP) is the right choice of the learning rate η, since it has an important effect on the training performance: if the value is too small, too many epochs are needed to reach an acceptable solution, but if it is too high, the minimum error may possibly not be reached. To deal with this challenge, some strategies are used, such as the use of adaptive learning rate or inverse scaling learning rate. Another important challenge is the local minima problem. These algorithms are designed to find a minimum error, which is not guaranteed to be the global minimum, and may depend on initialization parameters. Thus, they often stop training without being unable to find the global minimum. As a consequence, a number of studies have tackled this problem in various manners, including global optimization technics such as genetic algorithms (STANLEY; MIIKKULAINEN, 2002), by adding nodes on the hidden layer (CHOI; LEE; KIM, 2008), or by using a committee of machines, averaging many different results (TRESP, 2001).

To summarize, the multilayer perceptron trained with backpropagation (MLP-BP) is a powerful machine learning method for supervised learning that is able to solve complex problems stochastically, and is proven to be a universal function approximator (HORNIK; STINCHCOMBE; WHITE, 1989). It is also data driven, so no explicit

assumption is needed for the model between the inputs and outputs; it needs no assumptions on the distribution of input data, unlike statistical techniques; it can generalize, and produces good results even when facing new input patterns; it performs well even with the existence of noisy data; and the algorithm is easy to implement since the functions and its derivatives are well known. Lastly, in the recent experimental survey that compared 77 popular regression methods (FERNÁNDEZ-DELGADO *et al.*, 2019), the MLP-BP based model obtained remarkable results with small datasets by using a specific design with one hidden layer and few neurons on this hidden layer.

However, as challenges, we highlight that for more complex designs a (large) representative training set is needed, and that complex transformations may be costly to converge, and that there is a possibility of converging to a local minimum rather than a global minimum. Moreover, the results are seen as a “black box”, since the learning in the hidden layers has no direct relation with the inputs or outputs and are difficult to explain. Lastly, it may present *overfitting*, when the error is very low in the training set, although the solution is not generalizable, and errors are high for new examples.

For practical implementation, there are many choices and parameters for being tuned, which may lead to different results for the same input data. This need for many decisions leads to a high dependency on the expertise of the designer. As an alternative, research regarding the automatic selection of the model and setting of these parameters, also called hyperparameters, has begun to appear and these are called AutoML (HE; ZHAO; CHU, 2021).

2.3.2.2 General Regression Neural Networks – GRNN

General Regression Neural Networks (GRNN), as proposed by Specht (SPECHT, D.F., 1991), fall into the category of probabilistic neural networks (PNN) (SPECHT, Donald F., 1990). The main difference between PNNs and the other neural networks, such as MLP-BP, is that the activation function is replaced by one that is statistically derived, an exponential function. Hence, the resulting network is similar in structure to backpropagation and has the feature that the decision boundary implemented by the PNN, under certain easily-achieved conditions, asymptotically approaches the Bayes optimal decision surface. In addition, the use of probabilistic neural networks is especially advantageous because the network learns in one pass

through the data and is able to generalize from examples as soon as they are stored (SPECHT, D.F., 1991). Thus, it converges with only a few available training samples and is well suited for small datasets.

GRNNs are feed-forward networks based on the estimation of probability density functions (PDFs), and have a one-pass learning algorithm with a highly parallel structure. The GRNN algorithm is based on the estimation of the PDF from the observed samples using the Parzen window method (SPECHT, D.F., 1991). The PDF is the normal distribution, and stands on the following equation:

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y^i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}$$

(Eq. 2.16)

where,

$$D_i^2 = (X - X^i)^T (X - X^i)$$

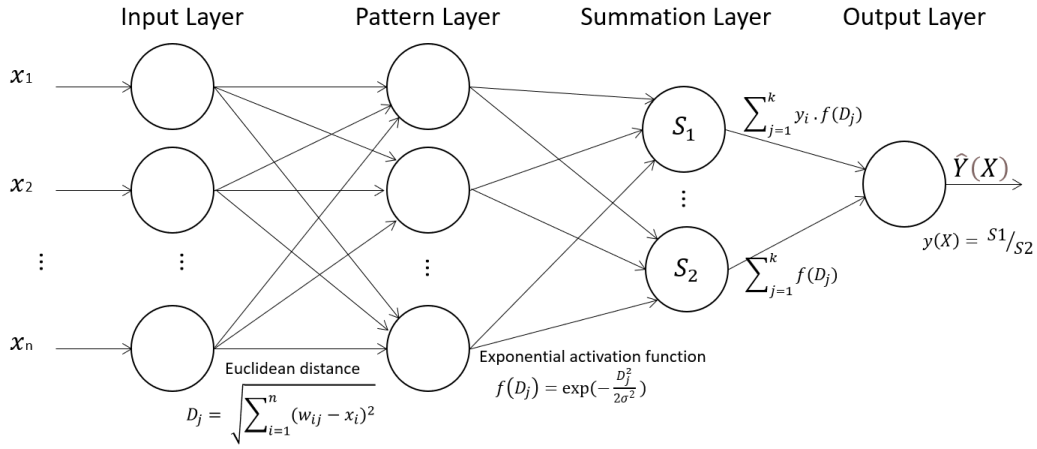
(Eq. 2.17)

and X is the input sample, X_i is the training sample, Y_i is the output of input sample X^i, D_i^2 is the Euclidean distance from X, $\exp\left(-\frac{D_i^2}{2\sigma^2}\right)$ is the activation function, $\hat{Y}$ is the estimated output value based on input X, and $\boldsymbol{A}^{\mathsf{T}}$ means the matrix transpose. The estimate $\hat{Y}(X)$ may be visualized as a weighted average of all the observed values, Y^i, where each observed value is weighted exponentially according to its Euclidean distance from X. The parameter σ is the smoothing parameter. When σ is large, the estimated density is forced to be smooth and in the limit it becomes a multivariate Gaussian. On the other hand, a smaller value for σ allows the estimated density to assume non-Gaussian shapes, but outliers may have too great an effect on the estimate. As σ becomes very large, $\hat{Y}(X)$ assumes the value of the sample mean of the observed Y^i, and as it approaches 0, $\hat{Y}(X)$ assumes the value of the Y^i associated with the observation closest to X. Although it is not possible to compute an optimum σ for a given number of observations *n*, it is easy to find on an empirical basis, for example, by minimizing the mean squared error between Y^j and the estimate $\hat{Y}(X^j)$ (SPECHT, D.F., 1991).

Neurons in the GRNN architecture are arranged in four layers, as didactically presented in (ANTANASIJEVIĆ et al., 2018) and shown in Figure 2.5.

The input layer is a distribution layer fully connected to the second layer. Its outputs are computed as the Euclidean distance (D_j) of the input from the stored patterns. On the pattern layer, the exponential activation function is used to calculate the output of each pattern neuron. Each pattern neuron is connected to the two neurons in the summation layer, which compute the sum of the weighted and unweighted outputs of the pattern neurons. Finally, the output neuron computes the predicted value by dividing the outputs of the summation layer, i.e., the weighted and unweighted sums.

Figure 2.5 – The GRNN architecture and calculations



Source: adapted from (ANTANASIJEVIĆ et al., 2018)

As summarized by Specht (SPECHT, D.F., 1991), the principal advantages of GRNN are fast learning and the convergence to the optimal regression surface as the number of samples become larger. GRNN is particularly advantageous with sparse data or small sample data, because the regression surface is instantly defined, even with just one sample. The one-sample estimate is that $\hat{Y}$ will be the same as the one observed value, regardless of the input vector X. A second sample will divide hyperspace into high and low halves with a smooth transition between them. Thus, the surface becomes gradually more complex with the addition of each new sample point. In an experimental setup, only 1% of the training was needed for the GRNN to achieve comparable accuracies for a MLP-BP model (SPECHT, D.F., 1991).

One further advantage is the reduced number of hyper parameters to be set, which is just one, a distinguishing difference from the MLP. GRNNs are also much faster to train than a MLP network, and they converge to a global minima, avoiding the convergence to local minima, which may occur with other techniques. Lastly, it is being currently used in many diverse domains, from estimation of traffic-related air pollutant

emissions (ANTANASIJEVIĆ *et al.*, 2018), to corrosion of steel embedded in soil (DING; RANGARAJU; POURSAEE, 2019).

The main disadvantage of the technique is the amount of computation required to estimate a new output vector, making it slower than MLP at estimating new cases. It also requires more memory space in order to store the model, and may be not practical for large datasets with complex models. Lastly, although Specht's lists this characteristic as an advantage, it may be seen like a disadvantage in regression problems: the estimate is bounded by the minimum and maximum of the observations, which may limit predicting new values.

2.3.3 Committee Machines

In committee machines, an ensemble of estimators—consisting typically of neural networks or decision trees—is generated by means of a learning process, and the prediction of the committee for a new input is generated in the form of a combination of the predictions of the individual committee members (TRESP, 2001).

This type of estimator's architecture is recognized as being useful in three different ways (TRESP, 2001). First, it can achieve a performance which is unobtainable by an individual committee member on its own, since the errors of the individual committee members, to some degree, are cancelled out when their predictions are combined. Typical representative approaches include *ensemble averaging*, *bagging*, and *boosting*. Second, it permits modularity, by enabling each estimator to focus on a particular region in the input space and the prediction of the committee can be obtained by a locally weighed combination of the predictions of the members. Its most representative approaches are the *mixture of experts* and its variants (HAYKIN, 1998). Third, it may reduce the computational complexity. Instead of training one estimator using all the training data, it may be more efficient at partitioning the data into several smaller data sets, training different estimators on the individual data sets and then combining the predictions of the individual estimators. Bayesian committee machines are examples of this case (TRESP, 2001).

This study uses only committee machines based on averaging, and therefore this section presents only this type of committee machines.

The basic idea behind averaging is to train a committee of estimators and combine the individual predictions, the goal of which is to achieve an improved

generalization performance when compared to the achievable performance with a single estimator (TRESP, 2001). In regression, the committee prediction for a test input *x* is achieved by a weighted sum of the predictions of the M committee members as following:

$$\hat{t}(x) = \sum_{i=1}^{M} g_i f_i(x)$$

(Eq. 2.18)

where $f_i(x)$ is the prediction of the i-th committee member at input x and g_i are weights associated to each member and often required to be positive and to achieve a sum of one. In the case of simple averages, weights are equal to each committee member, and are usually $1/M$.

The simple averaging approach is typically used with neural networks (TRESP, 2001). In the basic setup, the neural networks are all trained on the complete training set, and the decorrelation amongst the predictions is achieved by varying the initial conditions in training, such that different neural networks converge into different local minima of the cost function. In practice, it is a powerful solution for the disadvantage of achieving local minima on MLPs, by averaging results. With regards to efficiency, (PERRONE, 1993) and (KROGH; VEDELSBY, 1994) described that the error of the committee machine obtained by averaging is equal or less than the error of the committee members individually. As a recent practical result, an experiment comparing 77 popular regression models over 83 different datasets, (FERNÁNDEZ-DELGADO *et al.*, 2019) indicated that a simple average committee of neural networks composed of 5 equal MLPs trained using different random seeds was one of the well-performing models, despite being slow.

3 PREDICTING ELECTIONS WITH SOCIAL MEDIA DATA

“Humans are not very good at predicting the future, even just the next five minutes.”
(Gudmund Iversen)

The way politicians communicate with the electorate and run electoral campaigns was reshaped by the emergence and popularization of contemporary SM platforms. Due to the inherent capabilities of SM, such as the large amount of available data accessed in real-time, a new research subject has emerged, which focuses on the use of SM data to predict election outcomes. Although many studies have been conducted over the last decade, results have been controversial and often challenged. In this context, as initial steps of this thesis, we investigated and summarized how research on predicting elections based on SM data has evolved since its beginnings, so as to outline the state of both the art and the practice, and to identify research challenges and opportunities within this field.

In terms of method, we performed a systematic literature review analyzing the quantity and quality of publications, the electoral context of studies, the main approaches and characteristics of the successful studies, as well as their main strengths and challenges, and compared our results with previous reviews. We identified and analyzed 83 relevant studies, and the challenges were identified in many areas such as process, sampling, modeling, performance evaluation, and scientific rigor. The main findings include the low success rates of the most-used approach, namely volume and sentiment analysis on Twitter, and the better results of new approaches, such as regression methods trained with traditional polls. Lastly, a vision of future research on integrating advances in process definitions, modeling, and evaluation is also discussed, pointing out, amongst other items, the need to improve investigations into the application of state-of-the-art machine learning approaches.

The complete, detailed review, entitled “*A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions*” was accepted for publication in the IEEE Transactions on Computational Social Systems. At the time of this thesis's finishing, it is available as early access (BRITO,

K. dos S.; SILVA FILHO; ADEODATO, 2021). This Chapter presents a shorter version of the study, highlighting the main points.

The remainder of this Chapter is organized as follows: Section 3.1 presents the background related to the rise of election prediction with SM data, as well as an analysis of the main points of similar comparative studies. In Section 3.2, we present the main points of the review method and the procedure employed in this study, followed by Section 3.3, which provides an overall summary of the results. In Section 3.4, we summarize the main strengths, challenges, and future directions in the area. Lastly, Section 3.5 concludes and summarizes the outcomes.

3.1 RESEARCH BACKGROUND

3.1.1 The Rise of Election Prediction with SM Data

One of the first attempts that aimed at predicting election outcomes using data from SM may be attributed to Tilton (TILTON, 2008). In 2008, only two years after Facebook was launched for the general public, Tilton endeavored to predict election outcomes of a connected society, in this case a university, framed by the following research question: “Could Facebook be used to estimate the results of a student election?” Results showed that his model was able to predict into which place the candidates came in 21 out of 27 times in a given election. Probably because this was not related to a formal political scenario, Tilton’s study is seldom cited by studies in the area, although we consider it a very insightful preliminary study within this field.

Two studies may be considered seminal and have been cited by almost all the studies that followed. In 2010, Tumasjan et al. (TUMASJAN *et al.*, 2010) presented a study on the 2009 German federal election. They collected all the tweets that contained either the names of any of the six parties represented in the German parliament, or the most prominent politicians of these parties, and compared the volume of tweets with the election results. According to their results, they claimed that “*the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls.*” In the same year and with an improved approach via a sentiment detection of tweets, O’Connor (O’CONNOR *et al.*, 2010) observed that “*a relatively simple sentiment*

detector based on Twitter data replicates consumer confidence and presidential job approval polls."

Based on these two studies, the volume of tweets combined with automatic sentiment detection became the main approach for most further research around the world, such as in the Netherlands (SANG; BOS, 2012), Italy and France (CERON *et al.*, 2014), India (SINGHAL; AGRAWAL; MITTAL, 2015), Indonesia (PRASETYO; HAUFF, 2015), Colombia (CERON-GUZMAN; LEON-GUZMAN, 2016), Chile (RODRÍGUEZ *et al.*, 2018), and the U.S. (HEREDIA; PRUSA; KHOSHGOFTAAR, 2018). In general terms, researchers collected tweets referring to a candidate or party; performed a sentiment analysis to classify the post as positive, negative, or neutral; and attempted to correlate the volume of positive and negative posts with electoral results. In these studies, the main challenges were gathering data via an open search on Twitter and the sentiment analysis.

Despite being the most-commonly used approach, the analysis of the volume and sentiment of tweets engendered a number of criticisms just after their launch (GAYO-AVELLO, 2011; GAYO-AVELLO; METAXAS; MUSTAFARAJ, 2011; JUNGHERR *et al.*, 2017). In fact, by using these approaches, results may vary widely, as discussed by Jungherr (JUNGHERR; JÜRGENS; SCHOEN, 2012). After replicating Tumasjan's seminal study, Jungherr argued that "*the results are contingent on arbitrary choices of the authors,*" and indicated that simply including one more party or day of collection would greatly change the results.

Moreover, despite criticism, recent works have continued to use similar approaches to the volume and/or sentiment of tweets and have achieved a variety of results, both positive (BANSAL; SRIVASTAVA, 2018; SALARI *et al.*, 2018), negative (ANDY JANUAR WICAKSONO; SUYOTO; PRANOWO, 2016; SINGH; SAWHNEY; KAHLON, 2017) and even mixed (ANJARIA; GUDDETI, 2014; HEREDIA; PRUSA; KHOSHGOFTAAR, 2018). Additionally, novel approaches began to appear, such as models based on regression or time series methods (TSAKALIDIS *et al.*, 2015; ZHANG, X., 2018), and models using traditional polls for training or comparing results in order to calibrate the model (ISOTALO *et al.*, 2016).

3.1.2 Analysis of Previous Reviews

Due to the variety of approaches, with different achieved results even in replications of the same approach in the same context (JUNGHERR; JÜRGENS; SCHOEN, 2012), some researchers have tried to summarize the knowledge in this area.

In 2013, Kalampokis et al. (KALAMPOKIS; TAMBOURIS; TARABANIS, 2013) presented a systematic review aiming to understand the predictive power of SM, not only in the electoral context. By analyzing 52 studies, 11 regarding election predictions, they identified that the main approaches were based on volume, sentiment, and user profiling. In addition, the use of predictive analysis using linear regression was also identified, but not in the studies related to the political context. They also verified that 40% of the studies that had used sentiment-related variables challenged SM predictive power, i.e., it was not successful. This number increased to 65% in the case of lexicon-based approaches. Lastly, they emphasized the lack of predictive analytic evaluation and the controversial results of electoral predicting studies.

In the same year, Gayo-Avello (GAYO-AVELLO, 2013) presented a study that we consider to be the first review to specifically address predicting elections with SM, focusing on Twitter. By analyzing 10 previous studies from 2010 to 2013, he concluded that "*the presumed predictive power regarding electoral prediction has been somewhat exaggerated.*" Moreover, as in (KALAMPOKIS; TAMBOURIS; TARABANIS, 2013), he identified volume and sentiment analysis as the main approaches together with the need to use more up-to-date methods for sentiment analysis. He also expanded the list of challenges, such as the dependency on arbitrary decisions made by researchers regarding keywords, parties, and candidates, selection of the data collection period, and problems related to Twitter, such as demographic and self-selection bias, and bias related to spam, misleading propaganda and astroturfing. He ended the study by observing that regression models could be a future direction.

In 2015, studies by Prada (PRADA, 2015) and O'Leary (O'LEARY, 2015) presented the general lines of the main approaches for prediction using Twitter in many different domains, and briefly described a few studies related to election predictions (2 and 11 studies respectively). In 2018, Kwak (KWAK; CHO, 2018) presented the results from a survey including 69 papers, which supported the argument that SM may be used in order to understand political agenda, rather than election forecast. The most

recent studies (KOLI; AHMED; MANHAS, 2019)(BILAL *et al.*, 2019) have presented limited nonsystematic surveys, both analyzing 13 papers, adding some arguments to the original review from Gayo-Avello (GAYO-AVELLO, 2013). Koli (KOLI; AHMED; MANHAS, 2019) argued that prediction using Twitter was able to attain better results in developed countries, rather than developing countries, due to a higher literacy rate and more efficient internet access. In addition, Bilal (BILAL *et al.*, 2019) considered the challenges of sentiment analysis in languages other than English. However, despite these new arguments, recent studies have failed to identify novel approaches, as well as those that use SM other than Twitter and Facebook.

There is not yet a common consensus in the literature regarding well-established methods, processes, and tools for predicting election results based on SM data. Moreover, the SM landscape is undergoing continuous changes, as well as patterns of use. For example, Facebook has surpassed the number of active users of Twitter, and even new SM platforms have become more popular, such as Instagram (WE ARE SOCIAL; HOOTSUITE, 2020). Thus, a thorough review providing an understanding of the past and directions for future research is still needed and should be updated frequently until common bases can be defined.

3.2 METHODOLOGY

The method chosen for this research was a systematic literature review, which has proven to be a replicable and effective manner with which to identify, evaluate, interpret and compare studies that are relevant to a particular question or area (CENTRE FOR REVIEWS AND DISSEMINATION, 2009; DA SILVA *et al.*, 2011; KITCHENHAM; BRERETON, 2013; ZHANG, H.; BABAR; TELL, 2011). The method used in this research has followed the guidelines defined by (KITCHENHAM; BRERETON, 2013) and is fully described in the paper (BRITO, K. dos S.; SILVA FILHO; ADEODATO, 2021). This section presents the main points.

3.2.1 Research Questions

To define the research questions of this review, we returned to the main objective: “To provide a thorough review and investigation of the state of both the art and the

practice of predicting election outcomes based on SM data and to identify key research challenges and opportunities in this field".

Thus, the following research questions were derived:

- RQ1: In which electoral contexts is the research being performed?

This question aims at identifying the electoral contexts being studied, such as the year and country in which the election took place, and the type of election. This question is intended to ascertain whether the studies are best suited or giving attention to any particular electoral context.

- RQ2: What are the main approaches?

The objective of this question is to identify the main approaches used, their main characteristics, how they are modeled and applied to predict elections, and which metrics are used to assess their performance.

- RQ3: What are the main characteristics of successful studies?

The objective of this question is to identify the main characteristics of allegedly successful studies, in order to identify which specific contexts, which approaches, and which factors yield effective results.

- RQ4. What are the main strengths and challenges of predicting elections with social media?

After studying the context, approaches and characteristics of successful studies, the answer to this question aims to summarize the main perceived strengths, weaknesses, challenges, and opportunities in this new research area to guide future research.

3.2.2 Search Process

The rigor of the search process is one of the distinctive characteristics of systematic reviews (ZHANG, H.; BABAR; TELL, 2011). To implement an unbiased and strict search, two approaches were combined: (i) an automated search on indexing systems and (ii) a snowballing search on the references of studies found on the automated search.

The automated search was performed in four indexing systems: ACM Digital Library, IEEEXplore Digital Library, ISI Web of Science, and Scopus. The search was performed on the metadata of the papers: title, abstract, and keywords and aimed to

find studies focused on predicting elections based on SM data. After some initial refinements, the following search string was used in the automatic search:

(model OR method OR approach OR framework) AND (predict) AND (election*) AND (“social media” OR twitter OR facebook OR instagram).*

The snowballing search on the references was applied only at the end of the study selection so as to perform this search on already identified relevant studies only.

3.2.3 Quality Assessment

One initial difficulty regarding quality assessment is that there is no established manner with which to define study "quality." In this study, we have used the premise suggested by (CENTRE FOR REVIEWS AND DISSEMINATION, 2009), in which quality relates to the extent to which the study minimizes bias and maximizes internal and external validity. Thus, we have focused the quality assessment on the rigor of the study. Hence, we proposed the following quality assessment questions:

QA1: Are the aim(s)/objective(s) clearly identified?

QA2: Was the related work comprehensively reviewed?

QA3: Are the findings/results clearly reported?

QA4: Are bias and threats to validity clearly discussed?

QA5: Did the study compare the proposed solution and results with other works?

3.3 REVIEW RESULTS

The search procedure was last performed on July 31, 2020 and included all papers until 2019. The study selection resulted in a final set of 90 studies: 83 main primary studies and 7 surveys or literature reviews. Primary studies were analyzed and discussed to answer the research questions, whilst surveys were used in the discussion and comparison of this review’s results. This section summarizes the main results.

3.3.1 Quality Assessment

The objective of the quality assessment was not to exclude any study based on measured quality, but rather to understand the general quality of the published studies, and to detect possible strengths or weaknesses regarding methodology. Results demonstrated that studies generally satisfied questions QA1, QA2 and QA3. However, the main concerns reside in RQ4 and RQ5. Only 45% of the studies presented a discussion regarding the threats to validity, and only 11% of the studies conducted a clear comparison and discussion of their results with other previous research. These data lead to the conclusion that while many studies claim positive (or negative) results, it is difficult to support these results because no comparison with previous research has ever been conducted, and threats to validity are often not considered.

3.3.2 Electoral Contexts

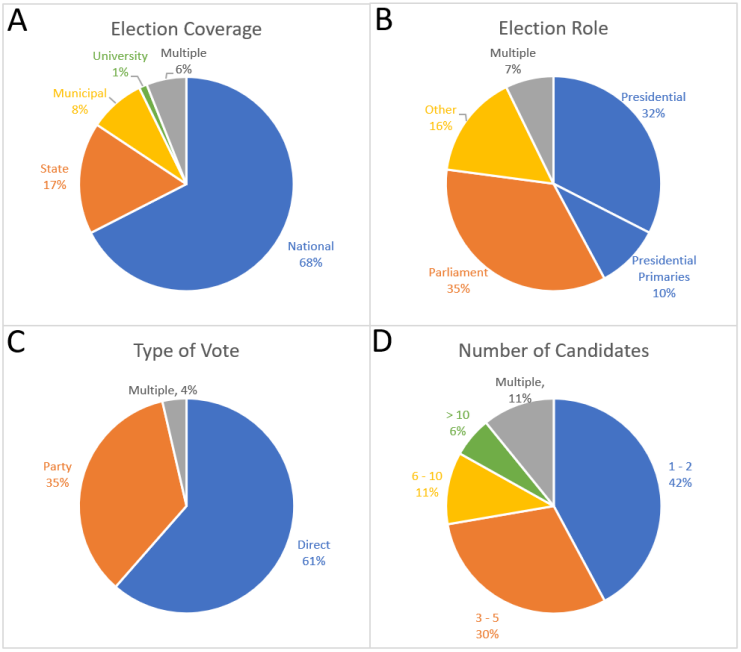
With regard to “RQ1: In which electoral contexts is the research being performed?”, we identified that most studies (72%) were performed in the context of a unique election, which may have impacted the applicability of their results, due to a lack of generalization. In addition, we identified that most studies were related to elections on a national level (68%), for the position of president (42%), and with a direct vote (61%) for a candidate. There were generally either only two candidates (42%), or a maximum of five candidates (72%). Figure 3.1 summarizes the general characteristics of the studied elections. These data are in line with the most-commonly studied scenario: U.S. presidential elections. It is important to highlight, that the prevalence of U.S. presidential elections may bias results, due to the specific characteristics of these elections, and the small number of studies on elections in Africa (only 2%) and Latin America (only 7%) illustrates that few assumptions may be made regarding elections in these regions.

3.3.3 Main Approaches

With regard to “RQ2: What are the main approaches?”, it was identified that approaches were grouped into five supermodel groups: (i) volume or sentiment; (ii) regression or time series; (iii) profile or posts interactions; (iv) topic analysis; and (v)

other unique approaches. Table 3.1 presents the number of studies classified according to each approach. The sum exceeds 100% because many studies used mixed approaches, and the table also reveals the number of studies that only used volume or sentiment approaches, not combined with anything else.

Figure 3.1 – Characteristics of studied elections: (a) coverage, (b) role, (c) type of vote, and (d) number of candidates.



Source: self-provided.

3.3.3.1 Volume or Sentiment

More than three-quarters of the studies were based on the detection of volume and/or sentiment of text on SM, which was the main approach used by the studies that included Twitter (61 out of 70 studies). Studies using this approach followed the proposal put forward by seminal studies such as those by Tumasjan (TUMASJAN *et al.*, 2010) and O'Connor (O'CONNOR *et al.*, 2010). Thus, the process they followed is: (i) Twitter data collection by pre-selected keywords; (ii) data cleaning, with the removal of tweets not addressing elections, together with duplicates or retweets; (iii) sentiment analysis; (iv) prediction based on sentiment counting analysis using a simple linear formula; and (v) performance evaluation. The linear formula in most cases was a direct correlation of the percentage of posts mentioning a candidate and his/her vote share.

The advantages of this are that it is a simple counting approach, it incurs low costs, is easily implemented, and generates fast results. Many authors, using this approach in a variety of scenarios have claimed success or promising results.

Table 3.1 – Main Approaches Identified

Prediction Model	**Studies**
Volume or Sentiment	64 (77%)
Volume or Sentiment solely	41 (49%)
Regression or Time Series	18 (22%)
Profile or Posts Interactions	14 (17%)
Topic Analysis	6 (07%)
Other approaches	6 (07%)

Source: self-provided.

We highlight two main challenges. First, the majority of studies focused on improving the sentiment analysis, and not on improving the actual prediction. However, lexicon-based analysis based on the presence of positive/negative words is the most common approach, and more sophisticated techniques based on the advances of artificial neural networks (ANN), including deep learning, are almost never used. The second main challenge is that the nature of this model leads to many biases, for example: (i) Twitter cannot be generalized as a good sample of all SM; (ii) collected data do not represent even a good sample of all tweets, due to platform constraints; (iii) it is too dependent on arbitrary decisions, such as search keywords and the selection of a period for data collection; and (iv) results are easily affected by volume manipulation from automated software, spammers, paid propaganda or even natural differences between online user behavior. Lastly, most studies using this approach performed what we term as one-shot predictions, *i.e.* just one prediction before elections, demonstrating a limited capability for use during campaigns.

3.3.3.2 Regression or Time Series

Regression and time series studies were grouped together because most time series models are, or share characteristics with, regression models. This was the second-most identified approach, present in 18 of the 83 studies (22%).

The main characteristic found in these studies is the use of traditional polls as additional input data, frequently used as ground truth for training predictive models. Moreover, in addition to Twitter data, studies also used data from Facebook, Google

Trends, Wikipedia and the candidates' home pages. Thus, new variables, such as Facebook likes and comments on official profile posts, number of page views, and metrics from Google Trends were added to Twitter volume and sentiment, in order to generate new sets of metrics. These metrics were then combined with offline poll results to train regression or time series models, capable of making predictions based on new instances of input data. The most commonly used models (44% of studies) were linear regression models, such as least square, ridge, and lasso. Moving average models, such as simple moving average (SMA), auto-regressive moving average (ARMA), and auto-regressive integrated moving average (ARIMA) were the second most used (33% of studies).

As advantages, these studies used machine learning and statistical methods for prediction. These methods are robust, well-grounded, and have been extensively tested in many other domains (SEBER; LEE, 2003; WEI, 2013). Also, by using traditional polls as ground truth for training, results are less affected by volume manipulation. Furthermore, the use of more SM platforms can reduce the inherent bias involved in using only Twitter as a data source, and focusing on official profiles reduces the bias regarding keyword selection. Lastly, this model seems to be more suitable for continuous predictions during the campaigns.

In terms of challenges, it is possible to identify a number of biases, such as the arbitrary selection of data sources, collected data, and the period of collection. For example, a different window size on the moving averaging techniques may totally change the results. Also, models chosen for regression and time series are limited for this context: linear regression may be not suitable, due to a possible nonlinear relationship between SM variables; and the ARIMA model is univariate, and therefore does not allow the combination of multiple variables. As a consequence, many of the studies analyzed each metric individually and chose the one with the best results, an experimental procedure that should be considered with caution.

3.3.3.3 Profile or Post Interactions

The number of interactions on posts or on the official profile of candidates or parties was also considered by a number of studies (17%). Three types of studies used this approach: (i) those that considered Facebook likes on posts made by official profiles as an approval rate or voter intention, similar to how volume/sentiment approaches

use mentions on Twitter; (ii) those that used a similar approach considering likes and dislikes on the Taiwanese PTT Bulletin Board System; and (iii) those that used likes or retweets as additional metrics in volume or sentiment models. These studies basically considered new metrics for prediction, not imposing novelty on the prediction model.

3.3.3.4 Topic or Event Detection

Topic or event detection and analysis are also supportive methods for other previously mentioned approaches. The six studies that used this approach did so as support or as a replacement for sentiment detection. By using Latent Dirichlet Allocation (LDA)(BLEI; NG; JORDAN, 2003), studies attempted to find the most important subjects being talked about in an election, the alignment of these topics with candidates, and then the volume and sentiment of public posts for or against the candidates. These studies may be considered as specializations of volume/sentiment approach, sharing their other characteristics.

3.3.3.5 Other Approaches

Unique studies include approaches based on prediction market, cluster detection, centrality score, statistical physics of complex networks, and analysis of groups of supporters, solely or in combination with previously described approaches.

3.3.4 Main Characteristics of Successful Studies

Less than two-thirds of studies (52 studies – 63%) were considered successful studies, 28% (23 studies) were considered unsuccessful, and 10% (8 studies) were categorized as having no clear results. Given that this type of research encourages the reporting of positive results, the low success rate of 63% is somewhat alarming, and puts into doubt the purported feasibility of predicting elections based on SM data. Moreover, if we consider the methodological limitations of most studies as the lack of replication in more than one context and the lack of statistical analysis of the results, it is plausible to consider that success may be obtained merely by chance, as directly argued in some of the studies. Table 3.2 presents the correlation of the characteristics and success rates of the studies.

From the data, we highlight a notable difference regarding the electoral year: in 35 studies regarding elections occurring between 2012 and 2015, there was a 77% success rate, in contrast to 47% of the 15 studies related to the years between 2008 and 2011, and 55% of the 33 studies on the years between 2016 and 2019. These data illustrate that, contrary to expectations, the success rate of studies does not increase over time. Studies on Asia (73% of success) and Latin America (71%) performed better than studies on Europe (63%) and Anglo-America (54%), despite the prevalence of studies being performed on the U.S. Moreover, studies on developing economies achieved greater success (74%) than on developed economies (57%), challenging the conclusions presented by Koli (KOLI; AHMED; MANHAS, 2019), who argued that predictions yield better results in developed countries.

Table 3.2 – Average Success Rate

Characteristic	Total of Studies	Success Rate
One Electoral Context	72	63%
Two Electoral Contexts	6	50%
Three Electoral Contexts	5	80%
Election Role: Presidential	27	63%
Election Role: Presidential Primaries	8	75%
Election Role: Parliament	29	69%
Election Role: Other	13	46%
Election Role: Multiple	6	50%
Type of Vote: Direct	51	63%
Type of Vote: Party	29	69%
Type of Vote: Multiple	3	0%
Number of candidates: 1 - 2	35	66%
Number of candidates: 3 - 5	25	64%
Number of candidates: 6 - 10	9	67%
Number of candidates: > 10	5	20%
Number of candidates: Multiple	9	67%
Election Year: 2008-2011	15	47%
Election Year: 2012-2015	35	77%
Election Year: 2016-2019	33	55%
Continent: Asia	26	73%
Continent: Anglo-America	24	54%
Continent: Europe	19	63%
Continent: Latin America	7	71%
Continent: Africa	2	50%
Continent: Oceania	1	100%
Continent: Multiple	4	25%
Economic Status: Developed	47	57%
Economic Status: Developing	34	74%
Economic Status: Multiple	2	0
Overall Success Rate	83	63%

Characteristic	Total of Studies	Success Rate
Model: Volume or Sentiment	64	55%
Model: Not Volume or Sentiment	19	89%
Model: Regression or Time Series	18	72%
Model: Not Regression or Time Series	65	60%
Model: Profile or Posts Interactions	14	64%
Model: Not Profile or Posts Interac.	69	62%
Model: Topic Analysis	6	83%
Model: Not Topic Analysis	77	61%
Model: Other	6	83%
Model: Not Other	77	61%
Social Network: Twitter	70	60%
Social Network: Not Twitter	13	77%
Social Network: Facebook	15	80%
Social Network: Not Facebook	68	59%
Social Network: Other Networks	13	85%
Social Network: Not Other Networks	70	59%
Other data input: Polls	17	76%
Other data input: Not Polls	66	59%
Days of collection: 1 - 30 days	33	61%
Days of collection: 31 - 90 days	28	64%
Days of collection: > 90 days	14	64%
Days of collection: Not specified	8	63%
Data volume: Less than 100,000	22	50%
Data volume: From 100,000 to 499,999	21	43%
Data volume: >= 500,000	26	73%
Data volume: Not specified	14	93%
Overall Success Rate	83	63%

Source: self-provided.

In terms of the approach used, volume or sentiment proved not to be a good approach: only 55% of the 64 studies that used this approach obtained success, in contrast to 89% of the 19 studies that did not use volume or sentiment. Reinforcing this finding, success was obtained by 72% of the 18 studies that used regression or time series approaches, 83% (of 6 studies) that used topic analysis, and 83% (of 6 studies) that used other specific approaches. These data enable us to argue that, despite being the most commonly used approach, the volume and sentiment approach is probably not the best way to predict elections based on SM, and more research needs to be conducted on other approaches, in special regression and time series, and topic analysis approaches.

In line with the previous conclusion, Twitter is not the best platform for data collection. While 60% of the 70 studies based on this SM platform were successful, 77% (of 13) not using Twitter achieved success. Moreover, better results were achieved with other platforms: 80% of the studies based on Facebook were successful (against 59% that did not use Facebook data), as well as 85% of studies using other data sources. Additionally, using polls to train the models also appears to be a promising practice: 76% of the 17 studies using polls as a data source were successful, compared to the 59% success rate amongst studies that did not use polls.

Regarding the number of data collection days, no significant differences were observed. Studies that collected data for less than 31 days, between 31 and 90 days, and for more than 90 days achieved success rates of 61%, 64% and 64% respectively. Finally, regarding the volume of microdata collected (e.g., number of tweets or Facebook posts), better results were obtained when a higher volume of data was collected, from the 47% success rate of the 43 studies that collected less than 500,000 data points to the 73% success rate of the 26 studies that collected more than 500,000 data points.

3.4 DISCUSSION OF MAIN STRENGTHS, CHALLENGES AND FUTURE DIRECTION

In this section, we aim to answer the final research question “RQ4. What are the main strengths and challenges of predicting elections with social media?” by summarizing and discussing the results presented in the previous sections. Thus, we present possible future directions for studies in this area.

3.4.1 Main Strengths of Predicting Elections with Social Media Data

As the main strengths of the analyzed studies, we highlight:

The use of new large amounts of available data: There is a large amount of data available on SM, including data on what people are saying about politicians or political parties, what politicians and parties are talking about, and the repercussion and reach of conversations. This data availability is unprecedented in human history and has changed the concept of media influencers. This change is from an era when the influence was mainly enjoyed by “big players” present on traditional media, mainly TV, to an era when ordinary people in small cities, with low or no budget, are able to exert a significant influence.

Real time data availability, collection, and analysis: In addition to having such a large amount of available data, these data can be collected and processed in real time. This capability opens up new opportunities in political campaigns, as these data may support quick adjustments to campaigns, policies, or speeches, e.g., in real time during a debate.

Low cost: Due to automated data collection and analysis, these approaches may be considered as low cost, relative to traditional offline polls, for which a coordinated operation with a large number of interviews is usually needed.

Advances of artificial intelligence: These approaches are strongly based on artificial intelligence. Fortunately, the last decade has witnessed substantial development in this area, including models and algorithms, as well as available hardware for model training and prediction execution, such as GPUs, distributed systems, grid computing and cloud computing. Thus, computations that, a few years ago, took weeks to execute, may currently be executed within a few minutes.

3.4.2 Main Challenges of Predicting Elections with Social Media Data

As main challenges of the studies, we highlight:

Lack of well-defined and replicable processes: Amongst the studies, it is difficult to find a definition of detailed, replicable processes, explaining and justifying the options and choices, in a manner that would yield replication in other scenarios by other researchers. Thus, as a consequence, despite certain efforts to replicate past

results with data from another or even the same election, the results achieved are usually quite different.

Lack of generalization: Combining the lack of replicable processes with the fact that most studies were applied to only one electoral context, there is little evidence as to whether the proposed approaches are applicable in other electoral contexts or whether they are generalizable. Thus, there is little evidence to determine whether positive results were obtained merely by chance, by overfitting the model to that specific election, or because it was a feasible predictive model. Moreover, due to the focus of the majority of studies on U.S. elections and the specific characteristics of this electoral context, it is hard to envision the results of application in other contexts. For example, authors of (ANJARIA; GUDDETI, 2014) applied the same approach in U.S. and India, and obtained success in the former but not in the latter.

Lack of prediction capabilities during the campaign: Almost all studies were performed after the election results were made public, and most studies were designed to perform only “one-shot” predictions, i.e., one prediction before elections, usually the day before. This design limits the applicability of approaches during campaign rallies, and there is little evidence that they are reliable for use during future campaigns. Indeed, most studies may be considered as posterior analyses of how the behavior on SN correlated to election results with descriptive goal, instead of how to perform predictions during electoral rallies.

Social media platforms do not represent a good population sample: Social media platforms cannot be considered a good sample of the population and should not be used as the only input data capable of generating generalizable results. For example, a recent report (WE ARE SOCIAL; HOOTSUITE, 2020) demonstrates that only 51% of the world population uses SN, the majority of whom are young and male. Additionally, another report published in 2019 indicates that Twitter users in the U.S. are younger, likelier to be identified as Democrats, more highly educated, and have higher incomes than overall U.S. adults (PEW RESEARCH CENTER, 2019). These data do not reflect world or U.S. demographics.

Twitter is most used but does not represent a good sample of SM platforms: The most used SM platform in most of the studies was Twitter (84%), and in many of them (75%), it was the only platform used as input. However, Twitter is not a good sample, even considering only SM users, due to its having very few active users (326 million), in relation to other platforms, such as Facebook (2.6 billion) and Instagram

(1,1 billion), according to a 2020 report (WE ARE SOCIAL; HOOTSUITE, 2020). Despite these data, it is hard to find a discussion on why the studies focused on Twitter. After analyzing the API of these SM platforms (FACEBOOK INC., 2020; TWITTER INC., 2020), we hypothesized that Twitter was chosen because it is easier for researchers to collect data on this platform. For example, starting in August 2018, the approval process for gathering data from Facebook and Instagram consisted of developing and deploying a fully functional system, creating and publishing a privacy policy and terms of use, recording a video showing all the functionalities related to Facebook and Instagram data collection, creating test accounts allowing Facebook employees to test the system and, in many cases, submitting formal documentation of an institution responsible for the system. By contrast, in August 2019, the Twitter approval process only involved completing a form with information about the system.

Collected data on Twitter do not represent a good sample of Twitter data: Twitter API may be used in two ways: streaming or query. By trying to gather large amounts of data, such as those used by Twitter-based approaches, developers may be limited in two ways: it returns a random sample of recent tweets published over the previous seven days; or the user is limited to 180 calls (which returns a maximum of 100 results by call) for a window of 15 minutes, which is usually insufficient to gather all tweets related to candidates.

Arbitrary data collection choices: In most studies, many of the choices involved in data collection were arbitrary, such as the data collection period, which usually varied from 3 days to 3 months before elections, and the keywords used for open search on volume/sentiment approaches. This created many problems, such as those presented by (GAYO-AVELLO, 2013), in which the performance was too unstable since it strongly depended on such parameterizations, and thus, unintentional data dredging could occur, due to post hoc analysis. This also reinforces the argument presented by Jungherr (JUNGHERR; JÜRGENS; SCHOEN, 2012) who, after replicating the seminal study of Tumasjan (TUMASJAN *et al.*, 2010), stated that "*the results are contingent on arbitrary choices of the authors,*" and indicated that simply including one more party or day of collection would greatly change the results.

High susceptibility to volume manipulation: Data volume manipulation on SM may be imposed in many ways, such as the use of automated software, known as BOTs (BESSI; FERRARA, 2016; FILER; FREDHEIM, 2017), spammers, paid

propaganda, astroturfing, or even natural differences between user behaviors (MUSTAFARAJ *et al.*, 2011).

Difficulties in crossing data from multiple networks: It is difficult, if not impossible, to implement the approach based on open search, used in Twitter-based studies, on other platforms, due to limitations of the API. For example, Facebook and Instagram do not allow an open search of general keywords. Similarly, even in studies considering high level metrics on regression or time series models, the models used are unsuitable for performing data analysis in an aggregated form. Thus, in these studies, each metric was analyzed and used for prediction in an independent manner, not allowing for the crossing of data from multiple networks and thereby limiting the effectiveness of results.

Lack of use of state-of-the-art machine learning: In studies based on volume or sentiment, there is more focus on improving sentiment analysis, rather than on the prediction model. Nevertheless, most studies relied on simple lexicon-based methods or on well-established methods, such as Naïve Bayes and Support Vector Machines (SVM). However, as mentioned above, these studies achieved little success. In addition, even in studies based on regression and time series, only simple and traditional methods were applied, with a prevalence of linear regression based on least squares, ridge, or lasso algorithms, and SMA/ARMA/ARIMA models for time series. Linear regressions are meant to describe linear relationships between variables, which cannot be assumed in this context. Also, ARIMA is a univariate model, and hence cannot exploit the leading indicators, nor combine multiple features as previously mentioned.

Recent advances have been made in machine learning models capable of dealing with these limitations, such as improvements in artificial neural networks, including recurrent neural networks or deep learning, although these have rarely been considered in current studies.

Technical modeling weaknesses: Despite being recognized by some authors that electoral prediction may be considered a time series forecasting problem with a very short series, authors have yet to bring to the fore data preprocessing techniques and AI time series modeling for the SM environment. This involves a precise characterization of the problem, the underlying mathematics of its dynamics and the approximations needed in the data analyses and preprocessing. This review was not able to reach papers dealing with these topics.

Additionally, it is well known that the results of using AI techniques and models may be very affected by the chosen parameters. However, very few studies take this into account, and the vast majority even made no mention of which parameters were used. From the studies that mention this aspect, 3 reported the use of default parameters of used tools, namely Weka (HALL, M. *et al.*, 2009) and Scikit-learn (PEDREGOSA *et al.*, 2011) and 4 made it evident that parameters were chosen by "trial-and-error". Only two papers presented discussions in this regard. This scenario presents a significant weakness in the area since failure may be related to parameter choice rather than the model itself.

Performance evaluation and scientific rigor: Additionally, the quality assessment and analysis of studies presented important drawbacks that could affect the reliability of the results: a lack of statistical analysis of the results; a lack of meaningful comparison of results with related works; and a lack of discussion regarding biases and threats to validity contained in the studies. The lack of such analyses and comparisons, when added to other challenges such as a lack of replicable processes and generalization, casts doubt onto the actual prediction capabilities of approaches based on SM.

The challenges presented above may be grouped into four categories, (a) process, (b) sampling, (c) modeling, and (d) performance evaluation and scientific rigor, as summarized and presented in Table 3.3.

Table 3.3 – Summary of Main Challenges on this research area

Category	Challenge
Process	Lack of well-defined replicable processes
	Lack of generalization
	Lack of prediction capabilities during the campaign
Sampling	Social networks do not represent a good population sample
	Twitter is the most used but is not a good sample of social
	Collected data on Twitter is not a good sample of Twitter data
	Arbitrary data collection choices
Modeling	High susceptibility to volume manipulation
	Difficulties in crossing data from multiple networks
	Lack of use of state-of-the-art machine learning
	Technical modeling weaknesses
Performance evaluation and Scientific rigor	Lack of statistical analysis of results
	Lack of meaningful comparison of results with related works
	Lack of discussion regarding bias and threats to validity

Source: self-provided.

3.4.3 Future Directions

The results indicate that research in this area is still in its infancy. Next, a discussion about its future is presented.

3.4.3.1 Future Directions in Process Definitions

As the most important direction for the future, we consider that studies should cease to be merely ad hoc initiatives and aim to become generalizable and repeatable processes. Thus, it may be possible to apply new approaches and models in many different electoral contexts, such as different countries and years, by proposing and testing improvements and comparing results. For this, processes defined for data mining and knowledge discovery may be used as a basis, such as CRISP-DM (SHEARER, 2000), SEMMA (ROGALEWICZ; SIKA, 2016) or DMLC (ALNOUKARI; EL SHEIKH, 2012). For example, CRISP-DM presents six phases: (i) business understanding, (ii) data understanding, (iii) data preparation, (iv) modeling, (v) evaluation, and (vi) deployment. Based on these phases, new approaches may benefit from detailing steps, inputs and outputs, models, and algorithms to be used in each phase, to become repeatable and generalizable.

Moreover, the process should also be adjusted to enable the use of approaches during campaign rallies, to increase their usefulness by opening new opportunities that support quick adjustment on campaigns, policies, or speeches in a continuous manner.

3.4.3.2 Future Directions in Model Definitions and Sampling

We agree with (KREISS; LAWRENCE; MCGREGOR, 2018), who stated that "*researchers should refrain from automatically generalizing the results of single-platform studies to social media as a whole,*" and results show that studies covering multiple SM platforms are necessary to provide a better frame for the prediction scenario. The research must also have certain characteristics. First, by using many SM platforms as input, studies should consider the different behavior of politicians and users on each platform. For example, while one politician may engage better on Twitter, others may perform better on Instagram. In an extreme case, one candidate may perform better on one SM platform at the beginning of a campaign, but later this

behavior may change. Second, data collection should be systematic and uniform in all the involved SM platforms, to allow a combination of different SM data as input data, and to avoid the common bias of arbitrary choices made by researchers. Third, new models should be resistant to volume manipulation, such as that threatened by spam, paid propaganda, bots, or even the different behavior of the electorate on the Internet.

As one possible direction, the use of state-of-the-art ML algorithms, for instance, based on ANNs may be a recommended approach, due to their characteristics: (i) ANNs can learn nonlinear mappings capturing complex relations amongst independent (input) and dependent (output) variables; (ii) ANNs do not need an explicit assumption for the model between the inputs and outputs; (iii) ANNs can generalize well; and (iv) ANNs do not require assumptions on the distribution of input data, unlike most statistical techniques. In particular, the multilayer perceptron (MLP) is likely to be useful in this research for having extra features such as being the most validated ANN, easy to use and a universal function approximator (HORNIK; STINCHCOMBE; WHITE, 1989). Also, to avoid volume manipulation, the training of ML algorithms on traditional polls is already presenting promising results.

Furthermore, the precise problem characterization, the underlying mathematics of the dynamics of the problem and the approximations needed in data analyses and preprocessing, which are already used in the fields of predictions and time series forecasting, should be addressed to leverage the quality of models. Finally, the proper addressing of precise parameter selection and tuning for models may also unlock a new level of reliability and robustness to the results.

3.4.3.3 Future Directions in Evaluation

To enable a better evaluation of the results of studies, future work could focus on establishing a common evaluation framework and common baselines. As discussed by (BEAUCHAMP, 2017), success must be measured statistically, not merely through description or mean average error, and must be relative to clear benchmarks, which may be previous election results, existing polls, or default assumptions, such as incumbency success. Thus, the application of statistical tests, such as Wilcoxon signed-rank tests, Wilcoxon–Mann–Whitney tests, Welch's t-test, or the paired t-test, to cite just a few, should be addressed.

Lastly, study reports should clearly discuss bias and threats to validity, together with the results.

3.5 CONCLUDING REMARKS

This chapter has presented a shortened version of our study, “A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions” (BRITO, K. dos S.; SILVA FILHO; ADEODATO, 2021). The study collected more than 500 articles, 90 of which focused on predicting elections based on SM data, and investigated and summarized how this new research field has evolved since 2008. Amongst these studies, 83 were primary studies aimed at predicting elections and seven were surveys or reviews of past studies.

The results show that the number of publications in this area is increasing and research has spread across 28 countries from every continent. Nevertheless, there cannot yet be found any prominent researchers, research groups, or clusters performing sustainable research in the area. Moreover, no common well-known forum for publication on this subject was identified, and results are spread across many forums.

With regard to electoral contexts, most studies were performed in the context of a unique election, which may have impacted the validity of the results. Also, most studies were related to presidential elections on a national level, with few candidates. Furthermore, the most studied scenario was the U.S. presidential scenario, which may impact generalization, due to its specificity.

Considering the main models used, we observed that most studies used the approach of volume/sentiment analysis only on Twitter, in a variety of data collection approaches. We also found that regression and time series analysis is increasing, with the use of multiple SM platforms, in addition to a number of supporting approaches, such as profile or post interactions and topic analysis.

By combining the characteristics and success of studies we observed that, despite being the most commonly used approach, volume/sentiment does not present high success rates, which is consistent with the conclusions of previous surveys. Thus, approaches such as regression or time series, or those based on profile/posts interactions may be a better choice to investigate and to introduce improvements; even completely new approaches, such as that based on statistical physics of complex

networks, may be tested. Lastly, studies based on Twitter achieved significantly lower success rates than studies based on other SM platforms, such as Facebook. Surprisingly, no studies based on Instagram were found.

Moreover, as main challenges, issues were identified in four areas. With regard to processes, we highlight the lack of well-defined, replicable and generalizable processes, as well as a lack of prediction capabilities during the campaign. In sampling, issues are mainly related to the fact that SM and Twitter data are not representative samples, and studies were performed with many arbitrary data collection choices. In modeling, we encountered difficulties in crossing data from multiple networks, the high susceptibility to volume manipulation, a lack of using state-of-the-art ML techniques and weaknesses in technical modeling. Considering the performance evaluation and scientific rigor of the studies, the lack of a statistical analysis of the results and of any meaningful comparison with related works are also main issues.

Finally, the study presented the authors' viewpoints on the future directions of predicting elections using SM data in three axes: process definitions, model definitions and sampling, and study evaluation. As the main directions, we would highlight the need for repeatable processes based on well-known methodologies, for example CRISP-DM or SEMMA; the use of state-of-the-art methods for regression based on machine learning that may combine data from multiple SM platforms, such as ANN; and the use of statistical tests for evaluating results, such as Wilcoxon signed-rank test or others.

One significant difference from previous studies concerns methodology. Only (KALAMPOKIS; TAMBOURIS; TARABANIS, 2013) followed a systematic approach, based on a Google Scholar search. Also, our study is the most extensive and complete to be found in the literature. We analyzed 83 primary studies and seven surveys focused on predicting elections, whilst other similar studies presented a significantly lower number of analyzed papers—at most 69, even including papers that did not strictly focus on the electoral context. This study also covered a broader set of data. For example, none of the previous studies had performed a quality assessment or analyzed and summarized the electoral contexts. Moreover, no analysis was found that focused on discovering correlations between study characteristics and successfulness, as performed in this study.

The results from this review contribute to the research field by providing the academic community, as well as practitioners, with a better understanding of the

research landscape and by identifying some of the gaps in the area that provide opportunities for future research. In addition to the future directions presented, this literature review may also, to a certain extent, be extended: a search extension may be performed to expand the search strategy and number of sources, thereby performing a broader study; a temporal update may be implemented without making modifications to the protocol, to expand the timeframe and compare results over time; and finally, both approaches may be combined.

4 PROBLEM DEFINITION AND METHODOLOGY

“Prediction is central in science and in evaluating alternative generalizations or models.”
(Arnold Zellner)

The main problem of this thesis may be clearly stated: Whether it is possible to predict elections using data from social media, and how to make these predictions within an acceptable error margin and during the campaign.

In the systematic literature review presented in Chapter 3, it was distinctly identified that this is an open problem that presents many challenges, and current research has achieved low success rates. Indeed, the same researchers applying the same approach in different contexts may achieve contradictory results (ANJARIA; GUDDETI, 2014; GOTO; GOTO, 2019), and different researchers applying the same approach in the same context, may also achieve opposite results (JUNGHERR; JÜRGENS; SCHOEN, 2012; TUMASJAN *et al.*, 2010).

For future directions, the review highlighted the need for generalizable and repeatable processes, as well as new models, especially including more than one SM platform and using more sophisticated nonlinear approaches of machine learning. Furthermore, some approaches based on using traditional polls as labeled data appear with promising results. Thus, the main goal of the present study becomes apparent:

Main Goal:
To define a process and create an ML model based on the SM performance of candidates, which is capable of making daily nowcasting and final predictions of election results with competitive results to traditional polls.

4.1 RESEARCH QUESTIONS AND HYPOTHESES

By considering the main goal of the thesis, the following research questions were defined.

- *RQ1: Is there a correlation between the SM performance of candidates and their electoral performance?*

This question aims to verify whether the SM and electoral performance are correlated, otherwise, it would not be possible to define a model for prediction. The main challenge for answering this question is how to model SM performance data in a meaningful and repeatable manner across elections.

- *RQ2: Is it possible to define a process and create an ML model capable of predicting election results based on the SM performance of candidates?*

This question aims to verify whether it is possible to predict the final results of an election within an acceptable error margin. The main challenge for answering this question is how to create a repeatable process and a model that is able to be trained with SM data and predict the election results within an acceptable error margin.

- *RQ3: Is it possible to define a process and create an ML model capable of performing daily nowcasting of election results based on the SM performance of candidates?*

This question aims to verify whether it is possible to nowcast the voting intentions of citizens in a continuous manner throughout the campaign within an acceptable error margin. For this question, the main challenge is how to adapt the model defined in response to RQ2 so as to be applied continuously throughout the campaign. Another challenge is how to define and measure an acceptable error margin, which is a challenge even in polling research.

This thesis is based on the postpositivist stance, by considering that, to answer the research questions, it is more productive to refute theories than to prove them, and we increase our confidence in a theory each time we fail to refute it (EASTERBROOK *et al.*, 2008). Thus, to answer the research questions, the following hypotheses were defined, followed by their respective null hypotheses, those that researchers aim to refute.

For RQ1, we based our hypothesis on Zajonc's exposition theory (MURPHY; ZAJONC, 1993; ZAJONC, 1980, 2001, 1968). Thus, we hypothesize that a different viewpoint regarding modeling SM data, which focuses on the user interactions on the profiles of candidates instead of the mentions to candidates on user profiles, would be successful. This leads to H_1 and H_1'.

H_1: It is possible to model the SM performance based on the interactions of users on the official profiles of candidates and find correlations between the SM and the electoral performances of candidates.

H_1': It is not possible to model the SM performance based on the interactions of users on the official profiles of candidates and find correlations between the SM and the electoral performances of candidates.

For RQ2, we hypothesize that it is possible to define a process and create an ML model for predicting the final results of elections. As input, the model will receive the SM performance data as features and traditional polls as labelled data, and produce the predictions as output, leading to H_2 and H_2'.

H_2: It is possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of predicting election results with competitive results to traditional polls.

H_2': It is not possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of predicting election results with competitive results to traditional polls.

For RQ3, we hypothesize that, as well as defining a process and creating a model to predict the final results of an election, it is also possible to nowcast public tendencies and perform accurate daily predictions, leading to H_3 and H_3'.

H_3: It is possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of making daily predictions of election results with competitive results to traditional polls.

H_3': It is not possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of making daily predictions of election results with competitive results to traditional polls.

4.2 RESEARCH METHODOLOGY

In order to reject the null hypotheses of H_1', H_2' and H_3' in favor of hypotheses H_1, H_2, and H_3, which answers the research questions, the following methodology was defined.

4.2.1 Rejecting H_1'

The methodology for rejecting H_1' consists of four steps:

S.1 – To define the concept of SM performance by defining and detailing the specific metrics related to performance;

S.2 – To collect the performance metrics on SM platforms;

S.3 – To collect the final vote share of candidates from the studied election;

S.4 – To run Pearson correlation tests for correlations between the SM performance metrics and the electoral performance.

It is important to highlight that this study has been performed after the elections. Therefore, the author must pay special attention to avoid unintentional bias related to arbitrary choices, as indicated in the analysis of previous studies. Thus, in order to avoid bias related to the arbitrary selection of the data collection period, correlation will be tested with data regarding many different collection periods before election day.

The Pearson correlation coefficient was chosen because it is a well known metric for covariance measure. The coefficient (*r*) ranges from -1 to 1. A value of 1 implies that a linear equation perfectly describes the relationship between the two variables, with all data points lying on a line for which one variable increases as the other also increases. A value of -1 implies a perfect negative correlation. A value of 0 implies that there is no linear correlation between the variables. There are guidelines for the interpretation of *r* (GOODWIN; LEECH, 2006), although it is dependent on context: a correlation of 0.8 is usually considered a high correlation, especially in the social sciences. However, the same correlation may be low if a physical law is verified using high-quality instruments. Even though it is commonly accepted that an $r >= 0.7$ indicates a high correlation, in this study we use the correlation coefficient rule of thumb, statistically justified in (KREHBIEL, 2004). In the text, considering *r* as the Pearson correlation coefficient and *n* the number of samples, three rules are presented:

- Rule of Thumb No. 1: If $|r_{xy}| \geq {}^{2}/_{\sqrt{n}}$, then a linear relationship exists.
- Rule of Thumb No. 2: If $|r_{xy}| \geq {}^{2}/_{\sqrt{n+1}}$, then a linear relationship exists.
- Rule of Thumb No. 3: If $|r_{xy}| \geq {}^{2}/_{\sqrt{n+2}}$, then a linear relationship exists.

As it is recognized by (KREHBIEL, 2004) that rule No. 1 is slightly conservative, to reject H_1' we will consider the threshold of rule No. 3 as indicating a correlation, the threshold of rule No. 2 as indicating a high correlation, and the threshold of rule No. 1 as indicating a very high correlation.

Thus, to reject H_1', at least one defined metric must present a correlation with the electoral results.

4.2.2 Rejecting H_2'

The methodology for rejecting H_2' follows the methodology of rejecting H_1', and the following steps will be performed:

S.5 – To define a framework composed of a process and ML modeling for election prediction;

S.6 – To collect traditional poll data regarding the same elections.

S.7 – To apply the framework on already collected data for prediction.

S.8 – To compare predictions with the electoral results and results of traditional polls.

As discussed in Chapter 3, most studies in this area do not compare results with strong baselines, and many merely claim that predictions are “close to” electoral results, that the prediction errors are low, or other such vague claims. However, to reject H_2', the error obtained with the defined approach needs to be measured with relevant metrics for the domain and compared with well-defined baselines.

The main metric used on the polls domain is the mean absolute error (MAE), described in Chapter 2, which is based on the error of each prediction, and will be the main metric used for evaluation. Thus, the errors obtained by predictions must be compared with well-defined baselines. The first is the historical MAE threshold, and the second is the errors obtained by the polls used for training the model.

In (JENNINGS; WLEZIEN, 2018), Jennings and Wlezien analyzed more than 30,000 national polls from 351 general elections in 45 countries between 1942 and

2017 and found a MAE of 1.83 for legislative elections (standard deviation of 1.56) and a MAE of 2.70 for presidential elections (standard deviation of 2.13), with almost no variation over the years. Thus, errors below the historical MAE or within 1 standard deviation (equivalent to 68% of historical values) will be considered in line with historical data.

In addition, the errors will be compared with those obtained by the last prediction of traditional polls individually, and with the final poll average.

Lastly, despite being rarely found in related studies, the predictions obtained with the proposed approach will be statistically tested in two manners. First, the predictions will be statistically tested with the electoral results to verify whether predictions are in accordance with the results. Next, the errors will be statistically tested with the errors obtained by polls, in order to verify whether the prediction errors are statistically equivalent, higher or lower, than the poll errors. For statistical tests, this is a scenario of matched samples: two continuous measures will be compared for the same sample, the vote share of each candidate or the measured error for each candidate. Within this scenario, a paired test is the most suitable, and two tests are natural candidates: the Wilcoxon signed rank test (WILCOXON, 1945) and the paired Student's t-test (STUDENT, 1908). The main difference between them is that the t-test is a parametric test, requiring that the sample means are normally distributed, which is not expected in our scenario with very few samples (number of candidates), and the nature of data. Thus, the Wilcoxon signed rank test will be used for statistical tests.

Although we could argue that each of these metrics may be sufficient to reject H_2', we will focus on the errors. Thus, to reject H_2', the statistical test of measurement errors must demonstrate that they are equivalent to or lower than the poll errors.

4.2.3 Rejecting H_3'

The methodology for rejecting H_3' follows the methodology for rejecting H_2'. Thus, after defining and testing a process and model for election prediction as one-shot prediction, the process and model will be adjusted for continuous prediction, according to the following steps:

S.9 – To adjust the defined process and ML modeling for daily nowcasting.

S.10 – To apply the framework on already collected data for prediction.

S.11 – To compare prediction results with the results of traditional polls.

To reject H_3', the error metrics will be similar to the error metrics used for rejecting H_2'. However, due to the characteristic of this context, it is difficult to obtain a conclusive, statistically demonstrated result. In order to reject H_2', we will compare our prediction with the final poll predictions, close to election day. However, to reject H_3', we do not have the "ground truth", but only an imprecise measure of it, the poll data.

This measure, and comparing the prediction accuracy during the campaign, are specific challenges, even for the field of poll research. There was no precise approach for this in any of the 83 studies papers, nor in studies focusing on election polling errors, such as (JENNINGS; WLEZIEN, 2018). Normally, only the polls close to election day are considered.

The main reason for this is that voters may simply change their minds on who to vote for or may only decide on their vote close to election day. Due to several factors, until the election there may be a high increase or decrease in a given vote share polled or predicted with 100% of accuracy on an arbitrary day before the election. Thus, the difference between the final vote share and the polled/predicted data may be high. Therefore, it would be never known if the polled/predicted data was accurate or not.

One possible way to compare the accuracy of different pollsters is **to measure the errors of each prediction related to the final election results**, in a similar manner to rejecting H_2'. The problem with this approach is that it rewards regularity and stability. The results will be consistent if the voting intentions for candidates remain stable during the whole campaign, but otherwise will present a strong bias. But this is not always the actual scenario. For example, in the 2018 Brazilian presidential campaign, the second most voted candidate (Haddad) was said to have 1.00% of voting intentions in January, but in October, received 20.75% of the vote. This difference is plausible since in January he was not even considered as being a possible candidate. However, if we measure the errors of each prediction related to the final election results, accurate predictions would be indicated as inaccurate.

Another approach may be to **compare predicted values with poll values over the entire period**. This approach may be useful to measure whether the predictions are similar to the polls, and would indicate that the predictions are good poll predictors. But this approach also presents certain drawbacks. There is a high variation of methodology and results amongst the pollsters, and some of them may be biased. The use of poll aggregation considers that pollster bias may compensate and cancel one another. Thus, averaging the errors of comparing predictions and polls using traditional

metrics, such as MAE, may be a suitable approach. However, to statistically test whether predictions and polls are similar would lead to wrong conclusions.

As a simple practical example of this challenge, let us consider that poll results during the campaign are biased in one direction and predicted results are biased in the other direction. If the last polls and predictions are the same and equally accurate, it is hard (or maybe impossible) to define statistically which one made better predictions during the campaign. Thus, we consider that it will not be possible to demonstrate statistically that our results would be similar or better than poll results.

It may be argued that it is possible to use approaches for polling aggregation, as presented in Chapter 2 (BLUMENTHAL, 2014; HILLYGUS, 2011; JACKMAN, 2005). One option would be to create moving averages: to average the results of polls performed on *N* days before the prediction, and compare the prediction with this value. However, this approach also presents drawbacks, because polls are not performed at equal intervals: as the election date approaches, the frequency of polls increases. The aim of this thesis however is not to investigate the best way to aggregate polls. Thus, a simpler approach will be used.

In this context, to reject H_3', the approach of comparing predicted results, polls, and the final vote share seems to be the most plausible option, although statistical tests on results may lead to wrong conclusions. Thus, we will perform two analyses. The first is a descriptive, qualitative analysis of the two most voted candidates regarding polls, predictions, and the final vote share. The second considers the measurement of prediction errors by considering polls as imprecise ground truth. Hence, competitive predictions will be considered as those that, compared with polls, present errors below the historical MAE of 2.7 percentual points, within 3.00 deviation, i.e., the error margin considered for most polls.

4.3 CONCLUDING REMARKS

In this Chapter, we have defined the main goal of this thesis, as well as the research questions, the null hypotheses, and the alternative hypotheses, the ones in favor of which the null hypotheses are rejected. We also defined the methodology for rejecting each null hypothesis and answering the research questions. We have discussed and detailed the metrics that will be used.

The following Chapter presents the basis and details of the three proposals created for rejecting the null hypotheses.

5 PROPOSALS

“So the problem is not so much to see what nobody has yet seen, as to think what nobody has yet thought concerning that which everybody sees.” (Arthur Schopenhauer, 1851)

This Chapter presents the basis and the details for the three proposals created to reject the null hypothesis, namely the following steps described in the methodology:

S.1 – To define the concept of SM performance by choosing and detailing the specific metrics related to performance;

S.5 – To define a framework composed of a process and ML modeling for election prediction;

S.9 – To adjust the defined process and ML modeling for continuous election prediction;

To define these proposals, we first analyze the specific domain challenges of the SM scenario and of the political scenario for ML modeling, presented in Section 5.1. Section 5.2 then presents the set of metrics defined for measuring SM performance. Section 5.3 presents the SoMEN, a **So**cial **M**edia framework for **E**lection **N**owcasting, composed of a process and model for predicting election results. Section 5.4 presents the SoMEN-DC, a **So**cial **M**edia framework for **E**lection **N**owcasting **D**uring the **C**ampaign, which is an execution strategy of SoMEN, enabling continuous prediction during campaigns. Lastly, Section 5.5 presents the concluding remarks.

5.1 DOMAIN CHALLENGES

In the scope of this thesis, there are two types of challenges that must be addressed: the social media challenges and the political scenario challenges for machine learning.

5.1.1 Social Media Challenges

The SM scenario presents certain strengths for the process of predicting election results. There is a large amount of available data, which may be collected and analyzed in realtime, at a low cost, when compared to traditional polling, as presented in Chapter 2. However, there are also a number of challenges, many of which were presented in Chapter 3, and summarized here.

The traditional approach is based on counting the volume of user comments related to a candidate, enhanced by the sentiment analysis of these posts, and directly correlating the percentage with the vote share, usually in a one (positive) post one vote correlation. This use is very similar to the straw polls of millions of people largely adopted before “the crisis of 1936” and with the worst results until then, based on a huge randomly selected sample and providing one complete report accumulated over a period. This may explain the low success rate of studies based on this approach. Additional challenges are that SM does not represent a good population sample and should not be used directly in this sense. Likewise, by considering the limits of data collection on official APIs, it is practically impossible to perform an open search and obtain all the user comments related to a candidate. Even if it were possible, after the Cambridge Analytica scandal (BERGHEL, 2018; ISAAK; HANNA, 2018), it became evident that performing user profiling, *i.e.*, inferring the voting intentions of individuals on SM based on their posts and comments, is undesirable due to questions of privacy. Also, volumetric approaches are highly susceptible to volume manipulation, which may be imposed in several ways, such as the use of BOTs, spammers, paid propaganda, astroturfing, or even natural differences between user behaviors.

Lastly, the rapidly changing SM landscape must also be considered. A given SM platform may be more prominent one year than in another, or even in the same year, and one platform may be relevant for one candidate but not for another.

5.1.2 Political Scenario Challenges

The political scenario, and electoral context, vary widely between countries. The most studied election context, the U.S. presidential elections, presents a specific scenario with specific characteristics, such as the indirect relation between vote share and election results, the existence of only two main political parties (Republicans and

Democrats), and the concept of safe states (those in which the victory of a particular party is already expected) and swing states (those that can reasonably be won by either the Democratic or Republican presidential candidate). Such characteristics make these elections very specific, and results on approaches designed for these elections may be hard to replicate in other scenarios. For example, in most Latin American countries, the presidential election races are run by many candidates from many parties, sometimes more than 10. The vote is direct, the concepts of safe or swing states do not exist, there are many parties, and even a small party may sometimes elect a president, as in the 2018 Brazilian presidential elections. However, the scarcity of studies related to Latin America suggests that very few claims can be generalized about this region.

Moreover, since presidential elections usually take place every four or five years, there are few available historical data. The scenario becomes even more challenging in the case of Latin America, where there is little party loyalty, changes of scenario between elections often occur due to scandals related to corruption, and the list of candidates may change even during the campaign. Thus, it is hard to obtain consistent historical data to train ML algorithms for election predictions.

Indeed, there is almost no available data for ML training: historical data are scarce, and there is only one actual labeled data, the final vote share, which is the aim of prediction. To address this main challenge, one direction is to use traditional polls as labeled data to train the models. However, the use of polls adds a new set of challenges. The polls themselves present a variety of errors, and their results are often challenged, as reported in Chapter 2. Also, there are many differences regarding pollsters, their methodology, presentation method, and results, which makes it challenging to group their results as a unique set of labeled data. Also, there are no evenly spaced time intervals between the polls, which usually decreases as election day approaches, imposing constraints on using traditional time-series approaches.

Lastly, predicting election results may be considered an activity of nowcasting, an estimation of the present or very near future, rather than a prediction or forecasting, since many electors only decide on their vote during the last few days before an election, or even on the very day of the election. As an example, two polls (INSTITUTO NACIONAL ELECTORAL, 2020) conducted one week before the 2018 Mexican Presidential elections presented the rates of 15% and 17% for "not answered," which includes the responses of "Don’t know" and "None of the candidates" for vote intention.

If just a small fraction of those people decide their vote for one candidate during the last week, it may drastically change the election results.

By considering these challenges, as well as the challenges and future directions of research in this area, observed in the systematic review in Chapter 3, we present the following proposals.

5.2 ENGAGEMENT METRICS FOR MEASURING SOCIAL MEDIA PERFORMANCE

As presented in Chapter 3, most studies measured performance on SM as the volume of posts (sometimes considering sentiment) from ordinary people talking about a candidate (usually on Twitter). Such studies are based on the seminal paper by Tumasjan (TUMASJAN *et al.*, 2010), who claimed that “the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls.” However, as previously discussed, this approach presents several drawbacks. This thesis presents an alternative.

In 1968 (ZAJONC, 1968) and beyond (MURPHY; ZAJONC, 1993; ZAJONC, 1980, 2001), Zajonc's studies on human psychology hypothesized that “mere repeated exposure of the individual to a stimulus object enhances his attitude toward it.” This effect, also termed the familiarity principle, has been demonstrated in many different contexts, such as paintings, sounds, geometric figures, and affective reactions. In agreement with this theory, Swap (SWAP, 1977) indicated that “overall, more frequently viewed others were preferred to those less frequently seen.” In other words, people tend to have better attitudes towards others whom they are used to seeing. Applying these theories in the electoral context, in 1986, Oppenheimer (OPPENHEIMER; STIMSON; WATERMAN, 1986) reported a correlation between the exposure of politicians and electoral performance, and Mondak (MONDAK, 1995) observed that “media exposure fuels political discussion.”

Unlike most common hypotheses, we have based our performance measurement on Zajonc’s mere-exposure theory by analyzing the number of people who pay attention to a candidate by interacting with their content and propagating their presence, within the context of SM. For this, we considered the official profiles of candidates on newsfeed platforms, such as Facebook, Twitter, and Instagram. Thus,

we initially considered two sets of engagement metrics: The first is the number and variation of the candidates' followers in each social network, and the second is the number of interactions on the candidates' posts.

The number of the candidates' followers in each social network is a direct measure of how many people subscribed to receive content directly from the candidates. It is expected that more subscribers lead to more people receiving content and paying attention to a candidate. However, this metric may fail to express how many people are paying attention to a candidate since not all the content of all accounts followed by someone is shown to them: SM algorithms prioritize showing user content which is more engaged with and more aligned to user preferences (LARS BACKSTROM; THE FACEBOOK, 2013).

The number of interactions on the candidates' posts consists essentially of the likes, comments, and shares on each post. These actions indicate that the user has seen and paid attention to the content and actively acted. One like may be considered a quick, easy endorsement of the content; a comment demands more cognitive effort and may be positive or negative; and a share replicates the content to the user's own network, thereby actively helping to propagate it. In the case of Facebook, a like has subtypes, such as "Like", "Love", "Haha", "Wow", "Sad" and "Angry". However, in practice there are no distinctions among these interactions, that may be considered as just one. This is because even negative reactions, such as "Sad" and "Angry", are usually negative regarding the content of the post, for example the reporting of a sad situation, and not a disagreement with whoever posted it.

Indeed, all these actions, not only sharing, help to propagate a candidate's presence online. As social network algorithms prioritize showing the content of users with more engagement (LARS BACKSTROM; THE FACEBOOK, 2013), this creates a snowball effect. As more people interact with a post, so it is shown to more people, leading to more people interacting with it. The end result of the exposure theory is that more engagement and more exposure may be correlated with a better attitude toward a candidate and more votes.

This approach was initially presented in (BRITO, K. *et al.*, 2019) and an extended version is presented in the study entitled "*Correlations of Social Media Performance and Electoral Results in Brazilian Presidential Elections*", accepted for publication in the Information Polity journal.

The approach was first tested with data regarding the 2018 Brazilian presidential elections, and results have shown strong correlations between the number of interactions on the candidates' posts and the electoral performance. On the other hand, since the number and variation of followers may fail to express the volume of people paying attention to a candidate, and did not present clear correlations on our previous studies, it was not considered in this thesis.

As a result, regardless of the social network under analysis, we have considered the metrics related to the number of likes, comments, and shares (or similar terms, such as retweets on Twitter as a synonym for shares) received by the official posts of candidates. We have considered the absolute numbers in a period and the relative numbers per post.

Let $p_c(d)$ denote the number of posts made by a candidate c on an arbitrary SM platform (such as Facebook) on day d, and $l(p_c(d))$ denote the total number of likes received by the posts made by the candidate c on day d. Similarly, let $s(p_c(d))$ denote the total number of shares received by the posts of the candidate, and $cm(p_c(d))$ the total number of comments received by the posts of the candidate c on day d. Thus, given a time window from $d = i$ to $d = f$, we consider the following metrics:

The number of posts P made by the candidate c on the SM platform within the time window:

$$P = \sum_{d=i}^{d=f} p_c(d)$$

(Eq. 5.1)

The number of likes L received by the posts of the candidate c on the SM platform within the time window:

$$L = \sum_{d=i}^{d=f} l(p_c(d))$$

(Eq. 5.2)

The number of shares S received by the posts of the candidate c on the SM platform within the time window:

$$S = \sum_{d=i}^{d=f} s(p_c(d))$$

(Eq. 5.3)

The number of comments CM received by the posts of the candidate c on the SM platform within the time window:

$$CM = \sum_{d=i}^{d=f} cm(p_c(d))$$

(Eq. 5.4)

The average number of likes per post (LP), shares per post (SP) and comments per post (CMP) received by the posts of the candidate c on the SM platform within the time window:

$$LP = {}^{L}/_{P}$$

(Eq. 5.5)

$$SP = {}^{S}/_{P}$$

(Eq. 5.6)

$$CMP = {}^{CM}/_{P}$$

(Eq. 5.7)

These metrics are generic and well suited for most newsfeed-based SM platforms. In the specific case of this study, we consider Facebook, Twitter, and Instagram as SM platforms, as justified in Chapter 6, and present the specific metrics in Table 5.1. However, if other relevant platforms are created or identified as being relevant in other election scenarios, their metrics may also be added by following the same rationale of interactions, even if they present slight differences. For example, in the case of YouTube, the number of visualizations, likes and comments on the videos posted by the candidates should be considered.

This new set of metrics deals with many challenges identified in studies in this scenario, because it is based on gathering data from many platforms in a well-defined, repeatable and generalizable way.

Furthermore, the presented set of metrics collects much less data, thousands of posts from less than a dozen of candidates, instead of millions of posts from the entire population talking about the candidates, which are collected by the mainstream approaches. This characteristic has become more important due to the increased limitations on data gathering on SM platforms. At the time of writing this thesis, the most popular platforms (Facebook, Twitter, and Instagram) allow around 3,500 of the most recent posts of individual accounts to be collected if the developer knows exactly from which accounts to gather data, as opposed to a sampling of all posts obtained by

open search on the platforms. Lastly, the unique arbitrary choice for data collection using this set of metrics is the time window of collection, and the challenge of choosing specific keywords for open search is dismissed.

Table 5.1 – Description of the metrics related to Facebook, Twitter and Instagram

Social Network	Metric	Description
Facebook	FBPosts	Sum of posts in the period
	FBLikes	Sum of likes in the period
	FBShares	Sum of shares in the period
	FBComments	Sum of comments in the period
	FBLikesPPost	Average of likes per post in the period
	FBSharesPPost	Average of shares per post in the period
	FBCommentsPPost	Average of comments per post in the period
Twitter	TTPosts	Sum of posts in the period
	TTLikes	Sum of likes in the period
	TTRetweets	Sum of retweets in the period
	TTLikesPPost	Average of likes per post in the period
	TTRetweetsPPost	Average of retweets per post in the period
Instagram	IGPosts	Sum of posts in the period
	IGLikes	Sum of likes in the period
	IGComments	Sum of comments in the period
	IGLikesPPost	Average of likes per post in the period
	IGCommentsPPost	Average of comments per post in the period

Source: self-provided.

5.3 SOMEN: A SOCIAL MEDIA FRAMEWORK FOR ELECTION NOWCASTING

Predicting elections with SM data presents many differences and additional challenges when compared to the usual ML problems and solutions. As presented in section 5.1, the problem definition is not crystal clear from input data space to target definition, and varies according to electoral context. Also, candidates and parties may vary from one election to another, and almost no historical assumptions may be made. One direction is to use traditional polls as labeled data to train the models, but this also imposes new challenges. Lastly, SM challenges such as the non-representativity of the population on the SM platforms, the susceptibility of volume manipulation, and the rapidly changing SM landscape must also be considered.

Hence, within this context, this proposal sets out to predict the final vote share of candidates in elections, which is a regression problem. The training and prediction will be based on the SM performance data as features. Polling data will be used as imprecise labeled data for supervised training during a time prior to elections. Thus,

the prediction will attempt to closely match the election vote share, which has only one sample.

Next, we present the defined process and the ML model.

5.3.1 The SoMEN Process

The SoMEN process is based on CRISP-DM (SHEARER, 2000), one of the best-known processes for data mining, in a way so that it is generic and may be adapted for presidential elections worldwide. The process contains five phases: (i) election understanding, (ii) data collection and understanding, (iii) data preparation, (iv) modeling and execution, and (v) evaluation. Due to the nature of the research, the deployment phase of CRISP-DM is not addressed in this thesis. All the other phases are presented below, and illustrated in Figure 5.1. This process was preliminarily presented in a recent paper (BRITO, K. dos S.; ADEODATO, 2020) and applied to the U.S. and Brazilian elections held in 2016 and 2018, respectively.

Figure 5.1 – SoMEN Process

Source: self-provided.

5.3.1.1 Election Understanding

The first phase is to understand the election scenario, which involves the election timeline, the candidates, SM platforms, pollsters, and polls.

Two main dates are visible on the election timeline: the official beginning of the campaign and election day. Therefore, data collection should at least consider the beginning of the campaign. Moreover, with this new SM scenario, politicians are usually in a permanent campaign, with neither geographic nor time constraints, and may begin mobilizing their voters long before the election period. Thus, it is important to define a reasonable date for initiating data collection even before the campaign. It is crucial that this decision is taken in advance because some SM platforms may limit the gathering of past data. We therefore suggest a date between 6 and 10 months before

elections, because this is normally when the list of official candidates starts to take form, and the initial polls start being published.

The list of probable candidates must be created at the beginning of data collection, and pruned until the election day. This is important because, in many cases, the candidates may change even during the official campaign, which in fact occurred during the last presidential elections in 2018 in Brazil and Mexico. The list of candidates should also be pruned because candidates with low voting intentions may bias the predictions in two ways. First, polls usually group minor candidates in the “other” category, which does not allow models to be trained with their data. Second, prediction errors may lead to misinterpreting the results, since an error of 0.5 percentage points regarding a candidate with 30% of votes is small, but the same error regarding a candidate with 0.1 percentage points is remarkably high. Thus, we suggest that candidates should only be considered if they consistently present at least 1% of the vote intentions in the polls.

It is also necessary to identify the main SM platforms used in the country where the election will be held, and to find the profiles of the candidates on each platform, preferentially verified accounts.

Lastly, a decision must be taken on which pollsters and polls to use in the training set. Some countries, such as the U.S., have a high number of publicly available polls and daily weighted averages created by news companies, such as those created by the Huffington Post (THE HUFFINGTON POST, 2016), New York Times (THE NEW YORK TIMES, 2016), and Real Clear Politics (REAL CLEAR POLITICS, 2016). Nevertheless, access to polls is a barrier in many countries. In Brazil, few polls are made publicly available, and in Mexico, data must be manually gathered from the national repository. Some strategies for pollster selection may be adopted, such as selecting the pollsters with the highest reputation and/or with the best results on previous elections. Results of poll aggregation sites may also be used. This decision must be taken carefully since it may directly affect the results since “garbage in, garbage out.”

To summarize this phase, the following steps must be performed:

1 – Discover the election timeline and decide on the window of data collection;

2 – Identify and prune the list of candidates;

3 – Identify the most relevant SM platforms used by candidates and their profiles on these platforms;

4 – Find and prune the list of pollsters to use as data input.

5.3.1.2 Data Collection and Understanding

The second phase involves collecting the two sets of data: SM data and poll data.

Collecting SM data is a challenging task for social and for machine learning researchers because it is difficult to find public datasets with this data. In order to gather data from SM platforms, complete information systems must be developed or acquired, and pass through the platform's verification process, as occurs with Facebook and Instagram and discussed in Chapter 2. After this, the candidates' posts, including the related metrics defined in the previous section, must be gathered on a daily basis.

Collecting poll data is also a challenge. The collection may be diverse and depends on manual collection from the websites of the electoral court, as in Mexico and Colombia, directly on the pollster websites, or even from document repositories such as Scribd (scribd.com), as in Argentina.

After data collection, the data understanding step should be performed. With regard to SM data, it is necessary to understand how candidates are using the SM platforms so as to identify which platforms are the most used and most relevant, and which should be considered in the prediction model. It is also important to verify data distribution and whether it is suitable for use in the chosen model. An understanding is required of poll data in order to prune the data and select which polls should be used.

To summarize this phase, the following steps must be performed:

1 – Obtain access to a software platform capable of gathering data from SM platforms;

2 – Collect data from the candidates' official profiles;

3 – Find and collect data regarding polls;

4 – Analyze data to identify the relevant platforms, and the suitability of data for the model.

5.3.1.3 Data Preparation

This phase aims to prepare the collected data to be used in the prediction model.

Because SM data used in this approach is public data collected from the official profiles of politicians, it should be collected using official application programming

interfaces (APIs). As a result, the data are complete and data cleaning is not necessary. However, the initial dataset should be enhanced and completely transformed to be used in the proposed ML approach.

Data are modeled so that the result r of a poll (or election results) at a specific date d is a function of the SM performance observed in the candidate's profiles in an aggregate window of w days prior d. In the specific case of this study, considering Facebook, Twitter, and Instagram as SM platforms, it is modeled as presented in Eq. 5.8.

$$r = f(F, T, I)_{d-w \ldots d-1}$$

(Eq. 5.8)

For the SM performance, we use the 17 features defined in Section 5.2 and described in Table 5.1. As an example, for a poll published on January 30 and considering a window of 28 days, input data are the individualized sum of all the posts, likes, comments, and shares/retweets from Facebook, Twitter, and Instagram from January 2 to January 29, and the ratio "per post" per platform for all of them.

The definition of the window size, i.e., the number of collection days before the target data, is central to data preparation and, most often, taken arbitrarily, as discussed in Chapter 3. However, in previous studies no correlations were found between the window size and success of predictions. Thus, two strategies may be adopted, both based on generating many datasets with different window sizes. The first, is to train/test the datasets with previously collected data (poll data) and use the dataset with the lowest errors on predicting polls. The second consists of creating a committee machine through an ensemble of estimators, each using one different dataset related to a different window, and averaging the result. The second approach may be promising, considering that the electorate's behavior is not uniform: some people may be used to accessing the SM platforms every day and interacting with the content of candidates on a daily basis, while others may access the platforms at different intervals. Thus, the use off an ensemble of estimators may be better for capturing different user behavior and avoiding the bias of selecting an arbitrary window size.

The small number of available polls in many elections leads to a small number of data samples for training and may lead to the well-recognized problems of high dimensionality (TRUNK, 1979) and violation of the VC-dimension (VAPNIK; LEVIN; CUN, 1994). Thus, it is desirable to use feature selection or dimensionality reduction

techniques, such as principal component analysis (PCA). In this context, the PCA is well suited because it eliminates the collinearity amongst features, which is likely in this scenario, while allowing dimensionality reduction.

To summarize this phase, the following steps must be performed:

1 – Define the observable window(s) to be used as input;

2 – Process data in order to generate the SM performance features;

3 – Generate the datasets by combining SM features and polls data;

4 – Perform feature selection/dimensionality reduction to generate the final datasets.

5.3.1.4 Modeling

This phase aims to choose an appropriate ML model and design an appropriate architecture for the predictions.

The candidate vote share prediction problem is characterized as a regression problem, because many continuous values are predicted, the vote shares. There are many regression methods, each with their own characteristics, strengths and weakness. Moreover, each method presents its own specific adjustable parameters, which may directly affect the results.

In order to choose an appropriate method, the small sample size of this domain, based on the number of available public polls, must be considered. The method should also be generalizable and should not depend on assumptions regarding the distribution of input data. Moreover, due to the low success of linear methods, as observed in the literature review on Chapter 3 and previously found in our own studies based on Brazilian data (BRITO, K. *et al.*, 2019)(BRITO, K. dos S.; ADEODATO, 2020), the model must be capable of drawing nonlinear mappings.

One challenge of the ML methods involves tuning its parameters. Hence, it is desirable to choose parameters by selecting similar problems in the literature or using techniques for automatically selecting parameters, such as grid or random parameter searches. The grid search for parameters is preferred since it may tune the parameter selection. First, the dataset is split into three sets: training, validation and test. It is then trained with the data on the training set and tested on the validation set, and the model with the lowest errors is selected. Finally, it is tested with the test set. The main disadvantage of this approach is the computational power needed to perform all the

calculations of all the parameter combinations. One way to reduce the computational power needed to find the best parameters is to use a randomized search, which randomly chooses some parameters from the grid search. It has been demonstrated that this approach may provide similar results to grid search, demanding less computational power (BERGSTRA; BENGIO, 2012).

Thus, this phase presents following steps:

1 – Choose an appropriate method and define an appropriate design;

2 – Choose the parameter selection strategy.

Section 5.3.2 discusses and suggests an appropriate ML model and architectural design for this problem.

5.3.1.5 Evaluation

Prediction evaluation is a challenge for the polling industry, as reported in the previous chapters. Evaluation must measure the difference between the predicted results and the candidate's final vote share. It must compare the errors obtained by predictions with the errors obtained by polls. Results must be measured with relevant metrics for the domain and compared with well-defined baselines.

A discussion regarding suitable evaluation metrics was presented in Chapter 4, Section 4.2.2. The same metrics and procedures used for rejecting H_2', are used for evaluating the results, with just a few additions.

The most commonly used metric for the polls domain is the mean absolute error (MAE), which is based on each prediction error, and is the main metric used for evaluation. However, other support metrics may also be used, such as the mean absolute percentage error (MAPE), which measures the percentage error, and the root mean squared error (RMSE). We consider the MAPE as being relevant because, for example, an error of 3 points in a vote share is much more relevant for a candidate with 2% of votes than for a candidate with 50% of votes, and this relevance is not captured by MAE. Moreover, the RMSE may also expose outliers. In addition, another metric used in the polls industry should also be used, the absolute error on margin (called AEOM in this thesis), which is the absolute value of the difference between the margin separating the two leading candidates in the prediction and in the actual vote share. This metric is relevant because it shows the error on the lead of the first candidate. Thus, while the main comparison metric is MAE, these other metrics should

be used in order to give more significance to the results, which may thus be observed from different viewpoints, and also in order to evaluate whether they are coherent in suggesting similar conclusions.

Thus, the following evaluation steps must be followed:

1 – Collect prediction and polls errors metrics related to the final vote share: MAE, MAPE, RMSE, and AEOM;

2 – Compare the MAE of prediction errors with the historical threshold of 2.7 (std. dev. = 2.13);

3 – Compare the MAE of predictions with the MAE of the last polls and poll average;

4 – Perform statistical tests (Wilcoxon signed rank test) to verify if predictions are in accordance with the electoral results;

5 – Perform statistical tests (Wilcoxon signed rank test) to verify if prediction errors are statistically equivalent, higher, or lower than the poll errors.

5.3.2 The SoMEN Model

This section proposes an ML model and an architectural design to be used in phase 4 of the SoMEN process, also involving decisions made in phases 2 and 3. It is important to note that many different choices may be made, and the objective is not to pursue the best model but rather one that is suitable and reasonable for this context.

Thus, for this modeling, the following decisions were taken:

- To reduce issues related to volume manipulation, such as the existence of astroturfing, spam, paid propaganda, and the use of BOTs, each candidate will be trained and predicted individually. In this way, the models are trained with the specific behavior of the supporters of each candidate;
- To reduce the dimensionality, as well as to allow the use of high-correlated features present on SM performance features, a PCA will be performed on the input dataset;
- Ten different datasets will be generated, with various windows sizes: 1..7, 14, 21, and 28 days;

- A committee machine composed of an ensemble of 10 predictors will be used: each will receive one dataset related to a specific window size, and the prediction will be the average of the members' predictions;
- The preferred ML methods for prediction are the MLP-BP and the GRNN.

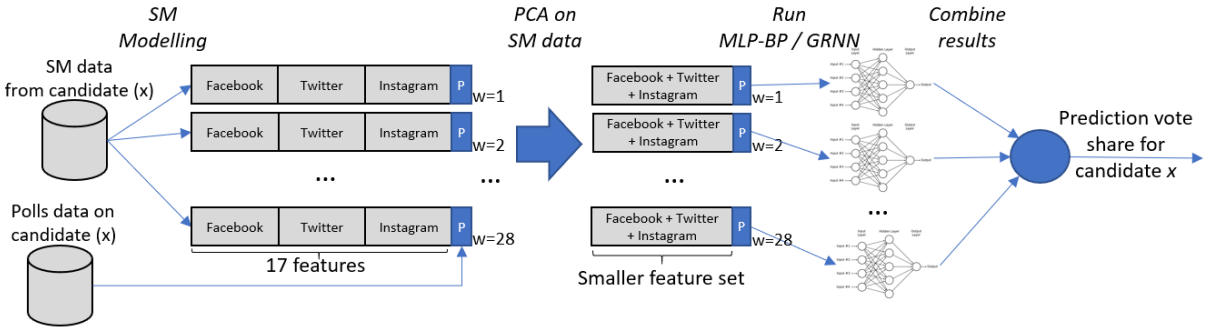
The MLP-BP was chosen because of its inherent characteristics, such as non-linearity, there is no need for assumptions on the distribution of input data, its good generalization capabilities, and its good performance even with the existence of noise data, plus it is proven to be a universal approximator (HORNIK; STINCHCOMBE; WHITE, 1989). The choice is also based on a recent extensive experimental survey of regression methods that compared 77 popular regression methods using 83 datasets (FERNÁNDEZ-DELGADO *et al.*, 2019). In the study, the MLP-BP based model designed with one hidden layer obtained remarkable results with small datasets. To avoid the well-known problem of local minima, the model consisted of another committee of 5 identical MLP-BP trained using different random seeds and was averaged to give a unique output.

The GRNN was chosen due to its main characteristics and advantages for this context. It is particularly advantageous with small sample data, because the regression surface is instantly defined even with just one sample. It also needs few examples for similar accuracy: in an experimental setup, only 1% of the training was needed for the GRNN to achieve comparable accuracies to a MLP-BP model (SPECHT, D.F., 1991). One of the other main advantages is that it only requires one hyper parameter to be set, a distinguishing difference from the MLP. It also converges to global minima, and is quicker to train, despite being lower for predictions.

There are some alternatives for the chosen ML models. An alternative for the selection of the best model and parameters may be the use of new research related to automated machine learning (AutoML) (HE; ZHAO; CHU, 2021), which promises automatically choose a good algorithm for a new dataset at hand, and also find their respective hyperparameters. Also, the use of recurrent neural networks, designed to learn sequential or time-varying patterns (MEDSKER; JAIN, 2001), would be suitable. Moreover, approaches of few-shot learning (WANG *et al.*, 2020) are being proposed to tackle the problem of small datasets.

Figure 5.2 illustrates the SoMEN instantiation, considering the data preparation and the modeling phases.

Figure 5.2 – SoMEN instantiation



Source: self-provided.

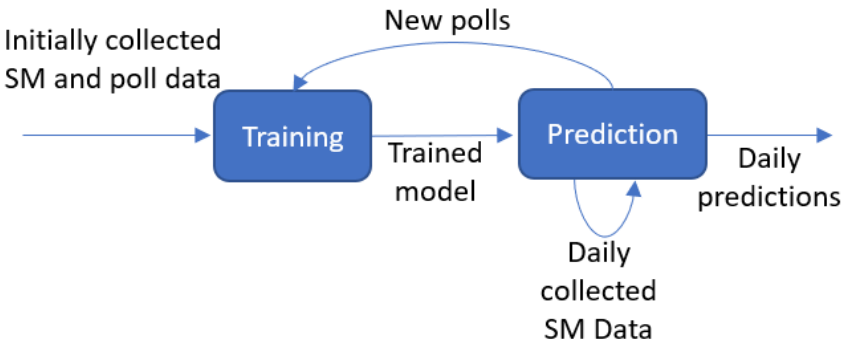
5.4 SOMEN-DC: A SOCIAL MEDIA FRAMEWORK FOR ELECTION NOWCASTING DURING THE CAMPAIGN

The SOMEN-DC is an execution strategy for SoMEN. The central point is that the defined process is linear: all SM and poll data are gathered, the model is trained, and the election results are predicted. Thus, the SoMEN-DC is a strategy to continuously repeat this process in a way that daily predictions may be made. Figure 5.3 illustrates the steps.

The execution consists of:

1. Gathering a minimum number of polls and the SM performance data;
2. Training the model with available data;
3. Collecting SM data daily and make out of sample, daily predictions;
4. Retraining the model with a new set of labeled data, when new poll data is released.

Figure 5.3 – The SoMEN-DC execution



Source: self-provided.

The first decision concerns the minimal number of polls needed to begin making predictions. Some regression models, such as the general regression neural network (GRNN) (SPECHT, D.F., 1991), are well suited for starting predictions just after one or two samples, although the majority of models need more data. It is also expected that when more poll data arrives and, consequently, more labeled data is available, prediction accuracy increases. This is exactly what occurs in the electoral context: as the election date approaches, more accurate predictions are made.

The second decision concerns retraining the model. The previously chosen models, MLP-BP and GRNN, are well suitable for this context. However, other approaches may be better suited, such as those based on online learning (LOSING; HAMMER; WERSING, 2018), because they may perform incremental learning, avoiding multiple training using all the datasets, thereby optimizing the process. In this study we will also use MLP-BP and GRNN for continuous predictions, although the use of these new approaches may be promising as future research.

Evaluating daily predictions is even more challenging than evaluating the final predictions, as discussed in Chapter 4. Thus, the approach for evaluating the SoMEN-DC performance will use the same metrics for rejecting H_3', based on comparing predicted results, polls and the final vote share.

Thus, the following evaluation steps must be followed:

1 – Descriptively and qualitatively analyze the two most voted candidates with regard to polls, predictions, and the final vote share;

2 – Measure prediction errors by considering polls as imprecise ground truth;

3 – Compare the MAE of prediction errors with an historical threshold of 2.7 within the traditional poll error margin of 3.0 points.

As discussed in Chapter 4, this procedure may suggest that prediction errors are competitive, or not, with polls, but statistical tests on these results may lead to wrong conclusions.

5.5 CONCLUDING REMARKS

This chapter began by highlighting the domain challenges of using ML for predicting election results based on SM data. We then presented the main proposals of this thesis: (i) a new set of metrics to measure SM performance; (ii) the SoMEN process to guide the steps from election understanding to prediction evaluation, and

the SoMEN ML model for prediction; and (iii) the SoMEN-DC, an execution strategy for SoMEN, enabling continuous prediction on a daily basis. All phases of the process were defined and detailed, and the main choices were discussed.

As presented in Chapter 2, in the study after the poll crisis of 1936, Crossley (CROSSLEY, 1937) indicated the characteristics of studies with the best and worst results, and highlighted what was considered at the time to be an ideal poll. It is worth noting that the most commonly used approach of volume/sentiment on Twitter share the main characteristics of studies with the worst results: they are based on a huge randomly selected sample and provides one complete report accumulated over a period. On the other hand, the proposal of this thesis may be considered compliant with what Crossley considered at the time to be the ideal poll: (i) it is flexible and not based on dated mailing lists, because it may be adapted for use with data from the most commonly used SM platforms at the time of election; (ii) a fairly small sample works properly, since we collect data from the official profiles of candidates, rather than the entire SM platforms population; (iii) the distribution of the sample is considered, since the proposals collect data from multiple platforms, use offline polls for training, and train the models for each candidate individually; and (iv) it is not cumulative, but repeated during the campaign and considering different time windows.

The next Chapter presents the instantiation and execution of the defined process and model. Experiments were performed with data from presidential elections, which took place in 2018 and 2019, in 4 major Latin American countries: Argentina, Brazil, Colombia, and Mexico.

6 EXPERIMENTS – LATIN AMERICAN PRESIDENTIAL ELECTIONS

"All models are wrong, but some are useful"
(George Box, 1987)

In this Chapter, we perform experiments to reject null hypotheses H_1', H_2' and H_3' in favor of the alternative hypotheses H_1, H_2, and H_3, which enables the research questions of this thesis, RQ1, RQ2, and RQ3 to be answered, all of which were detailed in Chapter 4, together with the methodology. For this, we applied the data modeling, the process, and the ML models described in Chapter 5.

As presented in Chapter 3, one of the main weaknesses of the related works is the application of defined processes and models in unique elections and, since studies are performed after elections, unintentional bias regarding the unique scenario may be introduced. Thus, the processes are frequently not replicable nor generalizable. In addition, since there is only a small number of studies on elections in Latin America, few assumptions may be made regarding elections in this region. For these reasons, in order to verify the generalization and replicability of our approach, and to study a relatively unstudied region, we chose to perform experiments on predicting the most recent presidential elections in Latin America.

Countries throughout Latin America all present a similar context: they are situated in the same region, they have similar historical origins, languages, characteristics and electoral procedures, such as the direct vote, the two-round system, many candidates running in the first round, and few available polls.

Presidential elections held in Colombia, Mexico, and Brazil in 2018 and Argentina in 2019 were chosen for the experiment. These are the four most populous countries in the region, home to 70% of the population (UNITED NATIONS, 2019), and were responsible for 81% of the GDP in 2019[8], which increases the validity of our experiments. We only considered the first round of elections, because the election context is different to the second round: it involves many candidates with varying strategies, because there is unequal exposure on TV and in the traditional media.

[8] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=ZJ&most_recent_value_desc=true Viewed on January 01, 2021.

Other presidential elections in the region were initially considered but were eventually discarded. The 2019 Bolivian elections were discarded due to disputes over transparency and legitimacy, which led to new elections. The 2018 Paraguayan and 2019 Uruguayan elections were also not included due to the researcher's difficulty in gathering poll data, which would be a manual process.

It should be noted that, in the systematic review presented in Chapter 3, none of the included studies considered more than three countries. This is also one of the rare studies that published results before the official release of the election results. A preliminary result predicting the 2019 Argentinian election, based on our preliminary methodology described in (BRITO, K. dos S.; ADEODATO, 2020), was published on Facebook[9] on election day and attained better results than the considered polls at that time. Also, results of applying the up-dated methodology described in this thesis to predict the vote share of the 2020 U.S. elections were published in LinkedIn on the day of elections[10], obtaining an MAE error 0.1 point lower than the RCP[11] poll average.

In the following sections, we present the definition and results of the experiments following the SoMEN and SoMEN-DC processes.

6.1 THE SOMEN EXECUTION

6.1.1 Election Understanding

The selected elections took place in 2018 and 2019, and the official campaigns lasted between two to four months. In all countries except Brazil, polls are not allowed to be released in the week before the election date.

To maximize the amount of collected data, mainly by considering that there are few available polls in the selected countries, we decided to begin collecting data 10 months before the elections (300 days), when the candidatures start to take form. The

[9] Available at https://www.facebook.com/notes/404228243926826/. Viewed on January 13, 2021.

[10] Available at https://www.linkedin.com/pulse/an%25C3%25A1lise-prevendo-o-resultado-das-elei%25C3%25A7%25C3%25B5es-nos-estados-kellyton-brito. Viewed on February 08, 2021.

[11] https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html. Viewed on January 13, 2021.

election timelines, including the election date, the campaign launch and the final day that pollsters are allowed to publish polls are presented in Table 6.1.

Table 6.1 – Election Timelines

Country	Election Date	Campaign launch	Final day to release polls	Start of data collection
Argentina	27/10/2019	07/07/2019	18/10/2019	31/12/2018
Brazil	07/10/2018	16/08/2018	-	11/12/2017
Colombia	27/05/2018	27/01/2018	20/05/2018	31/07/2017
Mexico	01/07/2018	30/03/2018	26/06/2018	04/09/2017

Source: self-provided.

The list of candidates is presented in Table 6.2. To avoid bias generated by candidates with small vote intentions, the study considered only candidates with more than 1% of vote intentions.

Table 6.2 – List of Candidates

Argentina	**Brazil**	**Colombia**	**Mexico**
Alberto Fernandez	Jair Bolsonaro	Iván Duque Márquez	Andrés Manuel López Obrador
Mauricio Macri	Fernando Haddad	Gustavo Petro	Ricardo Anaya Cortés
Roberto Lavagna	Ciro Gomes	Sergio Fajardo	José Antonio Meade Kuribreña
Nicolas del Caño	Geraldo Alckmin	Germán Vargas Lleras	Jaime Rodriguez Calderon
Juan José Gómez Centurión	João Amoêdo	Humberto De la Calle	
José Luis Espert	~~*Cabo Daciolo*~~	~~Jorge Antonio Trujillo~~	
	~~*Henrique Meirelles*~~	~~Viviane Morales~~	
	~~*Marina Silva*~~	~~Promotores Voto En Blanco~~	
	~~*Alvaro Dias*~~		
	~~*Guilherme Boulos*~~		
	~~*Vera Lúcia*~~		
	~~*Eymael*~~		
	~~*João Goulart Filho*~~		
6 Candidates, all considered	*13 candidates, 5 considered*	*8 Candidates, 5 considered*	*4 candidates, all considered*

Source: self-provided.

In order to identify the most relevant SM platforms, we considered the most commonly used SM platforms in all 4 countries, according to recent reports (KEMP; WE ARE SOCIAL; HOOTSUITE, 2020b, 2020d, 2020c, 2020a). They are listed in Table 6.3.

Because of the characteristics of our proposed model, the present study only considered news-feed based platforms. Consequently, Youtube, Facebook, Instagram, and Twitter were considered.

During the task of discovering the candidates' profiles on the most relevant platforms, it was identified that most candidates did not have an official account (nor

an account that could be unmistakably identified as the candidate's account) on Youtube. Also, even when they had an official account, the use of Youtube was irregular. Thus, Youtube was not taken into consideration, and Facebook, Instagram, and Twitter were.

Table 6.3 – Most used SM platforms in the analyzed countries

	Argentina	Brazil	Colombia	Mexico
1st	Youtube	Youtube	Youtube	Facebook
2nd	Whatsapp	Facebook	Facebook	Youtube
3rd	Facebook	Whatsapp	Whatsapp	Whatsapp
4th	Instagram	Instagram	Instagram	Facebook Messenger
5th	Facebook Messenger	Facebook Messenger	Facebook Messenger	Weixin / Wechat
6th	Twitter	Twitter	Twitter	Instagram

Source: self-provided.

The candidate profiles on these three networks are presented in Table 6.4. Those profiles marked with a star represent unverified profiles. At the time of the data collection for this thesis experiment, the profiles of the Mexican candidates Obrador and Cortés were personal profiles, rather than commercial profiles. This setup does not allow for data to be collected through the Instagram API. Thus, data on these candidates from Instagram were not used in this study.

Identifying pollsters was a manual task, and was performed on three sources, in this order: (1) On the official electoral court website, (2) On the Wikipedia entry describing the election in the official language of the country (Portuguese or Spanish) and in English, and (3) through an open search on Google using the words "pesquisa eleições 2018 Brasil" for the Brazilian elections, and "encuestas electorales <country> <year>" for the other countries, replacing <country> for the country name and <year> for the year of election. The findings and pruning of the list of pollsters and polls is detailed in the following section.

6.1.2 Data Collection and Understanding

6.1.2.1 Data from SM Platforms

Over the last eight years, a set of tools has been developed by the author for collecting data on politicians from open data repositories and SM platforms (BRITO, K. dos S. *et al.*, 2014a, 2014c, 2014b, 2015b, 2015a), using the concepts of social

machines (BRITO, K. dos S. *et al.*, 2020; BRITO, K. S. *et al.*, 2012; BURÉGIO, V. *et al.*, 2015; BUREGIO; MEIRA; ROSA, 2013; MEIRA *et al.*, 2011).

Table 6.4 – Candidate profiles on Facebook, Twitter, and Instagram

Candidate - Argentina	**Facebook**	**Twitter**	**Instagram**
Alberto Fernandez	alferdezok	alferdez	alferdezok
Mauricio Macri	mauriciomacri	mauriciomacri	mauriciomacri
Roberto Lavagna	LavagnaARG	Rlavagna	rlavagna
Nicolas del Caño	NicolasDelCano.PTS	NicolasdelCano	nico_del_cano
Juan José Gómez Centurión	juanjomalvinas	juanjomalvinas*	*not found*
José Luis Espert	JLEspert*	jlespert	joseluisespert
Candidate - Brazil	**Facebook**	**Twitter**	**Instagram**
Jair Bolsonaro	jairmessias.bolsonaro	jairbolsonaro	jairmessiasbolsonaro
Fernando Haddad	fernandohaddad	Haddad_Fernando	fernandohaddadoficial
Ciro Gomes	cirogomesoficial	cirogomes	cirogomes
Geraldo Alckmin	geraldoalckmin	geraldoalckmin	geraldoalckmin_
João Amoêdo	JoaoAmoedoNOVO	joaoamoedonovo	joaoamoedonovo
Candidate - Colombia	**FB**	**TT**	**IG**
Iván Duque Márquez	ivanduquemarquez	IvanDuque	ivanduquemarquez
Gustavo Petro	gustavopetrourrego	petrogustavo	gustavopetrourrego
Sergio Fajardo	SergioFajardoV	sergio_fajardo	sergiofajardovalderrama
Germán Vargas Lleras	GermanVargasLleras	German_Vargas	germanvargaslleras*
Humberto De la Calle	DeLaCalleHum	DeLaCalleHum	delacallehum*
Candidate - Mexico	**Facebook**	**Twitter**	**Instagram**
Andrés Manuel López Obrador	lopezobrador.org.mx	lopezobrador_	~~lopezobrador**~~
Ricardo Anaya Cortés	RicardoAnayaC	RicardoAnayaC	~~ricardoanayacortes**~~
José Antonio Meade Kuribreña	JoseAMeadeK	JoseAMeadeK	joseameadek
Jaime Rodriguez Calderon	JaimeRodriguezElBronco	JaimeRdzNL	jaimerodriguezcalderon

Source: self-provided.

* Unverified profiles. ** Not commercial profiles at the period of data collection. Thus, data collection on these profiles was not possible.

The social machine was originally defined by Tim Berners-Lee (BERNERS-LEE; FISCHETTI, 1999) as "processes in which the people do the creative work and the machine does the administration." Later, Meira (MEIRA *et al.*, 2011) published a seminal paper putting forward a new interpretation in a particular setting, as “A network of programmable machines that are connected to each other and that also connect people and institutions in a web of computing, communication and control.” By evolving Meira’s definition, Buregio (BURÉGIO, V. A. de A., 2014; BUREGIO; MEIRA; ROSA, 2013) characterized social machines as a result of the convergence of three different visions: (i) social software; (ii) people as computational units; and (iii) software as sociable entities, and thereby defined a new basis for the design and implementation of social systems.

The concepts of social machines are well fitted to the required software for collecting SM data, and were used to design and implement an information system towards this end. As the software engineering of this system is not the focus of this thesis, the system will not be detailed, but it is worth noting that it passed through all the necessary verification processes to gather data from the official APIs of Facebook, Instagram, and Twitter. This system also collects data on a daily basis on politicians from the end of 2017.

A total of 15,432 posts was collected from the Argentinian candidates: 3,224 from Facebook, 10,444 from Twitter, and 1,764 from Instagram. Although the majority were posted on Twitter, this network received the lowest total number of interactions, both raw interactions and the number of interactions per post. On the other hand, the lowest number of posts were on Instagram, although it presented the highest mean and median interactions per post. An overview of the collected data from the posts of Argentinian candidates is presented in Table 6.5.

Table 6.5 – Overview of the collected data from the posts of Argentinian candidates

Posts - Argentina	Total	Min	Max	Mean	Median	Std. Dev.
Facebook Posts	3,224	-	-	-	-	-
Facebook Likes	23,550,969	0	361,940	7,305	746	17,270
Facebook Comments	4,609,286	0	97,234	1,430	99	4,148
Facebook Shares	5,155,555	0	54,282	1,599	220	3,835
Twitter Posts	10,444	-	-	-	-	-
Twitter Likes	15,427,184	0	123,390	1,477	46	4,722
Twitter Shares	4,454,052	0	43,386	426	18	1,346
Instagram Posts	1,764	-	-	-	-	-
Instagram Likes	28,249,171	13	267,786	16,014	7,515	23,513
Instagram Comments	1,353,965	0	32,296	768	276	1,764

Source: self-provided.

A total of 19,586 posts was collected from the Brazilian candidates: 6,101 from Facebook, 10,181 from Twitter, and 3,304 from Instagram. As in Argentina, the majority were posted on Twitter, although this network received the lowest number of interactions. Instagram received the lowest number of posts but the highest mean and median interactions. An overview of the collected data is presented in Table 6.6.

A total of 22,542 posts was collected from the Colombian candidates: 3,746 from Facebook, 15,996 from Twitter and 2,800 from Instagram. As in Argentina and Brazil, the majority of posts were made on Twitter, but this network received the lowest number of interactions. However, a different result was obtained regarding the mean

and median of interactions per post, as it was obtained by posts on Facebook. An overview of the collected data is shown in Table 6.7.

Table 6.6 – Overview of the collected data from the posts of Brazilian candidates

Posts - Brazil	Total	Min	Max	Mean	Median	Std. Dev.
Facebook Posts	6,101	-	-	-	-	-
Facebook Likes	81,741,405	1	726,899	13,398	4,084	29,881
Facebook Comments	12,602,311	0	798,554	2,066	417	17,403
Facebook Shares	27,192,425	0	359,513	4,457	1,047	14,324
Twitter Posts	10,181	-	-	-	-	-
Twitter Likes	21,330,941	0	107,368	2,095	358	6,125
Twitter Shares	5,493,573	0	62,105	540	110	1,560
Instagram Posts	3,304	-	-	-	-	-
Instagram Likes	101,964,341	309	1,190,258	30,861	6,531	84,142
Instagram Comments	3,444,310	0	91,469	1,042	189	4,032

Source: self-provided.

Table 6.7 – Overview of the collected data from the posts of Colombian candidates

Posts - Colombia	Total	Min	Max	Mean	Median	Std. Dev.
Facebook Posts	3,746	-	-	-	-	-
Facebook Likes	14,548,883	5	117,504	3,884	945	8,566
Facebook Comments	3,488,095	0	59,747	931	115	2,953
Facebook Shares	5,485,706	0	171,825	1,464	206	5,135
Twitter Posts	15,996	-	-	-	-	-
Twitter Likes	10,865,730	0	35,230	679	200	1,467
Twitter Shares	4,635,743	0	12,540	290	103	567
Instagram Posts	2,800	-	-	-	-	-
Instagram Likes	5,052,473	28	32,534	1,804	712	2,868
Instagram Comments	134,758	0	1,694	48	20	96

Source: self-provided.

Lastly, a total of 9,843 posts was collected from the Mexican candidates: 2,308 from Facebook, 7,146 from Twitter, and only 389 from Instagram. As previously mentioned, we did not collect data from Instagram accounts of the two most voted candidates because these were personal profiles rather than commercial. As in other countries, the majority were posted on Twitter, but the highest interactions were on Facebook. Even only collecting data for the 3rd and 4th most voted candidates, their interactions per post on Instagram were higher than the interaction per posts of all the candidates on Twitter. An overview of the collected data is presented in Table 6.8.

A total of 67,403 posts was collected from the candidate profiles in all four countries, making an average of 16,851 per country. This is a very small number of

collected data if compared to the usual approaches: the review identified an average of 12.9 million and a median of 250 thousand datapoints (usually tweets) in current studies for a single election. If we consider the median, we collected 17,509 posts per country, which is only 7.0% of the data collected by other studies. Moreover, all the required data was collected over a long period (10 months), instead of just a sample of tweets during arbitrary periods, as obtained in the most commonly used approach.

Table 6.8 – Overview of the collected data from posts of the Mexican candidates

Posts - Mexico	Total	Min	Max	Mean	Median	Std. Dev.
Facebook Posts	2,308	-	-	-	-	-
Facebook Likes	38,046,763	15	357,567	16,485	5,605	30,696
Facebook Comments	7,366,475	1	132,499	3,192	945	7,821
Facebook Shares	10,440,845	0	165,399	4,524	1,469	10,652
Twitter Posts	7,146	-	-	-	-	-
Twitter Likes	8,081,106	0	35,744	1,131	174	2,731
Twitter Shares	3,585,676	0	15,191	502	71	1,100
Instagram Posts	389	-	-	-	-	-
Instagram Likes	870,331	127	16,207	2,237	1,293	2,417
Instagram Comments	29,765	1	719	77	39	101

Source: self-provided.

*Instagram posts were only collected for the 3rd and 4th most voted candidates.

To summarize, candidates in all countries mostly posted on Twitter, but the public response on this platform is lower than the response on Facebook and Instagram. In Argentina and Brazil, most interactions were on Instagram, while in Colombia, most interactions were on Facebook. Lastly, there was a large variation in the interactions on posts since they are able to receive between zero and one million interactions, and there is a large standard deviation in all metrics.

In order to obtain an idea regarding voter engagement on SM candidate profiles in the studied countries, we compared the interaction numbers with the total population in each country in 2019 (UNITED NATIONS, 2019). To avoid bias in the analysis, because of the lack of data on Instagram for two profiles of Mexican candidates, we created two summaries: one containing all the collected data, and another containing data only from Facebook and Twitter. Results are presented in Table 6.9 and demonstrate that the Argentinians were more engaged with the profiles of the candidates and the citizens from Mexico presented the lowest level of engagement.

Table 6.9 – Comparison of engagement on SM and the total population

All platforms

Country	Argentina	Brasil	Colombia	Mexico
Population	44,781,000	211,050,000	50,339,000	127,576,000
Total Posts	15,432	19,586	22,542	9,843
Interactions	82,800,182	253,769,306	44,211,388	68,420,961
Interactions/Post	5,365	12,957	1,961	6,951
Interactions/Habitant	*1.85*	*1.20*	*0.88*	*0.54*

Only Facebook and Twitter

Country	Argentina	Brasil	Colombia	Mexico
Population	44,781,000	211,050,000	50,339,000	127,576,000
Total Posts	13,668	16,282	19,742	9,454
Interactions	53,197,046	148,360,655	39,024,157	67,520,865
Interactions/Post	3,892	9,112	1,977	7,142
Interactions/Habitant	*1.19*	*0.70*	*0.78*	*0.53*

Source: self-provided.

6.1.2.2 Data from Polls

As mentioned in the previous step regarding business understanding, identifying the pollsters was a manual task, as were the tasks of collecting and pruning the polls.

The search for pollsters was undertaken on three sources: (1) the official electoral court website, (2) the Wikipedia entry describing the elections in the official languages of the countries (Portuguese or Spanish) and in English, and (3) an open search on Google.

The search for polls was undertaken on the following sources: (1) the official electoral court website, (2) the official pollster websites, (3) aggregator websites identified on the pollster search, and (4) through an open search on Google. Polls were found on the following sources: Mexican[12] and Colombian[13] polls from the electoral court websites, Argentinian polls from pollster websites and on the Scribd document repository (scribd.com), and Brazilian polls from a website aggregator, Poder 360[14].

[12] https://computos2018.ine.mx/#/presidencia/nacional/1/1/1/1 . Viewed in October, 2020.

[13] https://www.cne.gov.co/inventario-de-encuestas . Viewed in October, 2020.

[14] https://www.poder360.com.br/banco-de-dados/ . Viewed in October, 2020.

With the exception of data from Brazil, data for each survey was collected manually, including data referring to methodology.

Pollster and poll pruning was performed according to the following inclusion criteria:

I.1 – Only pollsters which had performed at least 5 polls, due to consistency and continuity.

I. 2 – The last poll must have been performed at least 15 days before the election (or limit) date, to enable a comparison of the last poll with the election result;

I.3 – National polls, performed on the broad public, for president in the first round.

The following exclusion criteria was defined:

E.1 – Polls performed only on the internet, on platforms based on self-selection.

The exclusion criteria were defined to reinforce the third criteria, polls performed on the broad public. The aim of using traditional polls for training the model is to avoid any bias related to the non-representativity of the population on the internet and on the SM platforms. Thus, training the model with polls performed only on the internet would reintroduce this bias.

Table 6.10 presents the pollsters and number of polls considered after poll selection and pruning.

Table 6.10 – The pollsters and number of polls collected for each country

Argentina		Brazil		Colombia		Mexico	
Pollster	**Polls**	**Pollster**	**Polls**	**Pollster**	**Polls**	**Pollster**	**Polls**
Federico González & Asociados	11	Ipespe	19	Centro Nacional de Consultoria	8	Parametria	9
Ricardo Rouvier & Asociados	8	Datafolha	11	Cifras Y Conceptos	7	Grupo Impacto	8
Gustavo Córdoba y Asociados	7	IBOPE	10	Datexco Company	6	Arias Consultores	7
Circuitos	5	DataPoder360	6	Guarumo, Ecoanalítica	6	Economista y Asociados	7
		FSB	6	Invamer Sas	6	El Financiero	7
		MDA	5	Yanhaas	6	Suasor Consultores	7
		Paraná Pesquisas	5			Conteo	6
Total	31	*Total*	62	*Total*	39	*Total*	63

Source: self-provided.

Table 6.11 presents a summary of the final poll data, from the final poll performed by each pollster before the election. The lowest standard deviation was observed in Argentina, signifying that the final polls somehow agreed, and that the predictions converged. On the other hand, the final predictions from Mexico were very different, ranging from 34.0 to 62.5 percentual points for the most voted candidate. Also, the poll average indicated a wrong list of the most voted candidates, since it

indicated Kuribreña as the second most voted candidate, who in fact was in third place. This high variation in the final polls in Mexico may have interfered in the results, since this data will be used for predictions.

Table 6.11 – Summary of the final week of polls

Country	Candidate	Period	Min	Max	Mean	Median	Std. Dev.
Argentina	Fernandez	Oct. 10-18	49.50	50.90	50.15	50.10	0.60
Argentina	Macri	Oct. 10-18	29.20	32.50	31.33	31.80	1.45
Argentina	Lavagna	Oct. 10-18	7.20	9.20	7.83	7.45	0.93
Argentina	del Caño	Oct. 10-18	2.20	2.80	2.53	2.55	0.28
Argentina	Centurión	Oct. 10-18	1.20	1.70	1.48	1.50	0.21
Argentina	Espert	Oct. 10-18	0.80	1.90	1.28	1.20	0.46
Brazil	Bolsonaro	Sep. 30 - Oct.07	30.00	36.70	34.37	36.00	2.71
Brazil	Haddad	Sep. 30 - Oct.07	21.80	25.00	22.97	22.00	1.32
Brazil	Gomes	Sep. 30 - Oct.07	9.00	15.00	11.19	11.00	2.14
Brazil	Alckimin	Sep. 30 - Oct.07	5.80	11.00	7.46	7.00	1.64
Brazil	Amoêdo	Sep. 30 - Oct.07	0.00	5.00	2.63	3.00	1.50
Colombia	Márquez	May 15-18	35.00	41.50	36.92	36.20	2.45
Colombia	Petro	May 15-18	24.00	34.80	27.58	26.50	4.07
Colombia	Fajardo	May 15-18	14.00	18.00	16.05	16.00	1.27
Colombia	Lleras	May 15-18	6.00	14.30	9.02	8.30	3.34
Colombia	Calle	May 15-18	1.90	4.00	2.80	2.65	0.84
Mexico	Obrador	Jun. 18-26	34.00	62.50	41.12	38.34	8.74
Mexico	Cortés	Jun. 18-26	14.00	24.00	19.67	21.00	3.82
Mexico	Kuribreña	Jun. 18-26	13.00	29.00	19.74	15.77	6.17
Mexico	Calderon	Jun. 18-26	1.00	9.00	3.77	2.49	2.71

Source: self-provided.

*The date of the poll was considered as one day after the last day of interviews

Table 6.12 presents a summary of all poll data, which presented high variations. This, however, was expected since voting intentions may vary throughout the campaign. As an example, at the beginning of the campaign in Brazil, Haddad was initially the candidate for vice-president and Lula was the candidate for president. However, after Lula's candidacy was denied by the Superior Electoral Court, Haddad became the presidential candidate just one month before the elections. Thus, it is expected that polls in the initial months would be very different from the polls close to the election.

All polls collected data, including the official documents, methodology, and values, is available at kellyton.com.br/somen.

Table 6.12 – Summary of all polls

Country	Candidate	Period	Min	Max	Mean	Median	StdDev
Argentina	Fernandez	Jan. 12 - Oct. 18	35.70	54.50	45.07	48.10	6.11
Argentina	Macri	Jan. 12 - Oct. 18	22.00	35.20	29.32	29.91	3.72
Argentina	Lavagna	Jan. 12 - Oct. 18	5.70	19.30	9.56	8.90	3.11
Argentina	del Caño	Jan. 12 - Oct. 18	0.40	4.97	3.18	3.05	1.09
Argentina	Centurión	Jan. 12 - Oct. 18	0.70	3.00	1.48	1.40	0.62
Argentina	Espert	Jan. 12 - Oct. 18	0.80	5.82	2.71	2.33	1.57
Brazil	Bolsonaro	Jan. 31 - Oct.07	17.00	36.70	25.60	24.00	5.29
Brazil	Haddad	Jan. 31 - Oct.07	1.00	25.20	11.04	7.00	8.67
Brazil	Gomes	Jan. 31 - Oct.07	8.00	15.00	10.57	10.45	1.60
Brazil	Alckimin	Jan. 31 - Oct.07	5.30	11.00	7.98	8.00	1.18
Brazil	Amoêdo	Jan. 31 - Oct.07	0.00	5.00	2.50	3.00	1.08
Colombia	Márquez	Jul. 31 - May 18	0.40	45.90	22.85	23.60	15.97
Colombia	Petro	Jul. 31 - May 18	9.00	34.80	20.73	22.00	7.17
Colombia	Fajardo	Jul. 31 - May 18	5.00	26.00	13.93	14.00	4.33
Colombia	Lleras	Jul. 31 - May 18	5.30	15.60	8.49	7.90	2.53
Colombia	Calle	Jul. 31 - May 18	1.00	11.00	4.29	4.00	2.38
Mexico	Obrador	Sep. 20 - Jun. 26	20.80	69.90	36.63	34.90	9.82
Mexico	Cortés	Sep. 20 - Jun. 26	12.00	28.00	20.18	21.00	3.58
Mexico	Kuribreña	Sep. 20 - Jun. 26	11.00	29.00	18.61	16.95	4.97
Mexico	Calderon	Sep. 20 - Jun. 26	0.60	9.00	2.61	2.00	1.84

Source: self-provided.

6.1.3 Data Preparation

A set of 10 independent datasets was generated for each of the considered candidates, with the features described in Chapter 5.2. Each dataset was generated with a different window size, w = [1..7, 14, 21, 28] days. Each sample is based on a poll day, as illustrated in Table 6.13.

The table presents part of the dataset used for training and for the predictions of the Brazilian candidate Jair Bolsonaro, considering a window of 7 days. The first four columns (Candidate, Window, Institute and Ref. Date) are metadata used to identify the data, in order to facilitate tests and the continuous predictions needed for RQ3. To create uniformity, the poll day was considered as one day after the last date of interviews, found in the poll methodologies. The last column, “Share”, is the vote share obtained by the candidate on that poll and is multiplied by 100 to facilitate calculations. Thus, a vote share presented as 1800 is 18.00. The other columns are the features, according to what was previously presented in Table 5.1.

Table 6.13 – An example of the generated dataset

Candidate	Window	Institute	Ref. Date	FBPosts	FBLikes	FBComm	FBShares	FBLikesPPost	FBCommPPost	FBSharesPPost	TTPosts	TTLikes	TTShares	TTLikesPPost	TTSharesPPost	IGPosts	IGLikes	IGComm	IGLikesPPost	IGCommPPost	Share
Jair Bolsonaro	7	Datafolha	31/01/2018	26	1240519	140364	313436	47712	5399	12055	25	118256	35344	4730	1414	12	537259	15602	44772	1300	1800
Jair Bolsonaro	7	MDA	04/03/2018	23	537136	31882	147563	23354	1386	6416	26	80495	19540	3096	752	10	396580	8838	39658	884	2000
Jair Bolsonaro	7	MDA	13/03/2018	17	551722	124801	143647	32454	7341	8450	14	57436	12739	4103	910	8	259977	8198	32497	1025	1830
Jair Bolsonaro	7	Datafolha	14/04/2018	18	623261	36655	198629	34626	2036	11035	22	141874	40562	6449	1844	12	619894	16871	51658	1406	1700
Jair Bolsonaro	7	DataPoder360	20/04/2018	31	1191262	95279	450268	38428	3074	14525	23	136287	38895	5926	1691	17	572379	12814	33669	754	2000
Jair Bolsonaro	7	Paraná Pesquisas	03/05/2018	24	630387	69941	170844	26266	2914	7119	23	120275	25267	5229	1099	10	373893	8185	37389	819	2050
Jair Bolsonaro	7	Ipespe	24/05/2018	23	751025	54831	292605	32653	2384	12722	22	111117	24830	5051	1129	12	607902	12014	50659	1001	2400
Jair Bolsonaro	7	DataPoder360	01/06/2018	23	985157	63514	540122	42833	2761	23484	25	223428	56971	8937	2279	16	847068	17931	52942	1121	2100
Jair Bolsonaro	7	Ipespe	07/06/2018	29	548229	44408	154040	18904	1531	5312	18	84539	18959	4697	1053	13	542434	11354	41726	873	2200
Jair Bolsonaro	7	Datafolha	08/06/2018	29	558771	56770	164586	19268	1958	5675	22	109107	24140	4959	1097	15	622199	12357	41480	824	1900
Jair Bolsonaro	7	Ipespe	14/06/2018	25	628023	53212	192918	25121	2128	7717	36	174688	37469	4852	1041	19	818896	15915	43100	838	2100
Jair Bolsonaro	7	Ipespe	21/06/2018	17	378416	53157	114595	22260	3127	6741	25	126528	26983	5061	1079	14	427468	11007	30533	786	2100
Jair Bolsonaro	7	Ibope	25/06/2018	18	335941	48307	81103	18663	2684	4506	27	119663	24010	4432	889	15	465948	11173	31063	745	1700
Jair Bolsonaro	7	Ipespe	28/06/2018	20	355359	20110	95139	17768	1006	4757	31	119681	26795	3861	864	14	451836	11856	32274	847	2200
Jair Bolsonaro	7	DataPoder360	30/06/2018	21	420003	43660	127425	20000	2079	6068	34	132761	30450	3905	896	12	430274	10898	35856	908	1800
Jair Bolsonaro	7	Ipespe	05/07/2018	23	480239	63075	159265	20880	2742	6925	38	137721	35383	3624	931	13	487702	12891	37516	992	2300
Jair Bolsonaro	7	Ipespe	12/07/2018	22	425956	31511	117593	19362	1432	5345	33	101393	21068	3073	638	13	619735	14069	47672	1082	2400
Jair Bolsonaro	7	Ipespe	19/07/2018	24	476907	48591	117211	19871	2025	4884	48	142688	39947	2973	832	16	746083	16115	46630	1007	2300
Jair Bolsonaro	7	Ipespe	26/07/2018	22	549986	97714	145233	24999	4442	6602	41	186313	48491	4544	1183	18	992659	26439	55148	1469	2300
Jair Bolsonaro	7	DataPoder360	29/07/2018	27	695770	121458	185136	25769	4498	6857	32	177635	48919	5551	1529	14	936657	24081	66904	1720	2000
Jair Bolsonaro	7	Paraná Pesquisas	31/07/2018	28	677526	68906	208503	24197	2461	7447	36	192135	52735	5337	1465	12	701269	20303	58439	1692	2360
Jair Bolsonaro	7	Ipespe	02/08/2018	33	891788	113599	246521	27024	3442	7470	47	245869	70407	5231	1498	13	1050238	31410	80788	2416	2200
Jair Bolsonaro	7	Ipespe	09/08/2018	26	1277601	201985	508580	49139	7769	19561	36	338023	91212	9390	2534	14	1318968	36357	94212	2597	2300
Jair Bolsonaro	7	Paraná Pesquisas	14/08/2018	29	1149187	84475	369659	39627	2913	12747	51	484047	128350	9491	2517	19	2130405	63658	112127	3350	2390
Jair Bolsonaro	7	Ipespe	16/08/2018	30	1218287	100324	376825	40610	3344	12561	53	562319	137206	10610	2589	18	2049869	64749	113882	3597	2300
Jair Bolsonaro	7	Ibope	21/08/2018	30	1676650	148840	649116	55888	4961	21637	57	593564	140600	10413	2467	18	1881190	52963	104511	2942	2000
Jair Bolsonaro	7	Datafolha	22/08/2018	30	1551864	137864	605475	51729	4595	20183	49	490379	117512	10008	2398	21	2041469	59108	97213	2815	2200
Jair Bolsonaro	7	Ipespe	23/08/2018	31	1635194	180258	641130	52748	5815	20682	49	486544	120529	9929	2460	20	1938577	62103	96929	3105	2300
Jair Bolsonaro	7	FSB	27/08/2018	40	1304886	118027	417540	32622	2951	10439	50	489140	135450	9783	2709	20	2154060	62952	107703	3148	2400
Jair Bolsonaro	7	Ipespe	30/08/2018	49	2186276	295101	734256	44618	6022	14985	50	688080	175834	13762	3517	22	3401249	135938	154602	6179	2300
Jair Bolsonaro	7	FSB	03/09/2018	43	2128302	341869	777898	49495	7950	18091	50	714020	177761	14280	3555	26	3887571	167743	149522	6452	2600
Jair Bolsonaro	7	Ibope	05/09/2018	40	2156879	323176	849381	53922	8079	21235	61	760752	208375	12471	3416	28	4254772	172230	151956	6151	2200
Jair Bolsonaro	7	Ipespe	06/09/2018	41	1655580	196135	565609	40380	4784	13795	60	652015	184482	10867	3075	28	3643613	102697	130129	3668	2300
Jair Bolsonaro	7	FSB	10/09/2018	27	1654257	174177	374052	61269	6451	13854	41	707857	162841	17265	3972	18	3607554	135572	200420	7532	3000
Jair Bolsonaro	7	Datafolha	11/09/2018	24	1639166	174486	350252	68299	7270	14594	32	717045	150681	22408	4709	17	3613751	138512	212574	8148	2400
Jair Bolsonaro	7	Ibope	12/09/2018	23	1620306	172839	292709	70448	7515	12726	30	762688	161024	25423	5367	17	3821001	146884	224765	8640	2600
Jair Bolsonaro	7	Ipespe	13/09/2018	20	1662463	131057	297505	83123	6553	14875	26	789300	161749	30358	6221	16	4351410	165258	271963	10329	2600
Jair Bolsonaro	7	Datafolha	15/09/2018	21	1739319	132170	304966	82825	6294	14522	25	802584	164671	32103	6587	18	5116625	168240	284257	9347	2600
Jair Bolsonaro	7	MDA	16/09/2018	25	2239618	188919	406544	89585	7557	16262	33	987708	201047	29931	6092	21	6131259	216621	291965	10315	2820
Jair Bolsonaro	7	FSB	17/09/2018	28	3241610	807893	854967	115772	28853	30535	40	1046685	232978	26167	5824	23	6999159	233346	304311	10145	3300
Jair Bolsonaro	7	Ibope	19/09/2018	30	3573053	845841	882049	119102	28195	29402	38	1030300	220968	27113	5815	24	7745563	273321	322732	11388	2800
Jair Bolsonaro	7	Datafolha	20/09/2018	26	3368854	835961	821256	129571	32152	31587	39	1030263	222822	26417	5713	22	7668873	283459	348585	12885	2800
Jair Bolsonaro	7	Ipespe	20/09/2018	26	3368854	835961	821256	129571	32152	31587	39	1030263	222822	26417	5713	22	7668873	283459	348585	12885	2800
Jair Bolsonaro	7	DataPoder360	21/09/2018	30	3664016	870115	1001058	122134	29004	33369	42	1069626	239146	25467	5694	24	8350473	348901	347936	14538	2600
Jair Bolsonaro	7	FSB	24/09/2018	31	3020139	291951	970218	97424	9418	31297	47	1237404	279101	26328	5938	23	8170704	353271	355248	15360	3300
Jair Bolsonaro	7	Ibope	25/09/2018	33	3094873	296943	1018722	93784	8998	30870	49	1276142	285432	26044	5825	23	8574361	365774	372798	15903	2800
Jair Bolsonaro	7	Paraná Pesquisas	26/09/2018	33	2883307	271172	1008228	87373	8217	30552	58	1375187	314499	23710	5422	24	8245189	341546	343550	14231	3120
Jair Bolsonaro	7	Ibope	27/09/2018	36	3059744	275569	1063377	84993	7655	29538	59	1408157	327294	23867	5547	25	7867717	321998	314709	12880	2700

Source: self-provided.

One additional sample was also generated, correlating SM features and the final vote share instead of the poll data. This sample is not used in the training, although the features are used for an out-of-sample prediction. This sample was also used to find correlations between the SM features and the election results, in order to answer RQ1, and to calculate the error measures.

In order to generate the final datasets to be used for training and predictions, the metadata were removed, and the features were normalized. Thus, as planned, due to the few training samples (from 31 in Argentina to 63 in Mexico) for the number of features (17), PCA was applied to each dataset independently. Component selection was set to cover a variance higher than 95%, a well-accepted level in the research literature. Thus, the number of components varied from four to seven in a 1-day window, and from two to four in a 28-day window, drastically reducing the number of input features to prevent the well-known problems of high dimensionality and collinearity. To compare and verify the gain obtained by applying PCA, experiments were applied twice: with and without the application of PCA.

6.1.4 Modeling

Three models were used for predictions. The MLP-BP artificial neural network, an alternative, well suited model, GRNN, and a baseline model, linear regression. The choices of MLP-BP and GRNN are justified in Chapter 5.

For the MLP-BP model, to avoid the well-known problem of local minima, a committee was created of 5 identical MLP-BPs trained using different random seeds, the results of which were averaged to give a unique output. Moreover, we initially considered two strategies for the parameter selection and optimization: manual parameter selection and grid search (LERMAN, 1980) for parameters.

For the manual parameter selection, data characteristics were considered, mainly the small number of samples, and the following parameters were chosen: one hidden layer with three neurons, to avoid overfitting; L-FBGS as a solver, which performs well with small samples; an alpha set to 0.05 for fast convergence, a constant learning rate for fast training, and logistic activation.

For the grid search, the considered parameters are presented in Table 6.14.

Both parameter selection approaches were applied in the preliminary version of this study regarding the 2018 Brazilian and 2016 U.S. elections (BRITO, K. dos S.;

ADEODATO, 2020). The data split for grid search used a time series split approach. This is a variation of the K-Fold (RODRIGUEZ; PEREZ; LOZANO, 2010) selection: in the *k*-th split, it returns the first *k* folds as the training set and the *(k+1)*th fold as the test set. Our preliminary results demonstrate that although the grid search parameters approach increased the execution time by 240 times (4 * 2 * 3 * 5 * 2 executions), it did not increase the prediction accuracy when compared with these manually selected parameters. Thus, the fixed parameter approach, with the abovementioned parameters, was used.

Table 6.14 – Values for the ANN grid search parameters

Parameter	Values
Hiddel Layer Sizes	3, 4, 5, 10
Activation Function	Logistic, Tanh
Solver	SGD, L-BFGS, ADAM
Alpha	0.00001, 0.001, 0.01, 0.05, 0.1
Learning Rate	Constant, Adaptive

Source: self-provided.

In the GRNN model, there is only one hyperparameter to be adjusted, the smoothing parameter. As there was no baseline value for this parameter, it was found by the grid search approach, varying from 0.1 to 4, in steps of 0.2.

To enable comparisons, a baseline technique, linear regression, was also applied in the same datasets.

As planned, a committee machine was created for each model, composed of an ensemble of 10 predictors. As input, each received a different dataset related to a different aggregated window size, and all predictions were averaged as the final prediction for that candidate. The machines were trained with all the available poll data until one day before the elections, and one prediction was made for the final vote share. The results were then compared with the actual vote share of each candidate.

To summarize, six sets of experiments were undertaken: linear regression, linear regression with PCA, MLP-BP, MLP-BP with PCA, GRNN, and GRNN with PCA.

6.1.5 Evaluation

All experiments were run in a standard laptop computer with the following configuration: Processor Intel Core i7-8550U, RAM 16GB, data on an SDD 128GB disk, O.S. Windows 10. The implementation was in python using Scikit-learn

(PEDREGOSA *et al.*, 2011) for most computations and pyGRNN[15] for GRNN implementation. All the outputs, including detailed outputs for each window size, are available at kellyton.com.br/somen.

The evaluation was performed according to the definitions in Chapters 4 and 5, with the collection and comparison of predictions, prediction errors, and poll errors. Statistical tests were also performed and are detailed in section 6.3.

6.2 THE SOMEN-DC EXECUTION

The execution of the SoMEN-DC was based on the SoMEN execution. The main difference is that instead of predicting only the final results, it made daily predictions, using the available data until one day before. The execution strategy consisted of the following stages:

1 – Initial setup: first training with 10 polls with data on all candidates;

2 – Begin daily predictions after the initial setup, considering SM and poll data until *D-1*;

3 – Out-of-sample predictions: the prediction of an arbitrary day *D* considers polls before *D*;

4 – When new poll data is available, retrain the model with the inclusion of this new data.

5 – Error metrics are based on predictions on days with available poll data.

The evaluation was performed according to the definitions in Chapters 4 and 5, and are detailed in section 6.3.

6.3 EXPERIMENT RESULTS

6.3.1 Research Question 1

After Phase 3, data preparation, we attempted to refute "H_1': It is not possible to model the SM performance based on the interactions of users on the official profiles of candidates and find correlations between the SM and the electoral performances of candidates*."* In addition to the ten datasets, we wished to investigate whether there

[15] Available at: https://github.com/federhub/pyGRNN . Viewed on: October 5, 2020.

were any correlations with other window sizes, for example, with the entire 10-month period. Thus, we generated additional datasets adding intervals of 30 days and ran Pearson correlation tests considering all the generated datasets.

For Argentina, 6 candidates were considered. Thus, based on the rules of thumb for Pearson correlation, set out in the methodology, the adopted thresholds were *r* >= .71 for showing a correlation, r >= .76 for a high correlation, and r >= .82 for a very high correlation. Table 6.15 presents the results. Cells in dark gray highlight the values above the highest threshold, and light gray highlights the values higher than the lowest threshold. The last column presents a simple average of the results from all windows, and the table is ordered by this average.

Table 6.15 – Pearson correlation results for the SM performance and the Argentinian election results

Metric - Argentina	Number of Days																				Average
	1	2	3	4	5	6	7	14	21	28	30	60	90	120	150	180	210	240	270	300	
IGLikesPPost	0.62	0.87	0.94	0.91	0.90	0.94	0.94	0.95	0.95	0.94	0.94	0.92	0.91	0.90	0.91	0.92	0.93	0.93	0.93	0.93	0.91
TTSharesPPost	0.71	0.80	0.94	0.93	0.92	0.83	0.82	0.92	0.93	0.95	0.93	0.93	0.90	0.91	0.92	0.86	0.81	0.76	0.73	0.69	0.86
FBSharesPPost	0.66	0.68	0.73	0.78	0.81	0.87	0.87	0.91	0.95	0.95	0.95	0.95	0.92	0.93	0.93	0.85	0.85	0.84	0.84	0.82	0.85
TTLikesPPost	0.70	0.82	0.97	0.95	0.94	0.81	0.81	0.91	0.92	0.94	0.92	0.92	0.88	0.88	0.90	0.84	0.79	0.74	0.71	0.67	0.85
FBLikesPPost	0.65	0.64	0.69	0.70	0.72	0.80	0.80	0.89	0.93	0.91	0.90	0.90	0.86	0.88	0.88	0.81	0.81	0.81	0.80	0.79	0.81
IGCommentsPPost	0.36	0.78	0.85	0.82	0.83	0.91	0.92	0.90	0.91	0.88	0.87	0.81	0.77	0.75	0.76	0.78	0.78	0.78	0.78	0.77	0.80
TTShares	0.68	0.85	0.71	0.76	0.76	0.72	0.72	0.75	0.76	0.76	0.76	0.79	0.86	0.90	0.90	0.88	0.87	0.86	0.85	0.83	0.80
TTLikes	0.65	0.85	0.75	0.78	0.77	0.69	0.70	0.74	0.75	0.75	0.74	0.78	0.84	0.87	0.87	0.85	0.84	0.83	0.82	0.80	0.78
FBCommentsPPost	0.11	0.61	0.72	0.70	0.74	0.86	0.87	0.90	0.91	0.89	0.88	0.86	0.79	0.82	0.82	0.75	0.75	0.74	0.73	0.72	0.76
FBShares	0.66	0.65	0.62	0.67	0.70	0.75	0.74	0.75	0.75	0.76	0.75	0.77	0.77	0.78	0.77	0.74	0.74	0.73	0.72	0.72	0.73
IGLikes	0.62	0.87	0.74	0.74	0.72	0.70	0.68	0.69	0.68	0.69	0.68	0.70	0.74	0.75	0.75	0.73	0.72	0.71	0.70	0.70	0.72
FBLikes	0.65	0.63	0.60	0.62	0.64	0.70	0.70	0.73	0.73	0.73	0.71	0.73	0.72	0.74	0.73	0.71	0.71	0.70	0.70	0.70	0.69
FBComments	0.11	0.61	0.62	0.63	0.65	0.75	0.74	0.74	0.72	0.71	0.70	0.70	0.68	0.70	0.69	0.67	0.67	0.66	0.65	0.65	0.65
IGComments	0.36	0.78	0.67	0.67	0.67	0.68	0.67	0.66	0.65	0.65	0.64	0.64	0.64	0.65	0.65	0.64	0.63	0.62	0.62	0.61	0.64
IGPosts	0.03	0.55	0.41	0.44	0.45	0.45	0.43	0.51	0.51	0.57	0.57	0.59	0.60	0.63	0.58	0.50	0.46	0.45	0.43	0.42	0.48
FBPosts	0.29	0.80	0.71	0.72	0.71	0.61	0.58	0.27	0.29	0.37	0.40	0.26	0.14	0.13	0.13	0.27	0.24	0.30	0.32	0.37	0.40
TTPosts	-0.36	-0.41	-0.04	-0.18	-0.37	-0.35	-0.36	-0.24	-0.13	-0.06	-0.06	-0.32	-0.43	-0.47	-0.53	-0.51	-0.47	-0.38	-0.33	-0.23	-0.31

Source: self-provided.

The data reveals that many of the defined variables presented a very high correlation with the results of the Argentinian elections, with emphasis on the variables related to the number of interactions per post on all three networks. Higher correlations were concentrated in the windows between 21 and 30 days. In opposition to this, the absolute number of posts on all three networks presented the lowest correlations, and the number of tweets presented a negative correlation. Moreover, the correlations considering the window of 1 day were notably lower than the other windows.

For Brazil and Colombia, 5 candidates were considered. Thus, the adopted thresholds were r >= .76 for showing a correlation, r >= .82 for a high correlation, and r >= .89 for a very high correlation. Table 6.16 and 6.17 present the results for Brazil and Colombia, respectively.

In the Brazilian scenario, although there was no correlation between the number of posts and votes, all the other variables presented a correlation on most windows,

and most variables presented at least a high correlation. However, a clear pattern could not be observed.

Table 6.16 – Pearson correlation results for the SM performance and the Brazilian election results

Metric - Brazil	Number of Days																			Average
	1	2	3	4	5	6	7	14	21	28	30	60	90	120	150	180	210	240	270	300	
FBComments	0.92	0.94	0.89	0.89	0.89	0.88	0.88	0.90	0.91	0.92	0.92	0.91	0.90	0.90	0.89	0.89	0.88	0.88	0.88	0.87	0.90
IGLikes	0.87	0.85	0.86	0.87	0.87	0.87	0.87	0.87	0.88	0.89	0.89	0.88	0.88	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
IGComments	0.83	0.87	0.87	0.88	0.88	0.88	0.88	0.87	0.88	0.88	0.88	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.86	0.87
IGLikesPPost	0.86	0.85	0.85	0.85	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.85	0.85	0.86	0.87	0.87	0.88	0.88	0.88	0.88	0.86
TTShares	0.86	0.77	0.80	0.87	0.86	0.87	0.88	0.87	0.88	0.88	0.88	0.87	0.87	0.86	0.86	0.85	0.85	0.85	0.84	0.84	0.86
FBCommentsPPost	0.84	0.85	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.85	0.86	0.84	0.84	0.85	0.85	0.85	0.85	0.85
FBSharesPPost	0.87	0.85	0.85	0.86	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.83	0.83	0.84	0.81	0.81	0.81	0.82	0.81	0.82	0.84
FBLikesPPost	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.84	0.83	0.84	0.84	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.84	0.84	0.84
IGCommentsPPost	0.87	0.85	0.84	0.84	0.83	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
TTLikes	0.80	0.70	0.74	0.82	0.83	0.83	0.84	0.85	0.86	0.87	0.87	0.86	0.86	0.86	0.85	0.85	0.84	0.84	0.84	0.84	0.83
FBShares	0.97	0.87	0.91	0.92	0.93	0.92	0.92	0.88	0.88	0.87	0.86	0.74	0.74	0.73	0.73	0.74	0.74	0.74	0.75	0.75	0.83
FBLikes	0.93	0.90	0.91	0.90	0.90	0.88	0.88	0.84	0.84	0.84	0.83	0.74	0.74	0.74	0.74	0.75	0.75	0.75	0.76	0.76	0.82
TTLikesPPost	0.77	0.83	0.83	0.84	0.84	0.79	0.79	0.80	0.81	0.82	0.82	0.82	0.81	0.80	0.80	0.80	0.79	0.79	0.78	0.77	0.81
TTSharesPPost	0.76	0.81	0.82	0.83	0.83	0.78	0.79	0.79	0.80	0.81	0.81	0.81	0.80	0.79	0.78	0.78	0.78	0.77	0.76	0.76	0.79
FBPosts	0.34	0.21	0.19	0.22	0.23	0.23	0.24	0.21	0.16	0.11	0.10	0.08	0.04	0.00	0.00	0.00	-0.02	-0.02	0.00	0.00	0.12
TTPosts	0.41	0.17	0.14	0.18	0.15	0.16	0.18	0.03	-0.01	-0.07	-0.07	-0.09	-0.08	-0.08	-0.08	-0.10	-0.11	-0.11	-0.11	-0.10	0.02
IGPosts	0.12	0.00	0.00	0.06	0.07	0.08	0.10	0.08	0.06	0.03	0.03	0.00	-0.03	-0.05	-0.06	-0.06	-0.07	-0.07	-0.06	-0.05	0.01

Source: self-provided.

Table 6.17 – Pearson correlation results for the SM performance and the Colombian election results

Metric - Colombia	Number of Days																			Average
	1	2	3	4	5	6	7	14	21	28	30	60	90	120	150	180	210	240	270	300	
FBComments	0.72	0.84	0.91	0.82	0.83	0.88	0.96	0.83	0.86	0.88	0.87	0.87	0.86	0.80	0.80	0.80	0.79	0.79	0.79	0.77	0.83
FBCommentsPPost	0.76	0.79	0.81	0.83	0.83	0.86	0.90	0.92	0.92	0.91	0.91	0.90	0.88	0.84	0.83	0.80	0.76	0.72	0.69	0.65	0.83
IGComments	0.35	0.35	0.47	0.56	0.61	0.73	0.79	0.88	0.94	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.92	0.92	0.92	0.92	0.80
IGCommentsPPost	0.72	0.79	0.89	0.92	0.94	0.95	0.92	0.96	0.90	0.84	0.85	0.82	0.78	0.72	0.66	0.62	0.58	0.55	0.51	0.47	0.77
IGLikesPPost	0.52	0.58	0.66	0.69	0.72	0.76	0.78	0.81	0.81	0.79	0.82	0.90	0.90	0.88	0.83	0.79	0.77	0.73	0.69	0.64	0.75
IGLikes	0.29	0.28	0.35	0.38	0.40	0.48	0.52	0.62	0.77	0.82	0.84	0.88	0.92	0.93	0.92	0.92	0.91	0.91	0.91	0.91	0.70
FBLikesPPost	0.99	0.91	0.77	0.71	0.71	0.63	0.61	0.58	0.58	0.56	0.58	0.65	0.71	0.73	0.71	0.66	0.61	0.58	0.54	0.52	0.67
FBSharesPPost	0.95	0.91	0.74	0.67	0.68	0.60	0.58	0.57	0.56	0.54	0.54	0.55	0.60	0.71	0.70	0.69	0.66	0.62	0.59	0.53	0.65
TTShares	0.81	0.56	0.50	0.48	0.53	0.58	0.55	0.53	0.57	0.54	0.52	0.53	0.50	0.50	0.51	0.51	0.51	0.51	0.51	0.50	0.54
TTSharesPPost	0.53	0.42	0.31	0.29	0.30	0.32	0.33	0.35	0.38	0.42	0.43	0.52	0.51	0.52	0.53	0.52	0.49	0.48	0.48	0.47	0.43
FBShares	0.59	0.27	0.21	0.24	0.25	0.27	0.28	0.29	0.32	0.35	0.33	0.32	0.42	0.57	0.59	0.63	0.65	0.65	0.64	0.60	0.42
FBLikes	0.34	0.19	0.18	0.22	0.24	0.27	0.28	0.26	0.32	0.34	0.34	0.34	0.45	0.55	0.55	0.56	0.55	0.55	0.55	0.54	0.38
TTLikesPPost	0.22	0.10	0.10	0.10	0.14	0.17	0.18	0.24	0.28	0.32	0.33	0.38	0.39	0.40	0.40	0.39	0.37	0.35	0.35	0.34	0.28
TTLikes	0.20	0.01	0.05	0.02	0.10	0.19	0.25	0.31	0.36	0.37	0.35	0.35	0.34	0.34	0.34	0.34	0.34	0.33	0.33	0.32	0.26
IGPosts	-0.19	-0.39	-0.20	-0.29	-0.23	-0.20	-0.20	-0.26	-0.18	-0.15	-0.18	-0.29	-0.25	-0.18	-0.08	0.04	0.07	0.15	0.21	0.30	-0.13
TTPosts	0.05	0.09	0.13	0.11	0.08	0.10	0.14	0.06	-0.14	-0.33	-0.40	-0.84	-0.82	-0.81	-0.78	-0.67	-0.54	-0.35	-0.24	-0.17	-0.27
FBPosts	-0.87	-0.89	-0.88	-0.88	-0.85	-0.87	-0.85	-0.91	-0.95	-0.96	-0.96	-0.95	-0.94	-0.94	-0.93	-0.90	-0.88	-0.86	-0.83	-0.80	-0.89

Source: self-provided.

The performance of the SM variables regarding the Colombian elections was lower than the results for Argentina and Brazil. The highest correlations were related to comments, both the absolute number as well as the number averaged per post, on Facebook and Instagram. The lowest correlations were related to Twitter, and a high negative correlation was related to the absolute number of posts on Facebook.

For Mexico, only 4 candidates were considered. Thus, the adopted thresholds were r >= .82 for showing a correlation, r >= .89 for a high correlation, and r = 1.0 for a very high correlation. Table 6.18 presents the results.

As in the case of Argentina, the number of interactions per post presented the highest correlations in all windows, close to one, especially when considering the Twitter platform. In agreement with results in other countries, the total number of posts presented a negative correlation. As expected, the data related to Instagram presented

high negative correlations because it was not possible to collect Instagram data from the accounts of the most voted candidates. As this data is not complete, it has been struck through in Table 6.18.

Table 6.18 – Pearson correlation results for the SM performance and the Mexican election results

Metric - Mexico	Number of Days																				Average
	1	2	3	4	5	6	7	14	21	28	30	60	90	120	150	180	210	240	270	300	
TTLikesPPost	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
TTSharesPPost	0.93	0.95	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.97
FBSharesPPost	0.98	0.94	0.89	0.87	0.90	0.65	0.70	0.88	0.84	0.90	0.92	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.90
FBLikesPPost	0.92	0.86	0.84	0.86	0.81	0.77	0.77	0.70	0.66	0.66	0.69	0.77	0.82	0.85	0.85	0.86	0.86	0.85	0.84	0.84	0.80
TTLikes	0.98	0.89	0.94	0.86	0.94	0.93	0.93	0.57	0.57	0.59	0.59	0.73	0.70	0.75	0.70	0.69	0.66	0.67	0.69	0.71	0.75
TTShares	0.96	0.80	0.77	0.59	0.69	0.73	0.73	0.22	0.23	0.29	0.29	0.44	0.40	0.45	0.44	0.44	0.43	0.45	0.46	0.48	0.51
FBShares	0.93	0.43	-0.21	-0.35	-0.15	-0.42	-0.32	-0.21	-0.22	0.21	0.36	0.86	0.94	0.95	0.96	0.97	0.98	0.98	0.97	0.97	0.43
FBLikes	0.81	0.30	0.10	-0.05	0.02	-0.21	-0.09	-0.09	-0.06	0.07	0.13	0.32	0.42	0.48	0.53	0.58	0.62	0.65	0.65	0.67	0.29
FBCommentsPPost	0.38	0.48	0.31	0.38	0.33	0.12	0.08	0.04	0.02	0.04	0.07	0.12	0.16	0.20	0.24	0.27	0.29	0.30	0.27	0.30	0.22
FBComments	0.20	0.01	-0.44	-0.55	-0.54	-0.79	-0.85	-0.39	-0.32	-0.29	-0.27	-0.13	-0.09	-0.07	-0.04	-0.03	0.01	0.03	0.02	0.05	-0.23
~~IGCommentsPPost~~	~~-0.32~~	~~-0.38~~	~~-0.41~~	~~-0.38~~	~~-0.38~~	~~-0.39~~	~~-0.40~~	~~-0.40~~	~~-0.39~~	~~-0.40~~	~~-0.40~~	~~-0.38~~	~~-0.39~~	~~-0.39~~	~~-0.40~~	~~-0.42~~	~~-0.43~~	~~-0.44~~	~~-0.45~~	~~-0.45~~	~~-0.40~~
~~IGLikesPPost~~	~~-0.41~~	~~-0.44~~	~~-0.45~~	~~-0.43~~	~~-0.41~~	~~-0.42~~	~~-0.42~~	~~-0.39~~	~~-0.38~~	~~-0.38~~	~~-0.39~~	~~-0.39~~	~~-0.40~~	~~-0.40~~	~~-0.40~~	~~-0.41~~	~~-0.42~~	~~-0.42~~	~~-0.43~~	~~-0.43~~	~~-0.41~~
~~IGComments~~	~~-0.47~~	~~-0.57~~	~~-0.68~~	~~-0.64~~	~~-0.67~~	~~-0.72~~	~~-0.73~~	~~-0.74~~	~~-0.75~~	~~-0.75~~	~~-0.75~~	~~-0.69~~	~~-0.62~~	~~-0.61~~	~~-0.58~~	~~-0.58~~	~~-0.56~~	~~-0.56~~	~~-0.56~~	~~-0.56~~	~~-0.64~~
~~IGLikes~~	~~-0.67~~	~~-0.67~~	~~-0.73~~	~~-0.70~~	~~-0.72~~	~~-0.74~~	~~-0.75~~	~~-0.73~~	~~-0.74~~	~~-0.73~~	~~-0.74~~	~~-0.70~~	~~-0.64~~	~~-0.63~~	~~-0.59~~	~~-0.57~~	~~-0.53~~	~~-0.53~~	~~-0.53~~	~~-0.53~~	~~-0.66~~
~~IGPosts~~	~~-0.73~~	~~-0.74~~	~~-0.72~~	~~-0.73~~	~~-0.71~~	~~-0.70~~	~~-0.70~~	~~-0.70~~	~~-0.69~~	~~-0.69~~	~~-0.69~~	~~-0.71~~	~~-0.73~~	~~-0.74~~	~~-0.75~~	~~-0.76~~	~~-0.77~~	~~-0.77~~	~~-0.77~~	~~-0.77~~	~~-0.73~~
FBPosts	-0.76	-0.99	-0.95	-0.94	-0.91	-0.87	-0.80	-0.90	-0.89	-0.87	-0.86	-0.76	-0.53	-0.51	-0.46	-0.52	-0.51	-0.52	-0.53	-0.58	-0.73
TTPosts	-0.58	-0.75	-0.95	-0.77	-0.75	-0.74	-0.75	-0.81	-0.79	-0.80	-0.80	-0.79	-0.79	-0.77	-0.77	-0.78	-0.78	-0.80	-0.82	-0.85	-0.78

Source: self-provided.

Considering that most of the defined metrics presented a correlation with the electoral results in all four countries, the presented data therefore **rejects** the null hypothesis H_1', which validates the alternative hypothesis "H_1: It is possible to model the SM performance based on the interactions of users on the official profiles of candidates and find correlations between the SM and the electoral performances of candidates", and answers affirmatively the "RQ1: Is there a correlation between the SM performance of candidates and their electoral performance?"

In addition to rejecting the null hypothesis, other conclusions may also be drawn. Even being of the same region and having similar characteristics, each electoral context and use of SM platforms is particular. In Argentina, the metrics with the highest correlation with votes were the number of likes and shares per post on all three platforms, the same as Mexico (excluding Instagram). On the other hand, in Colombia, the most correlated metrics were related to comments: the absolute number of comments and the number of comments per posts, chiefly on Facebook and Instagram. In Brazil, all platforms presented high correlations, but interactions per post on Twitter presented lower correlations, an opposite result to that of Mexico.

As this data can only be obtained after the official election results, we conclude that we should not define, a priori, the most suitable SM platform or metrics for predicting elections, since each election may be more related to a specific platform or metric, which may be discovered only after the election. Thus, it reinforces the argument of this thesis that data from all the most used platforms must be collected,

and the prediction model must be sufficiently flexible to use all these data and be capable of identifying, throughout the process, the most suitable features. This approach was used in the prediction model. As these correlations were only known after the elections, we used all the defined metrics for the prediction procedure.

6.3.2 Research Question 2

After running all the phases of the process, we attempted to refute "H_2': It is not possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of predicting election results with competitive results to traditional polls." For this, we compared our predictions with the actual results, the MAE error of our predictions with the historical threshold of 2.7 (standard deviation of 2.13), and also with the errors obtained by each pollster and by the poll average. We then performed two statistical tests, by comparing the predictions and the electoral results, to verify whether they were in accordance with the results, and the predicted errors with the errors obtained by the poll average. For all the measures, we considered raw results, since calculating the relative percentual may bias the results: mobilization on the candidate profiles was low at the beginning of the campaign, as was the number of people who had decided upon their vote. Thus, the sum would not reach 100%.

In Argentina, the election was held on October 27, 2019, but the final poll prediction was from October 18. Thus, our predictions were performed with data trained until 10 days before the elections. Table 6.19 presents the predictions and final vote share, as well as the average of the poll predictions. Table 6.20 presents the error metrics, including each pollster individually and the poll average.

Table 6.19 – Predictions for the Argentinian elections

Candidate	Election result	Polls Average	Predictions					
			LR	LR PCA	MLP	MLP PCA	GRNN	GRNN PCA
Fernandez	47.02	50.15	53.27	51.73	47.44	47.00	47.41	46.82
Macri	39.27	31.33	43.07	34.39	31.14	31.14	30.70	30.66
Lavagna	5.99	7.83	8.74	8.25	7.68	7.75	7.95	7.96
Caño	2.10	2.53	3.95	2.59	2.70	2.33	2.76	2.77
Centurión	1.66	1.48	1.89	1.34	1.31	1.20	0.94	0.92
Espert	1.43	1.28	7.07	2.41	2.78	2.79	1.86	1.91

Source: self-provided.

Table 6.20 – Error metrics for the Argentinian predictions

Argentina - Error Comparisons	MAE	MAPE	RMSE	AEOM
Ricardo Rouvier & Asociados	1.89	0.15	3.00	9.25
MLP-BP PCA	1.99	0.31	3.45	8.11
MLP-BP	2.09	0.32	3.45	8.55
GRNN PCA	2.11	0.28	3.64	8.41
GRNN	2.12	0.27	3.62	8.96
LinearRegression PCA	2.27	0.29	2.96	9.59
Poll Average	*2.28*	*0.17*	*3.57*	*11.08*
Circuitos	2.41	0.26	3.50	11.35
Gustavo Córdoba y Asociados	2.46	0.21	3.58	10.75
Federico González & Asociados	2.53	0.16	4.33	12.95
LinearRegression	3.42	0.94	4.01	2.45

Source: self-provided.

Analyzing the Argentinian results, one pollster obtained a remarkable result, with an MAE of 1.89, but with a AEOM higher than our models. All MLP-BP and GRNN models obtained better MAE results than the other pollsters, and were also better than the poll average. All predictions, except the linear regression, were under the historical MAE threshold of 2.70.

In Brazil, the election was held on October 7, 2018, and we considered the date of the final predictions until October 6. Thus, our predictions were performed with data trained until 1 day before the elections. Table 6.21 presents our predictions, the average of poll predictions and the final vote share. Table 6.22 presents the error metrics.

Table 6.21 – Predictions for the Brazilian elections

Candidate	Election result	Polls Average	Predictions					
			LR	LR PCA	MLP	MLP PCA	GRNN	GRNN PCA
Bolsonaro	32.63	34.37	37.90	34.69	30.65	32.13	34.57	34.56
Haddad	20.75	22.97	36.37	30.76	20.68	20.93	23.77	23.78
Gomes	8.84	11.19	14.24	11.38	10.83	10.85	10.04	10.00
Alckmin	3.37	7.46	6.14	7.26	7.27	7.04	7.20	7.53
Amoêdo	1.77	2.63	2.25	2.63	2.87	2.58	2.41	2.44

Source: self-provided.

Analyzing the Brazilian results, the MLP models obtained the best results. Indeed, the prediction errors for the two most voted candidates using the MLP-BP PCA model were less than or equal to 0.5 percentage points. The GRNN models also obtained a good performance, better than the poll average and most of the pollsters. Moreover, all results generated by the MLP and GRNN models were under the MAE

threshold of 2.70, although only four of the 7 pollster results were under this threshold. Lastly, the linear regression models obtained the poorest results, as expected.

Table 6.22 – Error metrics from the Brazilian predictions

Brazil - Error Comparisons	MAE	MAPE	RMSE	AEOM
MLP-BP PCA	1.43	0.36	1.92	0.68
MLP-BP	1.81	0.41	2.20	1.91
Paraná Pesquisas	1.85	0.43	2.22	1.22
Ibope	2.13	0.32	2.48	2.12
GRNN	2.13	0.37	2.42	1.08
GRNN PCA	2.19	0.39	2.53	1.10
Poll Average	2.25	0.43	2.49	0.48
MDA	2.27	0.28	2.62	0.82
Ipespe	2.33	0.44	2.54	2.12
Datafolha	2.73	0.48	3.00	2.12
FSB	3.18	0.86	4.05	4.88
Datapoder 360	3.69	0.61	3.98	6.88
LinearRegression PCA	3.87	0.49	5.04	7.95
LinearRegression	5.91	0.52	7.86	10.35

Source: self-provided.

In Colombia, the election was held on May 27, 2018, but the final poll prediction was from May 20. Thus, our predictions were performed with data trained until one week before the elections. Table 6.23 presents our predictions, the average of poll predictions and the final vote share. Table 6.24 presents the error metrics.

Table 6.23 – Predictions for the Colombian elections

Candidate	Election result	Polls Average	Predictions					
			LR	LR PCA	MLP	MLP PCA	GRNN	GRNN PCA
Márquez	39.34	36.92	61.59	56.71	36.73	37.85	38.29	38.43
Petro	25.08	27.58	36.38	28.87	26.36	26.41	27.16	27.13
Fajardo	23.78	16.05	64.85	14.22	12.52	13.65	15.57	15.53
Lleras	7.3	9.02	7.32	7.95	8.52	8.54	8.89	8.84
Calle	2.05	2.80	4.64	4.44	2.85	3.01	2.61	2.58

Source: self-provided.

Analyzing the Colombian results, the GRNN models obtained the best results, and were the only results below the threshold of 2.7. However, all results from the MLP-BP were within one historical standard deviation (4.83). The MAE with MLP-BP model with PCA where better than four of the six pollsters, very close to the MAE of the poll average. The MAE with the MLP-BP without PCA was in the middle: lower than 3 pollsters and higher than the other three, thereby presenting competitive results with

the pollsters. Finally, the linear regression models obtained the poorest results, as expected.

Table 6.24 – Error metrics from the Colombian predictions

Colombia - Error Comparisons	MAE	MAPE	RMSE	AEOM
GRNN PCA	2.66	0.18	3.89	2.96
GRNN	2.70	0.19	3.89	3.13
Guarumo, Ecoanalítica	2.87	0.21	3.95	0.86
Invamer SAS	2.98	0.14	4.02	2.26
Poll Average	3.02	0.22	3.88	4.92
MLP-BP PCA	3.03	0.23	4.67	2.82
Centro Nacional de Consultoria	3.14	0.35	3.45	5.26
MLP-BP	3.43	0.23	5.24	3.89
Yanhaas	3.46	0.24	4.86	5.26
Datexto Company	4.32	0.19	5.74	12.66
Cifras y Conceptos	4.33	0.43	5.13	3.26
LinearRegression PCA	6.75	0.45	9.10	13.58
LinearRegression	15.45	0.80	21.52	10.95

Source: self-provided.

In Mexico, the election was held on July 1, 2018, but the final poll prediction was from June 26. Thus, our predictions were performed with data trained until one week before the elections. Table 6.25 presents our predictions, the average of poll predictions and the final vote share. Table 6.26 presents the error metrics.

Table 6.25 – Predictions for the Mexican elections

Candidate	Election result	Polls Average	Predictions					
			LR	LR PCA	MLP	MLP PCA	GRNN	GRNN PCA
Obrador	53.19	41.12	40.58	37.96	36.74	37.66	38.87	39.06
Cortés	22.28	19.67	18.36	20.10	19.42	20.20	20.21	20.21
Kuribreña	16.41	19.74	18.46	17.92	17.55	17.31	18.34	18.36
Calderon	5.23	3.77	6.75	3.66	2.85	3.00	2.93	2.90

Source: self-provided.

Poll and prediction errors in Mexico were higher than in the other countries. None of the polls or predictions achieved an MAE even close to the threshold of 2.7, and only two of the nine pollsters achieved errors within one historical standard deviation (4.83). Some pollsters presented an MAE higher than 7.0, which is unacceptable for this context. Also, despite a final difference of 5.87 percentual points between the second and third most voted candidates, most of the final polls and the poll average missed the ranking of candidates. The polls marked in the table with a star indicated the third most voted candidate, Kuribreña, as being the second most voted.

Table 6.26 – Error metrics from the Mexican predictions

Mexico - Error Comparisons	MAE	MAPE	RMSE	AEOM
Parametria	3.41	0.13	4.48	4.91
Reforma	3.59	0.19	5.61	10.99
*Poll Average**	*4.87*	*0.21*	*6.44*	*9.46*
Arias Consultores*	4.89	0.19	6.27	17.59
LinearRegression	5.03	0.21	6.72	8.69
GRNN PCA	5.12	0.23	7.3	12.06
LinearRegression PCA	5.12	0.19	7.77	13.05
GRNN	5.16	0.23	7.39	12.25
MLP-BP PCA	5.18	0.22	7.93	13.45
MLP-BP	5.71	0.24	8.45	13.59
Suasor Consultores*	6.43	0.27	8.75	15.91
El Financiero*	6.53	0.31	8.44	7.48
Economista y Asociados	6.83	0.35	9.65	18.71
Grupo Impacto	6.91	0.37	8.16	7.91
Conteo*	8.43	0.43	10.53	18.91
Pop Group*	9.32	0.50	11.69	17.91

Source: self-provided.

This result was expected because of the wide variation in the final polls, ranging from 34.0 to 62.5 for the most voted candidate, from 14.0 to 24.0 to the second most voted, and from 13.0 to 29.0 to the third most voted, as previously presented in Table 6.11. Thus, as the model was trained with this data, the results were in line with the polls: the MAE of all predictions were poorer than 3 pollsters and the poll average, and better than 6 pollsters. Surprisingly, the linear regression model obtained the best results amongst models, although it is difficult to identify why, due to the high variation of polls. Lastly, we highlight that data from Instagram of the two most voted candidates were not collected. Thus, if we had obtained access to this data, different results would have been obtained.

By summarizing the results on all countries, we verified that:

- In Argentina, predictions with the MLP-BP PCA model obtained the second best results, better than the poll average and with a lower AEOM;
- In Brazil, predictions obtained with the MLP-BP PCA model obtained the best results;
- In Colombia, predictions with the GRNN PCA model obtained the best results, and predictions with the MLP-BP PCA obtained competitive results;
- In Mexico, predictions with all models obtained competitive results, but results were impaired by high prediction errors of the final polls.

Table 6.27 summarizes the MAE obtained by the models in all countries, allowing a comparison with the poll average of each country. It demonstrates that, on average, the MLP-BP model with PCA obtained the best results. The two GRNN models also obtained better results than the poll average, and the MLP-BP model without PCA obtained competitive results. In all countries, the MLP-BP with PCA obtained better results than without PCA. The GRNN model with PCA obtained better results than the version without PCA in three of the four countries.

Table 6.27 – MAE errors of all countries, compared with the respective poll average

MAE Comparison	Argentina	Brasil	Colombia	Mexico	Average
MLP-BP PCA	1.99	1.43	3.03	5.18	2.91
GRNN PCA	2.11	2.19	2.66	5.12	3.02
GRNN	2.12	2.13	2.70	5.16	3.03
Polls Average	*2.28*	*2.25*	*3.02*	*4.87*	*3.11*
MLP-BP	2.09	1.81	3.43	5.71	3.26
Linear Regression PCA	2.27	3.87	6.75	5.12	4.50
Linear Regression	3.42	5.91	15.45	5.03	7.45

Source: self-provided.

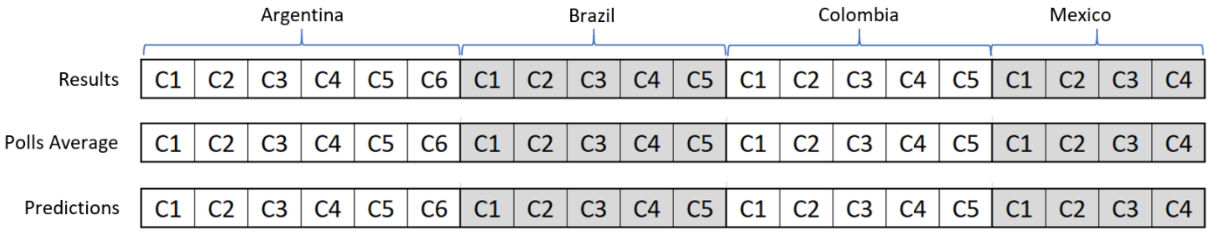
Following the defined methodology, we also performed Wilcoxon signed-rank tests to verify: (i) whether the predicted values were statistically different from the election vote shares; (ii) whether the final poll values were statistically different from the election vote shares; and (iii) whether the errors of the predicted values were statistically different from the errors of the poll values, either higher or lower.

For this, we worked with seven series, one for each prediction model and one additional with the poll averages. Due to the low number of samples (ranging from 4 candidates in Mexico to 6 candidates in Argentina), we concatenated the values of each country in series containing 20 points, related to the predictions of the candidates of all countries: 6 from Argentina, 5 from Brazil, 5 from Colombia and 4 from Mexico. Thus, we have more data so as to perform better statistical tests considering all predictions at once. Figure 6.1 illustrates this modeling.

The first test compared the election results with our predictions, as well as with the poll averages. This data is presented in Table 6.28 and gives the *p* values of all the tests. The *Two Sided* column presenting a *p* value <= *.05* indicates that the results are not equivalent, the *Greater* column with a *p* <= *.05* indicates that the election results are statistically greater than the predictions, i.e., predictions are biased to present lower values; and the column showing *p* <= *.05* indicates the opposite. All calculations

were performed using the *Wilcoxon* function of the *scipy.stats.wilcoxon*[16] python module.

Figure 6.1 – Data modeling for the statistical tests



Source: self-provided.

The results illustrate that the only predictions not equivalent to the election results were the predictions using linear regression, presenting a $p = .00$ on the two-sided test, and on the lower test. Thus, it indicates that linear regression predictions are biased and that the election results presented lower values than the predictions. All other values, including the poll averages, may be considered as not statistically different from the elections results.

Table 6.28 – Wilcoxon signed rank test between the election results, and the predictions and polls

Comparison with Election Results	Two Sided	Greater	Lower
Election Results x Linear Regression	**0.00**	1.00	**0.00**
Election Results x Linear Regression PCA	0.19	0.91	0.09
Election Results x MLP-BP	0.67	0.34	0.68
Election Results x MLP-BP PCA	0.87	0.43	0.58
Election Results x GRNN	0.81	0.61	0.41
Election Results x GRNN PCA	0.87	0.58	0.43
Election Results x Polls Average	*0.65*	*0.69*	*0.32*

Source: self-provided.

By comparing the absolute errors of our predictions and the poll averages, we found the data presented in Table 6.29. This demonstrates that the poll averages presented lower errors than both the linear regression approaches, although all other errors are similar. We highlight that besides the linear regression, the most significant result is $p = 0.14$, indicating that the errors of the poll averages may be higher than the errors obtained with the MLP-BP PCA approach, although the p-value is not strong enough to confirm this.

[16] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

Table 6.29 – Wilcoxon signed rank test between the errors obtained with the prediction and polls

Comparison	Two Sided	Greater	Lower
Polls Average x Linear Regression	0.02	0.99	0.01
Polls Average x Linear Regression PCA	0.09	0.95	0.05
Polls Average x MLP-BP	0.75	0.62	0.38
Polls Average x MLP-BP PCA	0.28	0.14	0.86
Polls Average x GRNN	0.90	0.45	0.55
Polls Average x GRNN PCA	0.99	0.51	0.49

Source: self-provided.

The presented data illustrates that, considering the traditional and most used error metric in this scenario, the MAE, the results obtained with the two proposed models, the MLP-BP and the GRNN presented competitive, or even better, results with the poll results in all countries. Also, the statistical analysis demonstrates that, statistically, the predicted results are competitive with the poll results. The presented data therefore **rejects** the null hypothesis H_2', which validates the alternative hypothesis "H_2: It is possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of predicting election results with competitive results to traditional polls", and answers affirmatively the "RQ2: Is it possible to define a process and create an ML model capable of predicting election results based on the SM performance of candidates?".

In addition to rejecting the null hypothesis, additional conclusions may also be drawn. The countries that presented higher correlations between the SM performance and the vote share, most notably Brazil and Argentina, were those that presented the best prediction results, with a MAE lower than 2.0. As the data of higher correlation with the final vote share is only available after the elections, we used all the features to perform the predictions, avoiding the insertion of bias. However, one promising strategy may be to test the SM performance with the poll performances in order to choose the most suitable features.

Moreover, although the model presented a good performance even with the existence of a high variation in data on the SM features, its results depend on accurate poll data for training, as expected. Thus, training the model with accurate polls, as in Brazil and Argentina, led to even more accurate predictions than the best polls, while

training the models with inaccurate polls, such as in the Mexican scenario, led to high prediction errors. Thus, advanced approaches for pollster and poll pruning, such as identifying and removing outlier polls, would lead to better results. Lastly, this approach may be very useful for internal use by parties in countries that do not allow polls to be released during the last week before elections, thereby enabling candidates and parties to have an estimate of their performance by only using the SM data and previous polls.

6.3.3 Research Question 3

After running all phases of the process, we attempted to refute the "H_3': It is not possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of making daily predictions of election results with competitive results to traditional polls." For this, according to the methodology, we will perform two analyses: a descriptive and qualitative analysis of the two most voted candidates with regard to polls, predictions and the final vote share; and the measurement of prediction errors by considering the polls as imprecise ground truth. Thus, competitive predictions will be considered as those that, compared with the polls, present errors below the historical MAE of 2.7 percentual points, within a deviation of 3.00, which is considered the error margin of most polls.

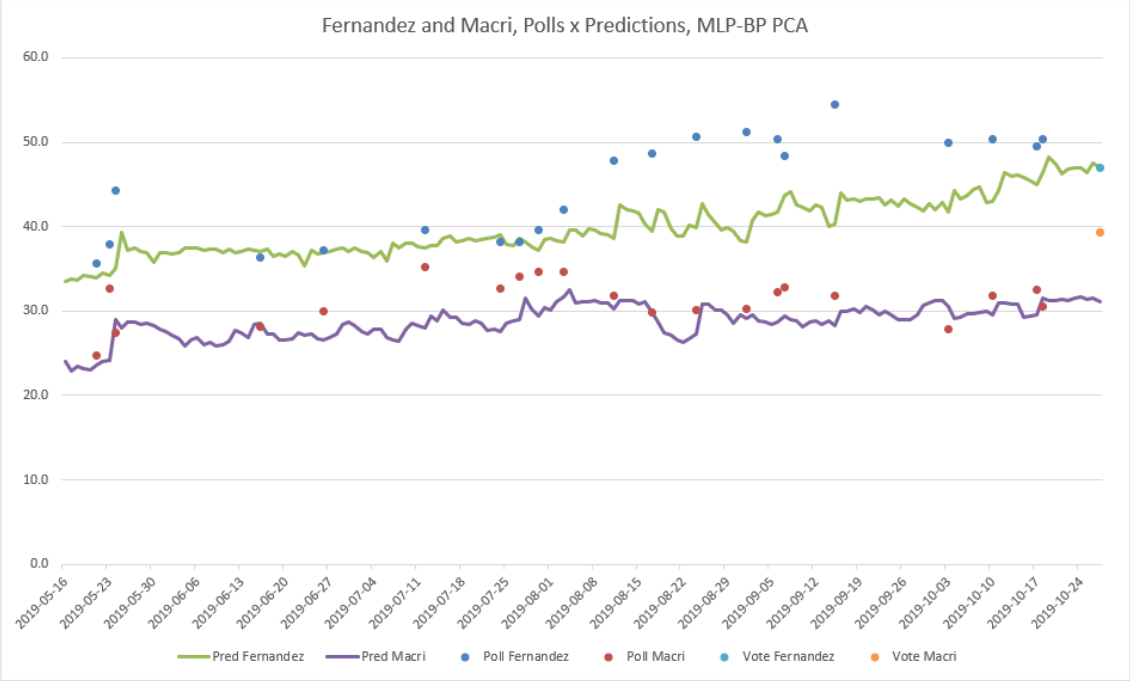
To illustrate the predictions, Figures 6.2-6.5 presents the daily predictions, obtained by the MLP-BP PCA model, of the two most voted candidates in Argentina, Brazil, Colombia and Mexico respectively, as two lines. They also present the polled data as dots, and the final election result as the last dots.

The Argentinian data, presented in Figure 6.2, demonstrates that for the most voted candidate, Fernandez, after the start of the campaign, the polls overestimated his vote share, and our predictions underestimated while achieving the exact result on the final prediction day, as previously presented in Table 6.19. Moreover, both polls and predictions similarly underestimated the second most voted candidate.

In the Brazilian scenario, presented in Figure 6.3, it may be observed that before the last month of the campaign, predictions varied between higher and lower values than the polls, suggesting unbiased values. However, in the last month, almost all the polls for both candidates, and especially for Haddad, were higher than our predictions,

and higher than their final vote share. However, our final predictions were extremely close to the final vote share, as presented in Table 6.21. This data, although not conclusive, and difficult to be proved or statistically tested, reveals that polls may present a bias towards both candidates and overestimate their vote share, which did not occur with our predictions.

Figure 6.2 – Polls and predictions in Argentina: Fernandez and Macri, predictions using the MLP-BP PCA model and the final vote share



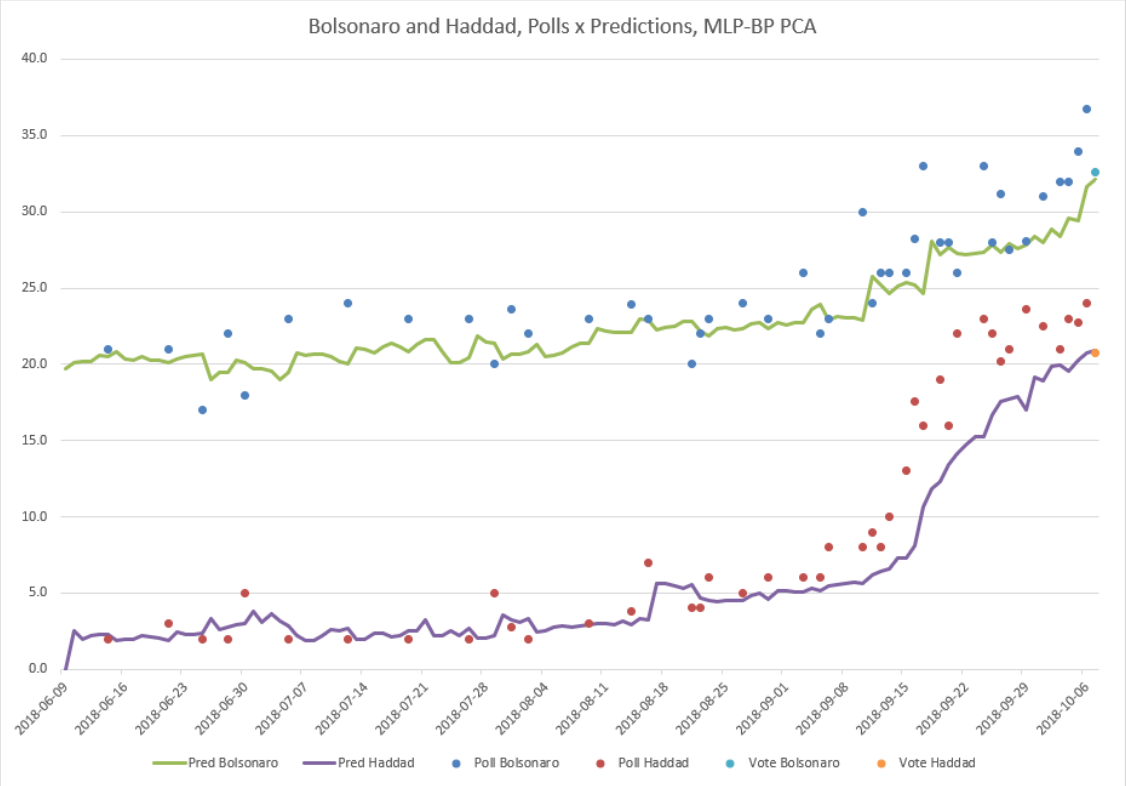
Source: self-provided.

In the Colombian scenario, shown in Figure 6.4, both polls and predictions presented close results nearer to election day, especially for the second most voted candidate. However, many strong outliers were found on polls for both candidates, some of them have been circled in the figure. Furthermore, some spikes were found on predictions for Márques on specific days. These spikes may be explained by the instability of polls on days close to the spikes, but this effect needs to be investigated further.

Figure 6.5 presents the polls and predictions for the Mexican scenario. The polls and predictions for the second most voted candidate, Cortés, remained close during the whole period. Exceptions were some poll outliers on June 4 and 8. Otherwise, the high variation in the poll results close to the elections, from 34.0 to 62.5 percentual points during the last week, as presented in Table 6.11, made it difficult to undertake accurate predictions for Obrador. Thus, although the predictions had obtained results

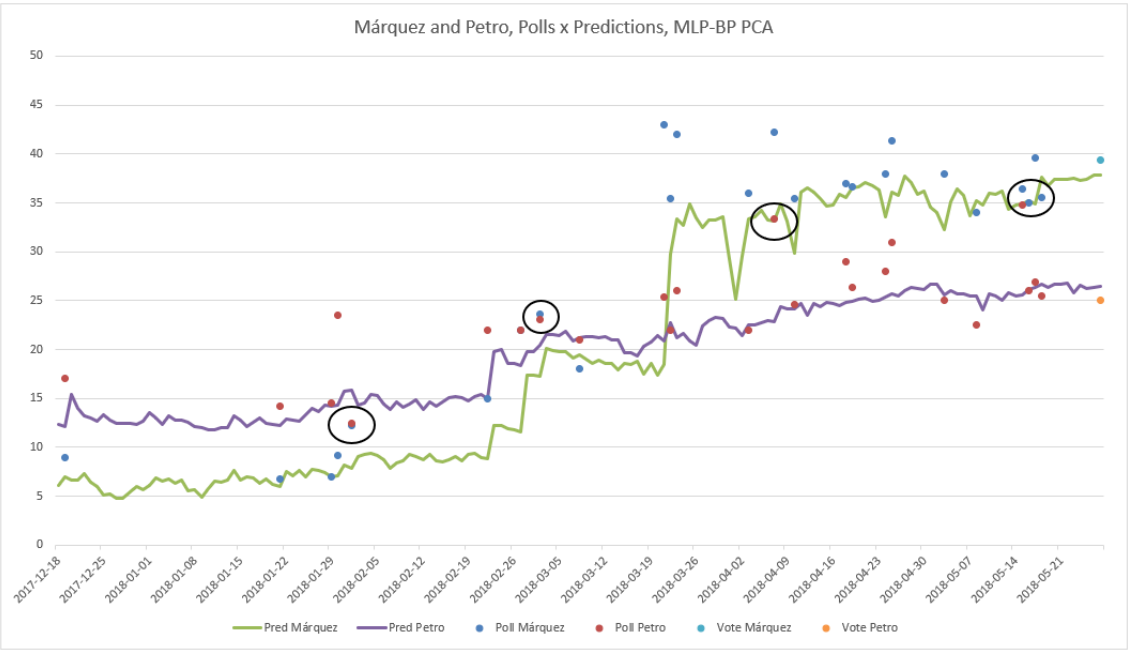
that were competitive with the polls, this result reinforces the need for a better approach for poll pruning.

Figure 6.3 – Polls and predictions in Brazil: Bolsonaro and Haddad, predictions using the MLP-BP PCA model and the final vote share



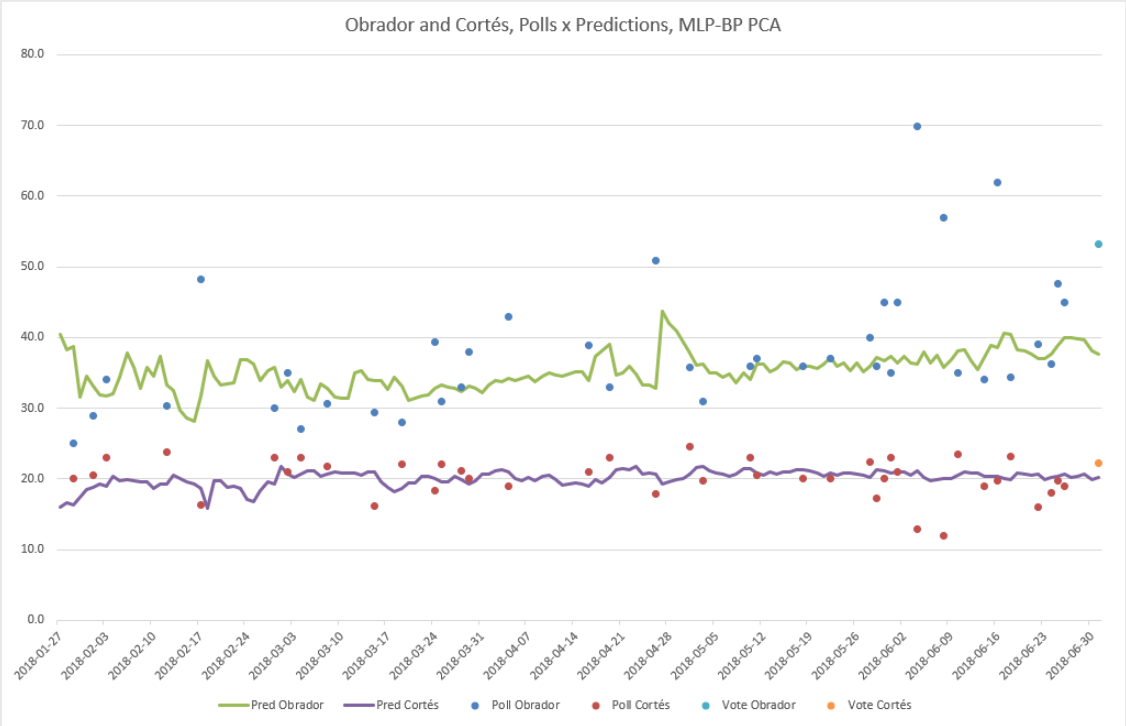
Source: self-provided.

Figure 6.4 – Polls and predictions in Colombia: Márquez and Petro, predictions using the MLP-BP PCA model and the final vote share



Source: self-provided.

Figure 6.5 – Polls and predictions in Mexico: Obrador and Cortés, predictions using the MLP-BP PCA model and the final vote share



Source: self-provided.

By summarizing this analysis on all countries, we have verified that:

- In Argentina: for the first candidate, results suggests that the prediction bias was lower than the poll bias, since the predictions in the last week were closer than the polls. For the second most voted candidates, predictions and polls seemed equally biased;
- In Brazil, results for both candidates suggest a higher poll bias than prediction bias, since the prediction curve smoothly achieved very accurate final predictions;
- In Colombia, polls and predictions presented similar results, despite some high poll outliers and some prediction spikes, which should be investigated further;
- In Mexico, the polls and predictions were close with regard to the second most voted candidate. However, the high variance in the polls, particularly close to elections, and the high errors obtained both by polls and prediction, prevents any conclusions regarding the results.

The MAE comparing all predictions and polls during the campaign is presented in Table 6.30. The calculation considers polls as the baseline, and the error is the difference between the predicted results (on the days with polls) and polls.

Table 6.30 – MAE comparing predictions and polls. This shows how close the predictions are to the polls

Model	Argentina	Brazil	Colombia	Mexico	Average
LinearRegression	151.93	13.30	46.11	8.70	55.01
LinearRegression PCA	2.60	1.44	3.26	4.01	2.83
MLP-BP	2.21	1.60	2.96	3.56	2.58
MLP-BP PCA	2.29	1.54	2.88	3.59	2.58
GRNN	2.35	1.33	3.00	3.62	2.58
GRNN PCA	2.34	1.35	2.96	3.61	2.57

Source: self-provided.

As expected, the linear regression model presented the poorest results, even though the linear regression with PCA presented unexpectedly good results. Moreover, results with neural networks in Argentina, Brazil and Colombia presented a MAE lower than or equal to the usual poll error margin of 3 percentage points, and most values were lower than the historical threshold of 2.7 percentage points, with remarkable low values in Brazil. This data reveals that, on average, these intermediary predictions may be seen as correct predictions of polls.

The presented data therefore **rejects** null hypothesis H_3', which validates the alternative hypothesis "H_3: It is possible to define a process and create a model based on the SM performance of candidates, using an ML approach and trained with traditional polls, which is capable of making daily predictions of election results with competitive results to traditional polls," and answers affirmatively the "RQ3: Is it possible to define a process and create an ML model capable of performing daily nowcasting of election results based on the SM performance of candidates?".

6.4 RESULTS DISCUSSION AND COMPARISON

6.4.1 Research Area Challenges Addressed

This thesis has found correlations between SM metrics and election results. This fact is not a novelty in itself, since it has already been stated by many previous studies

claiming success in predicting elections with SM, ever since the seminal paper by Tumasjan stating that “the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share”. The main difference concerns the definition of the SM performance. Most studies have measured performance based on how many people are talking about a candidate (generally on Twitter), but our approach has considered how many people are paying attention to a candidate by interacting with his/her profiles on SM. This change of perspective addresses many of the sampling challenges observed in previous studies.

This thesis is not the first claiming to be able to predict elections results. However, the defined process and the presented modeling has addressed many of the challenges found in previous studies, already presented in Table 3.3. The manner in which this proposal has addressed each challenge is discussed below.

The **process challenges** were all addressed by the definition of the SoMEN and the SoMEN-DC processes. The processes were run in four different countries and performed well in different contexts. The SoMEN-DC was defined exactly to enable predictions during a campaign.

The **sampling challenges** were also addressed. The challenge of using social networks as a population sample is addressed by using traditional polls to train the models. The representativity problem of Twitter, as being the sample of all platforms, was addressed by the capability of the proposal to use data from different SM platforms, those that the candidates and the population use most, and is also naturally prepared for use even with platforms that are not yet popular.

Moreover, due to a change in the concept of SM performance, from measuring how many people are talking about a candidate, to measuring how many people are paying attention to a candidate, the defined approach requires the collection of much less data, a thousand posts from less than a dozen candidates, instead of millions of posts from the entire population. Thus, all required data can be collected, rather than just a sample. Furthermore, very few choices have to be made by researchers regarding data collection. Since data is collected from the official profiles of candidates, no keyword for data collection needs to be defined. We also proposed that data collection would occur for a long period before elections (10 months), and the use of this data combined with many different window sizes exempts the need of arbitrary decisions regarding how many days of SM data should be collected and used. Lastly, the small amount of data to be collected and processed thereby requires a very low

computational power in order to perform the predictions. Indeed, all training and predictions were performed on a standard laptop computer.

In the case of the **modeling challenges,** the high susceptibility to volume manipulation was addressed by training and predicting the candidate results individually. Thus, the model was trained with the specific behaviour of the candidate's supporters, and even the behaviour of his/her network of BOTs or paid propaganda, should this exist.

Also, to the best of our knowledge, as presented in the systematic review, this is the first approach that is capable of crossing data from multiple platforms, particularly Facebook, Instagram, and Twitter.

We also considered the state-of-the-art machine learning and technical modeling. The chosen models, the MLP-BP and the GRNN, although they are not the most recent proposals, are well suited for this context and are capable of presenting nonlinear relationships between inputs and outputs, unlike other approaches based on linear relationships. In particular, the averaged MLP-BP presented remarkable results on small samples in a very recent comparative study (FERNÁNDEZ-DELGADO *et al.*, 2019).

In relation to one of the most relevant technical modeling weaknesses, the choice of parameters for the ML models, this thesis has clearly presented the strategy for parameter choice. The GRNN only needs one parameter, which was found by a grid search strategy. In addition, the MLP-BP parameters were not only chosen based on the characteristics of the sample, but also based on a preliminary study, which compared the chosen parameters with a grid search strategy for the choice of the best parameters and achieved similar results with both strategies.

Lastly, the challenges of **performance evaluation and scientific rigor** were also addressed. Statistical analyses of the results were proposed and performed, and have been presented in this chapter. Thus, the following sections compare the results with other related works, and discusses the bias and threats to validity.

6.4.2 Comparison with Related Works

In addition to comparing our approach with the aggregated results of the systematic review, we also performed an analysis and comparison with related studies. No studies were found predicting elections for the same countries during the same

years for direct comparison. Thus, we compared our approach and results with studies containing similar characteristics: predicting elections in the studied countries, using polls for training, and/or predicting elections in more than one country.

Cerón-Guzmán and León-Guzmán (CERON-GUZMAN; LEON-GUZMAN, 2016) studied the 2014 Colombian presidential election using a slightly similar approach to that used in this thesis: the use of polls and SM data for training several regression models. They mostly used linear models for regression (Ordinary Least Square, Ridge, Lasso, and Support Vector Regression), and their input data were based on the volume/sentiment approach, enhanced by spammer detection and an improved sentiment analysis of Spanish tweets. However, their results wrongly presented the list of the most voted candidates, and they concluded that “the obtained results show that inference methods based on Twitter data are not consistent…”.

In terms of Brazilian elections, (JUSTINO GARCIA PRACIANO *et al.*, 2019) developed an approach also based on volume/sentiment to detect the winner of the second round of the 2014 presidential elections. The authors considered the approach as being successful by correctly indicating the winners in 19 out of 26 states. However, since they did not indicate the vote share, and only the winner in each state, the approach would be better suited to scenarios such as in the U.S.

With regard to studies analyzing multiple elections, Gaurav et al. (GAURAV *et al.*, 2013) used a Twitter based volume approach to predict elections in Ecuador, Paraguay and Venezuela in 2013. They found that “counting the tweets that mention a candidate’s conventional name is not sufficient to obtain good predictions,” but after some enhancements, such as ignoring multiple tweets from a single user, they achieved better results, with MAEs of approximately 3.0. In addition to all the previously discussed drawbacks on the volume counting approach, in this study they collected 1.2 billion tweets, which is impracticable on the current state of platforms.

Anjaria et al. (ANJARIA; GUDDETI, 2014) employed four ML techniques (SVM, Naïve Bayes, maximum entropy and MLP-BP) for a Twitter sentiment analysis to predict the 2012 U.S. presidential elections and the 2013 Karnataka (India) state assembly elections. It should be noted that the parameter selection was “achieved by trial-and-error method”, without any details. Their best result in the U.S. presented a MAE of 3.44, which is an acceptable result, better than our results in Mexico and worse than the others. Nevertheless, they obtained a MAE of 13.60 in the Indian elections considering four parties, which largely differs from the actual vote share.

No studies were identified using polls for training, not mainly based on Twitter volume/sentiment, and applied in more than one country. However, one study similar to ours, although only applied to the 2016 U.S. presidential election, is described in (ISOTALO *et al.*, 2016). In this study, 13 different variables available online were used, including polls, Facebook page likes, Google trends, Wikipedia page traffic and betting sites, amongst others. A linear regression was then performed for training and prediction. While the study observed correlations between Facebook page likes and betting with polls, there were no correlations with other variables. Despite claiming good results and presenting new ideas for variables, the lack of detail regarding the used variables and the presence of arbitrary decisions, such as the use of days 1-3 of the independent variables for explaining polls on day 5, prevents its replication.

The study we consider the closest to ours was presented by Tsakalidis et al. (TSAKALIDIS *et al.*, 2015). They used Twitter data to predict the 2014 EU election results in Germany, the Netherlands, and Greece. Despite using the traditional Twitter volume and sentiment model, they used 11 volume/sentiment derived variables combined with one poll-based feature for training regressors. Three algorithms were applied for regression: linear regression, Gaussian process, and sequential minimal optimization, and the output average of the three was used as the final estimate in this combined regressor. They used 26 polls from Greece, nine from Germany, and 13 from the Netherlands. In terms of results, their approach achieved a good performance, and obtained a MAE below 2.0 in the three countries. Their good results are in line with our arguments that generating domain-based derived variables and training ML algorithms with these data combined with polls is able to achieve good results. They also performed the Wilcoxon (two-tailed) test, but it was only used to test the differences between the variations of their approach, and not to test the prediction results nor compare their results with polls.

Despite the good results, Tsakalidis's study presents certain drawbacks, most of them well known in volume/sentiment approaches. The most prominent was the selection of keywords for searching on Twitter. They reported the inclusion of many keywords, including some possible misspellings, but without listing them. They also excluded several ambiguous keywords to reduce noise "for example, the abbreviation of the Dutch party 'GL' could stand for 'good luck'", again without listing them. Moreover, the study used an arbitrary 7-day window, despite having tested other values. They also recognized that due to the restrictions of the Twitter Streaming API

at that time, no more than 1 percent of the public tweets had been gathered. Even with this limitation, their approach collected on average 350,175 posts (361,713, 452,348, and 263,465 tweets from Germany, the Netherlands, and Greece, respectively). Notwithstanding, our approach does not present these problems of arbitrary choices, is able to collect all the required data, and needs to gather only a fraction of the posts (16,851 posts by country, 4.8% of data).

6.5 LIMITATIONS AND VALIDITY DISCUSSION

Despite the rigor with which this study was conducted, it is possible that it may have been affected by the threats of validity. Next, we discuss the internal, external, construct, and conclusion validities.

Internal validity considers whether the experimental design is able to support conclusions on causality or correlations. In this study, although the theory that inspired the definition of the set of SM performance metrics suggests a causality relation between the exposure and enhancing of attitudes regarding an individual, the objective was to find whether correlations—not causality—existed. The correlations found between the SM metrics and votes does not necessarily signify that SM impacts voting. Offline events, the behavior of candidates in debates, the effectiveness of their propaganda, and many other facts may equally impact both electoral results and the SM performance, leading to such a correlation. In this sense, we argue that measuring the SM performance may be a quick, easy manner with which to measure public opinion, complementing traditional polling methods but not replacing them.

The external validity of the study measures its capability of being affected by generalization, i.e., the capability of repeating the same study in other research groups. In this sense, it is one of the few studies that applies exactly the same process, models and choices in different contexts, and is the only one that we have found that was applied in the elections of four different countries. However, the context of the countries is similar, since the selected countries are the most populous countries in Latin America. Thus, there are no guarantees that the approach proposed in this study may be used in different contexts, such as in Europe or Asia.

However, we have evidence that the approach may be used in other contexts, both considering the country as well as the SM scenario changes. It was designed not to be dependent on a specific SM platform, so that it may capture the most commonly used

platform in the country and in the year of a given election, even if this platform does not yet exist, but is based on newsfeed. As an example, a preliminary version of this approach was applied equally to the 2016 U.S. and the 2018 Brazilian elections (BRITO, K. dos S.; ADEODATO, 2020), and presented similar results. It was also repeated in the 2020 U.S. elections and the results were published informally[17] on a SM platform (LinkedIn) on the morning of election day, and also presented accurate results.

Construct validity considers whether the models and metrics used in a study are a valid abstraction of the real world under study. In this sense, the study was performed using real-world data from the selected elections. A complete information system was developed for this task, as well as the publicly available polls.

In terms of metrics, for correlation, the well-known Pearson coefficient correlation was used. For prediction evaluation, the most used metrics used in the polling domain were used. Lastly, for a statistical assessment of the results, the Wilcoxon signed-rank test was used. To reduce the bias, which may be introduced by using a small sample size for the Pearson correlations, as well as the choice of an arbitrary value to determine acceptable values for correlations, we used the correlation coefficient's "rules of thumb", statistically justified correlation coefficient. In addition, because the small sample size was a difficulty in the Wilcoxon statistical test of predictions, we ran the statistical test with data from all four countries aggregated, using a longer series of 20 samples instead of a smaller series of 4 to 6 samples.

Conclusion validity is concerned with the relationship between the treatment and the outcome, and determines the capability of the study to generate conclusions. Considering the ML model, we used three different models: MLP-BP, GRNN, and linear regression. The first two were chosen because they are well suited for this context. The third was used as a baseline model.

Moreover, some different decisions could be taken. All studies were based on a combination of the three SM platforms: Facebook, Instagram and Twitter. It could be argued that the use of only one or two of the most appropriate platforms would lead to better results. We prefer not to use these combinations because in fact we may only know the most correlated platform with the final vote share after the election. It would

[17] Available at https://www.linkedin.com/pulse/an%25C3%25A1lise-prevendo-o-resultado-das-elei%25C3%25A7%25C3%25B5es-nos-estados-kellyton-brito. Viewed on: February 08, 2021.

also limit the generalization capabilities of our proposal. Moreover, in the preliminary results of this thesis regarding elections in Brazil and the U.S. (BRITO, K. dos S.; ADEODATO, 2020), it was not possible to make conclusive assumptions regarding the best number and combination of platforms used on the input model. Indeed, results show that the most suitable platform is dependent on the country being analyzed. The same occurred with selecting the size of the window for data gathering. Because it is hard to decide the window size beforehand, without adding a bias or arbitrary decision, we decided to use several different window sizes in an ensemble of networks.

6.6 CONCLUDING REMARKS

In this Chapter we have presented practical experiments that rejected the null hypotheses H_1', H_2' and H_3' in favor of the alternative hypotheses H_1, H_2, and H_3, which enabled the research questions of this thesis, RQ1, RQ2, and RQ3, to be affirmatively answered. Thus, we conclude that:

H1: It is possible to model the SM performance based on the interactions of users on the official profiles of candidates and find correlations between the SM and the electoral performances of candidates;

H2: It is possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of predicting election results with competitive results to traditional polls; and

H3: It is possible to define a process and create a model based on the SM performance of candidates, using an ML approach trained with traditional polls, which is capable of making daily predictions of election results with competitive results to traditional polls.

For this, we instantiated the defined processes SoMEN and SoMEN-DC in the context of the most recent presidential elections of the four most populous countries, with the highest GDPs in Latin America, namely: Argentina, Brazil, Colombia, and Mexico.

Strong correlations were observed between the defined SM performance metrics and the electoral results. Also, the errors in predicting the final results of the elections were lower than or equivalent to the errors obtained by traditional polls. Finally, predictions made during the campaign period were close to the poll predictions, within the error margins of the polls.

Furthermore, it was presented how the proposals address the challenges of this research subject, and a direct comparison with related works was also performed. Lastly, we discussed study validity.

The next chapter presents this thesis' concluding remarks.

7 CONCLUSION

“There are things we know we know. We also know there are known unknowns.”
(Donald Rumsfeld)

This thesis has investigated the correlations between SM performance and the electoral results of presidential elections, as well as the feasibility of using SM data and ML models for predicting elections. For this, a new set of metrics was defined, based on the engagement of citizens with the official profiles of candidates. A process was also defined to perform election predictions, and a machine learning model was created composed of an ensemble of artificial neural networks. Experiments were performed with data from the most recent presidential elections in four of the most populous Latin American countries, Argentina, Brazil, Colombia, and Mexico.

The results of the experiments show that (i) there is a correlation between the SM performance of candidates and their electoral performance; (ii) it is possible to define a process and create a model capable of predicting election results based on the SM performance of candidates; and (iii) it is possible to define a process and create a model capable of performing daily nowcasting of election results based on the SM performance of candidates.

In addition to answering the research questions, this thesis was also strongly based on the results of a systematic review of this new area of research. Thus, the proposed process and model attempted to deal with the main challenges recognized in previous studies, including the challenges presented by process, sampling, modeling, and performance evaluation and scientific rigor.

We believe that this thesis may change the direction of future studies in this area. We have identified that the most commonly used approach based on volume/sentiment on Twitter posts might not be the most efficient, and have presented a new way of measuring performance on SM. Instead of considering how many people are talking about a candidate, the new approach considers how many people are paying attention to a candidate. We have also presented a new process which is capable of being repeated in different elections with minor adjustments. Moreover, the framework (process and model) is able to make predictions with data from different SM platforms,

even those that have not as yet become popular, using nonlinear ML models. Lastly, we went in pursuit of “suitable” models, and not the best. Thus, we believe that there is still a long road ahead in order to improve the election prediction scenario.

We highlight that we made public two predictions before the final release of results: the prediction of the 2019 Argentinian election was published on Facebook[18] on election day and attained better results than the considered polls at that time, and the prediction of the 2020 U.S. elections was published in LinkedIn[19] on the day of elections, obtaining an MAE error 0.1 point lower than the RCP[20] poll average. These results reinforce the claims that the proposed approach may be used in predicting elections in real-time, rather than only working with data from the past.

Next, we summarize the contributions, list the publications arising from this work, and finish discussing possible future research topics to evolve.

7.1 CONTRIBUTIONS

The main contributions of this thesis are summarized as follows:

- **First:** This thesis proposed a new approach for modelling SM performance, changing the focus from how many people are talking about a candidate to a new approach considering how many people are paying attention to a candidate. Thus, we defined a new set of metrics that may be used to measure performance on SM to be used as features in prediction. The correlation of this new set of metrics was tested on four electoral contexts, the most recent presidential elections in four Latin American countries, and high correlations were found. Thus, the hypothesis that there is a correlation between SM engagement and electoral results has been validated. It is also important to highlight that this new set of metrics is malleable, not dependent on specific SM platforms, and may be easily adjusted for different contexts where different SM platforms are most used.

[18] Available at https://www.facebook.com/notes/404228243926826/. Viewed on: January 13, 2021.

[19] Available at https://www.linkedin.com/pulse/an%25C3%25A1lise-prevendo-o-resultado-das-elei%25C3%25A7%25C3%25B5es-nos-estados-kellyton-brito. Viewed on: February 08, 2021.

[20] https://www.realclearpolitics.com/epolls/2020/president/us/general_election_trump_vs_biden-6247.html. Viewed on: January 13, 2021.

- **Second:** A framework (SoMEN) was defined, composed of a process and an ML model capable of nowcasting electoral results, as well as an adaptation to nowcasting intermediary results during the campaign (SoMEN-DC). The processes defined the steps and the main decisions needed to perform predictions. The model defined an ANN architecture, composed of an ensemble of ANN, well suited to this kind of prediction. The SoMEN and SoMEN-DC achieved results that were competitive with traditional polls. Thus, the framework may be used both by the academia and the industry to perform electoral predictions, both the final vote share as well as nowcasting during the campaigns. It may also be very helpful for election campaigns so as to measure the impact of the campaign and the voting intention on a daily basis, which was not yet possible in Latin America due to the low numbers of polls performed in this region.
- **Third:** To the best of our knowledge, this thesis has performed the most extensive, complete systematic review of the area of predicting elections based on SM data. Thus, new conclusions have been achieved, such as the low success rate achieved by the main approach based on Twitter volume/sentiment, and new possible approaches have been identified. We also identified the main challenges and indicated future directions in the areas of process definitions, model definitions, sampling, and evaluation.
- **Fourth:** To the best of our knowledge, this is the second attempt at using the same approach for predicting multiple elections on Latin America, following on from the work of (GAURAV *et al.*, 2013), but it is the first attempt at publishing results on the day of elections and before the official results, as we undertook with the Argentinian preliminary results. It is also the first attempt at using data from Instagram to predict the vote share of presidential elections.

This is a multidisciplinary thesis and involves knowledge from the areas of polling and electoral predictions, social media studies, and machine learning. Thus, we emphasize the specific contributions for these areas.

For the **machine learning** area, this thesis studied and applied ML for a new, as yet underexplored context, the prediction of electoral results. This context presents very specific challenges, such as the lack of historical data, the rapid change of context, a small number of samples and the existence of only one actual labeled data, the final result that is the aim of prediction. Also, this context presents external interferences,

such as the existence of bots or even the natural difference of the behavior of supporters on SM. Thus, we consider the proposals, including the new set of features, the training of each candidate individually, and the ensemble of many ANN for different windows, allowing accurate predictions with a smaller number of collected data points, contributions in the area of ML.

For **social media** studies, this thesis has directly studied the correlation of SM performance and electoral performance, finding a high correlation. This result is, in itself, a relevant contribution for the area of social media studies, by finding correlations between online behavior and offline outcomes. This kind of result is difficult to be achieved by social scientists because they frequently face difficulties in accessing and processing the high volume of data needed for these calculations. Thus, by knowing this correlation, social researchers may derivate other studies, for example, to study if there is causality on the found correlations.

For the subject of **electoral predictions**, the results of this thesis contribute by proposing an approach able to complement traditional polling. It may be used as in this thesis, by performing nowcasting on days with no polling, or may be included in the methodology of traditional pollsters, by updating their methodology to use both data from interviews and from SM.

7.2 PUBLICATIONS ARISING FROM THIS WORK

This section presents the publications that have arisen from this work. The research began long before the studies for this thesis was formalized, through the development of information systems needed for data collection on SM, based on the concepts of the web of social machines. These papers are listed in Chapter 6, Section 6.1.2.1: Data Collection and Understanding section. Nevertheless, only the publications over the past four years, the formal period of this PhD study, have been included in this section.

Articles Published in Journals

- Brito, K. dos S., de Lima, A. A., Ferreira, S. E., de Arruda Buregio, V., Garcia, V. C., Meira, S. R. L. (2020). Evolution of the Web of Social Machines: A Systematic Review and Research Challenges. **IEEE Transactions on**

Computational Social Systems, 7(2), 373–388. https://doi.org/10.1109/TCSS.2019.2961269

Articles Accepted for Publication in Journals

- Brito, K. dos S., Silva Filho, R., 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**
- Brito, K. dos S., Meira, S. R. de L., Adeodato, P. J. L. (2021). Correlations of Social Media Performance and Electoral Results in Brazilian Presidential Elections. **Information Polity**.

Full Papers Published in Conference Proceedings

- Brito, K. dos S., Paula, N., Fernandes, M., & Meira, S. (2019). Social Media and Presidential Campaigns – Preliminary Results of the 2018 Brazilian Presidential Election. **In 20th Annual International Conference on Digital Government Research on - dg.o 2019** (pp. 332–341). New York, New York, USA: ACM Press. https://doi.org/10.1145/3325112.3325252
- Brito, K. dos S., & Adeodato, P. J. L. (2020). Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data. **In 2020 International Joint Conference on Neural Networks (IJCNN)**. Glasgow, United Kingdom. Retrieved from https://ieeexplore.ieee.org/document/9207147

7.3 FUTURE RESEARCH

We believe that this thesis may change the direction of future work in this area. Thus, we consider that many future works may be performed, such as:

- Applying the proposed framework to other electoral contexts, such as presidential elections in Europe and Asia, as well as in other types of elections, e.g., parliamentary elections. Thus, it may be verified whether the approach is well suited to other kinds of elections and to other regions. Also, as our experiments were based on presidential elections of populous

countries, studies regarding the use of the framework for small elections, such as those for state governors or city councils, and to discover the minimal amount of data needed for predictions should also be investigated. Lastly, the correlations between the accuracy of results and contextual data from the place of elections, such as human development index or internet user penetration, may also be investigated.

- Adjusting the process with different options. We believe that neither the process nor each of its phases, from election understanding through to evaluation, are final. Thus, they may be improved or tuned. We would make particular mention of the fact that the election understanding is focused on presidential elections that follow the model applied in Latin America, which will be different if applied to the U.S. Also, as discussed in Chapter 4, the evaluation remains a challenge, even for the polling research. Thus, more research regarding the evaluation of continuous polling is needed and must be tracked by researchers aiming to improve this research field.
- Adjusting the model with different options. For example, instead of using PCA to reduce the dimensionality, different approaches to dimensionality reduction or feature selection may be used. Also, one more effective way of pruning polls, excluding outliers, must be defined and will probably improve the results. Moreover, instead of using an ensemble of machines using the datasets with many windows, different approaches may be tested. For example, each dataset may be tested individually with the polls, and the one with the best results could be used for future predictions.
- Testing different ML learning strategies to improve the accuracy of predictions, as discussed on Chapter 5. Approaches based on online learning (LOSING; HAMMER; WERSING, 2018) avoids multiple training using all the datasets, thereby optimizing the process of daily predictions. Recurrent neural networks (MEDSKER; JAIN, 2001) are designed to learn sequential or time-varying patterns and may present good results in this context. Models based on the state space of latent variables, such as Kalman Filter (NÓBREGA; OLIVEIRA, 2019), may also be considered. Few-shot learning (WANG *et al.*, 2020) approaches may be further investigated to tackle the problem of small datasets. Approaches of AutoML (HE; ZHAO; CHU, 2021), focused on automatically identify the best model

and hyperparameters may be also further investigated. Lastly, the use of explainable AI (GOEBEL *et al.*, 2018), “white-box” approaches capable of explaining how the results were achieved, may lead to easy adoption of the predictions in practice.

- As elections are dynamic, there is a possibility that changes may occur in the correlation patterns regarding the behavior of citizens on SM platforms and voting intentions, or even active attempts by politicians to manipulate and inflate their data. The framework presented in this thesis is able to capture and adapt to behavior changes or attempts at data manipulation in different ways, especially if they occur in a consistent manner. However, to allow the framework to better capture and adapt to sudden, unforeseeable changes of SM data patterns, the study and addition of concept drift approaches (LU *et al.*, 2018) within the framework may also be promising.

REFERENCES

ABRAMOWITZ, Alan I. When Good Forecasts Go Bad: The Time-for-Change Model and the 2004 Presidential Election. **Political Science and Politics**, [*s. l.*], vol. 37, no. 04, p. 745–746, 2004. Available at: https://doi.org/10.1017/S1049096504045056

ALNOUKARI, Mouhib; EL SHEIKH, Asim. Knowledge Discovery Process Models: From Traditional to Agile Modeling. *In*: BUSINESS INTELLIGENCE AND AGILE METHODOLOGIES FOR KNOWLEDGE-BASED ORGANIZATIONS: CROSS-DISCIPLINARY APPLICATIONS. [*S. l.*]: IGI Global, Hershey, PA, USA, 2012. p. 72–100. Available at: https://doi.org/10.4018/978-1-61350-050-7.ch004

ANDY JANUAR WICAKSONO; SUYOTO; PRANOWO. A proposed method for predicting US presidential election by analyzing sentiment in social media. *In*: , 2016. **2016 2nd International Conference on Science in Information Technology (ICSITech)**. [*S. l.*]: IEEE, 2016. p. 276–280. Available at: https://doi.org/10.1109/ICSITech.2016.7852647

ANJARIA, Malhar; GUDDETI, Ram Mohana Reddy. A novel sentiment analysis of social networks using supervised learning. **Social Network Analysis and Mining**, [*s. l.*], vol. 4, no. 1, p. 181, 2014. Available at: https://doi.org/10.1007/s13278-014-0181-9

ANTANASIJEVIĆ, Davor *et al.* Multiple-input–multiple-output general regression neural networks model for the simultaneous estimation of traffic-related air pollutant emissions. **Atmospheric Pollution Research**, [*s. l.*], vol. 9, no. 2, p. 388–397, 2018. Available at: https://doi.org/10.1016/j.apr.2017.10.011

ARDABILI, Sina; MOSAVI, Amir; VÁRKONYI-KÓCZY, Annamária R. Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods. *In*: [*S. l.: s. n.*], 2020. p. 215–227. Available at: https://doi.org/10.1007/978-3-030-36841-8_21

ARZHEIMER, Kai; EVANS, Jocelyn. A New Multinomial Accuracy Measure for Polling Bias. **Political Analysis**, [*s. l.*], vol. 22, no. 1, p. 31–44, 2014. Available at: https://doi.org/10.1093/pan/mpt012

BANSAL, Barkha; SRIVASTAVA, Sangeet. On predicting elections with hybrid topic based sentiment analysis of tweets. **Procedia Computer Science**, [*s. l.*], vol. 135, no. 2018, p. 346–353, 2018. Available at: https://doi.org/10.1016/j.procs.2018.08.183

BEAUCHAMP, Nicholas. Predicting and Interpolating State-Level Polls Using Twitter Textual Data. **American Journal of Political Science**, [*s. l.*], vol. 61, no. 2, p. 490–503, 2017. Available at: https://doi.org/10.1111/ajps.12274

BERG, Joyce *et al.* Results from a Dozen Years of Election Futures Markets Research. *In*: [*S. l.: s. n.*], 2008. p. 742–751. Available at: https://doi.org/10.1016/S1574-0722(07)00080-7

BERGHEL, Hal. Malice Domestic: The Cambridge Analytica Dystopia. **Computer**, [*s. l.*], vol. 51, no. 5, p. 84–89, 2018. Available at: https://doi.org/10.1109/MC.2018.2381135

BERGSTRA, James; BENGIO, Yoshua. Random Search for Hyper-Parameter Optimization. **Journal of Machine Learning Research**, [*s. l.*], vol. 13, p. 281–305, 2012.

BERNERS-LEE, T; FISCHETTI, M. **Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web**. New York: Harper Collins, 1999.

BESSI, Alessandro; FERRARA, Emilio. Social bots distort the 2016 U.S. Presidential election online discussion. **First Monday**, [*s. l.*], vol. 21, no. 11, 2016. Available at: https://doi.org/10.5210/fm.v21i11.7090

BILAL, Muhammad *et al.* Predicting Elections: Social Media Data and Techniques. *In*: , 2019. **2019 International Conference on Engineering and Emerging Technologies (ICEET)**. [*S. l.*]: IEEE, 2019. p. 1–6. Available at: https://doi.org/10.1109/CEET1.2019.8711854

BIMBER, Bruce. Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment. **Journal of Information Technology & Politics**, [*s. l.*], vol. 11, no. 2, p. 130–150, 2014. Available at: https://doi.org/10.1080/19331681.2014.895691

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. **Journal of Machine Learning Research**, [*s. l.*], vol. 3, no. Jan, p. 993–1022, 2003.

BLUMENTHAL, Mark. Polls, Forecasts, and Aggregators. **PS: Political Science & Politics**, [*s. l.*], vol. 47, no. 02, p. 297–300, 2014. Available at: https://doi.org/10.1017/S1049096514000055

BRAGA, Antonio de Pádua; CARVALHO, André Ponce de Leon F. de; LUDERMIR, Teresa Bernarda. **Redes Neurais Artificiais - Teoria e Aplicações**. 2nd editioed. [*S. l.*]: LTC, 2007.

BRICK, J. Michael. The Future of Survey Sampling. **Public Opinion Quarterly**,

[*s. l.*], vol. 75, no. 5, p. 872–888, 2011. Available at: https://doi.org/10.1093/poq/nfr045

BRITO, Kellyton *et al.* Social Media and Presidential Campaigns – Preliminary Results of the 2018 Brazilian Presidential Election. *In*: , 2019, Dubai, United Arab Emirates. **Proceedings of the 20th Annual International Conference on Digital Government Research**. Dubai, United Arab Emirates: ACM, 2019. p. 332–341. Available at: https://doi.org/10.1145/3325112.3325252

BRITO, Kellyton dos Santos *et al.* Assessing the Benefits of Open Government Data: the Case of Meu Congresso Nacional in Brazilian Elections 2014. *In*: , 2015a, New York, NY, USA. **Proceedings of the 16th Annual International Conference on Digital Government Research**. New York, NY, USA: ACM, 2015. p. 89–96. Available at: https://doi.org/10.1145/2757401.2757422

BRITO, Kellyton dos Santos *et al.* Brazilian government open data: implementation, challenges, and potential opportunities. *In*: , 2014a, New York, New York, USA. **Proceedings of the 15th Annual International Conference on Digital Government Research - dg.o '14**. New York, New York, USA: ACM Press, 2014. p. 11–16. Available at: https://doi.org/10.1145/2612733.2612770

BRITO, Kellyton dos Santos *et al.* Evolution of the Web of Social Machines: A Systematic Review and Research Challenges. **IEEE Transactions on Computational Social Systems**, [*s. l.*], vol. 7, no. 2, p. 373–388, 2020. Available at: https://doi.org/10.1109/TCSS.2019.2961269

BRITO, Kellyton dos Santos *et al.* Experiences Integrating Heterogeneous Government Open Data Sources to Deliver Services and Promote Transparency in Brazil. *In*: , 2014b. **2014 IEEE 38th Annual Computer Software and Applications Conference**. [*S. l.*]: IEEE, 2014. p. 606–607. Available at: https://doi.org/10.1109/COMPSAC.2014.87

BRITO, Kellyton dos Santos *et al.* Is Brazilian Open Government Data Actually Open Data? An Analysis of the Current Scenario. **International Journal of E-Planning Research**, [*s. l.*], vol. 4, no. 2, p. 57–73, 2015b. Available at: https://doi.org/10.4018/ijepr.2015040104

BRITO, Kellyton dos Santos *et al.* Using parliamentary Brazilian open data to improve transparency and public participation in Brazil. *In*: , 2014c, New York, New York, USA. **Proceedings of the 15th Annual International Conference on Digital Government Research - dg.o '14**. New York, New York, USA: ACM Press, 2014. p. 171–177. Available at: https://doi.org/10.1145/2612733.2612769

BRITO, Kellyton dos Santos; ADEODATO, Paulo Jorge Leitao. Predicting Brazilian and U.S. Elections with Machine Learning and Social Media Data. *In*: , 2020, Glasgow, United Kingdom. **2020 International Joint Conference on Neural Networks (IJCNN)**. Glasgow, United Kingdom: IEEE, 2020. p. 1–8. Available at: https://doi.org/10.1109/IJCNN48605.2020.9207147

BRITO, Kellyton dos Santos; SILVA FILHO, Rogerio Luiz Cardoso; ADEODATO, Paulo Jorge Leitao. A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions. **IEEE Transactions on Computational Social Systems**, [*s. l.*], p. 1–25, 2021. Available at: https://doi.org/10.1109/TCSS.2021.3063660

BRITO, Kellyton Santos *et al.* How People Care about Their Personal Data Released on Social Media. *In*: , 2013, Tarragona, Spain. **2013 11th Annual Conference on Privacy, Security and Trust, PST 2013**. Tarragona, Spain: [*s. n.*], 2013. Available at: https://doi.org/10.1109/PST.2013.6596044

BRITO, Kellyton Santos *et al.* Implementing Web Applications as Social Machines Composition: A Case Study. *In*: , 2012, Redwood City, USA. **24th International Conference on Software Engineering & Knowledge Engineering (SEKE'2012)**. Redwood City, USA: [*s. n.*], 2012. p. 311–314.

BUREGIO, Vanilson; MEIRA, Silvio; ROSA, Nelson. Social Machines: A Unified Paradigm to Describe Social Web-Oriented Systems. *In*: , 2013, New York, New York, USA. **Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion**. New York, New York, USA: ACM Press, 2013. p. 885–890. Available at: https://doi.org/10.1145/2487788.2488074

BURÉGIO, Vanilson *et al.* Towards Government as a Social Machine. *In*: , 2015. **Proceedings of the 24th International Conference on World Wide Web**. [*S. l.: s. n.*], 2015. p. 1131–1136.

BURÉGIO, Vanilson André de Arruda. **Social Machines: A Unified Paradigm to Describe, Design and Implement Emerging Social Systems**. 187 f. 2014. - Pernambuco Federal University, [*s. l.*], 2014.

CENTRE FOR REVIEWS AND DISSEMINATION. **Systematic reviews: CRD's guidance for undertaking reviews in health care**. [*S. l.*]: CRD, University of York,York, UK, 2009. *E-book*.

CERON-GUZMAN, Jhon Adrian; LEON-GUZMAN, Elizabeth. A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential

Election. *In*: , 2016. **2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)**. [*S. l.*]: IEEE, 2016. p. 250–257. Available at: https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.47

CERON, Andrea *et al.* Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. **New Media & Society**, [*s. l.*], vol. 16, no. 2, p. 340–358, 2014. Available at: https://doi.org/10.1177/1461444813480466

CHOI, Bumghi; LEE, Ju-Hong; KIM, Deok-Hwan. Solving local minima problem with large number of hidden nodes on two-layered feed-forward artificial neural networks. **Neurocomputing**, [*s. l.*], vol. 71, no. 16–18, p. 3640–3643, 2008. Available at: https://doi.org/10.1016/j.neucom.2008.04.004

COGBURN, Derrick L.; ESPINOZA-VASQUEZ, Fatima K. From Networked Nominee to Networked Nation: Examining the Impact of Web 2.0 and Social Media on Political Participation and Civic Engagement in the 2008 Obama Campaign. **Journal of Political Marketing**, [*s. l.*], vol. 10, no. 1–2, p. 189–213, 2011. Available at: https://doi.org/10.1080/15377857.2011.540224

COUPER, M. P. The Future of Modes of Data Collection. **Public Opinion Quarterly**, [*s. l.*], vol. 75, no. 5, p. 889–908, 2011. Available at: https://doi.org/10.1093/poq/nfr046

CROSSLEY, Archibald M. Straw Polls in 1936. **Public Opinion Quarterly**, [*s. l.*], vol. 1, no. 1, p. 24, 1937. Available at: https://doi.org/10.1086/265035

CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals, and Systems**, [*s. l.*], vol. 2, no. 4, p. 303–314, 1989. Available at: https://doi.org/10.1007/BF02551274

DA SILVA, Fabio Q.B. *et al.* Six years of systematic literature reviews in software engineering: An updated tertiary study. **Information and Software Technology**, [*s. l.*], vol. 53, no. 9, p. 899–913, 2011. Available at: https://doi.org/10.1016/j.infsof.2011.04.004

DING, Ling; RANGARAJU, Prasad; POURSAEE, Amir. Application of generalized regression neural network method for corrosion modeling of steel embedded in soil. **Soils and Foundations**, [*s. l.*], vol. 59, no. 2, p. 474–483, 2019. Available at: https://doi.org/10.1016/j.sandf.2018.12.016

EASTERBROOK, Steve *et al.* Selecting Empirical Methods for Software Engineering Research. *In*: SHULL, F; SINGER, J (eds.). **Guide to Advanced Empirical Software Engineering**. London: Springer London, 2008. p. 285–311. Available at: https://doi.org/10.1007/978-1-84800-044-5_11

EFRON, Brandley *et al.* Least Angle Regression. **The Annals of Statistics**, [*s. l.*], vol. 32, no. 2, p. 407–4099, 2004.

ELTANTAWY, Nahed; WIEST, Julie B. Social Media in the Egyptian Revolution: Reconsidering Resource Mobilization Theory. **International Journal of Communications**, [*s. l.*], vol. 5, p. 1207–1224, 2011.

ERIKSON, R. S.; WLEZIEN, C. Are Political Markets Really Superior to Polls as Election Predictors? **Public Opinion Quarterly**, [*s. l.*], vol. 72, no. 2, p. 190–215, 2008. Available at: https://doi.org/10.1093/poq/nfn010

FACEBOOK INC. **Facebook API**. [*S. l.*], 2020. Available at: https://developers.facebook.com/docs/graph-api. Accessed at: 21 Nov. 2020.

FAIR, Ray C. The Effect of Economic Events on Votes for President. **The Review of Economics and Statistics**, [*s. l.*], vol. 60, no. 2, p. 159, 1978. Available at: https://doi.org/10.2307/1924969

FERNÁNDEZ-DELGADO, M. *et al.* An extensive experimental survey of regression methods. **Neural Networks**, [*s. l.*], vol. 111, p. 11–34, 2019. Available at: https://doi.org/10.1016/j.neunet.2018.12.010

FILER, Tanya; FREDHEIM, Rolf. Popular with the Robots: Accusation and Automation in the Argentine Presidential Elections, 2015. **International Journal of Politics, Culture and Society**, [*s. l.*], vol. 30, no. 3, p. 259–274, 2017. Available at: https://doi.org/10.1007/s10767-016-9233-7

FIORINA, Morris P. Economic Retrospective Voting in American National Elections: A Micro-Analysis. **American Journal of Political Science**, [*s. l.*], vol. 22, no. 2, p. 426, 1978. Available at: https://doi.org/10.2307/2110623

FLAXMAN, Seth; GOEL, Sharad; RAO, Justin M. Filter bubbles, echo chambers, and online news consumption. **Public Opinion Quarterly**, [*s. l.*], vol. 80, no. S1, p. 298–320, 2016. Available at: https://doi.org/10.1093/poq/nfw006

FRANCIA, Peter L. Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump. **Social Science Computer Review**, [*s. l.*], vol. 36, no. 4, p. 440–455, 2018. Available at: https://doi.org/10.1177/0894439317730302

GALLUP, George. Polls and the Political Process-Past, Present, and Future. **Public Opinion Quarterly**, [*s. l.*], vol. 29, no. 4, p. 544, 1965. Available at: https://doi.org/10.1086/267358

GAURAV, Manish *et al.* Leveraging candidate popularity on Twitter to predict election outcome. *In*: , 2013, New York, New York, USA. **Proceedings of the 7th Workshop on Social Network Mining and Analysis - SNAKDD '13**. New York, New York, USA: ACM Press, 2013. p. 1–8. Available at: https://doi.org/10.1145/2501025.2501038

GAYO-AVELLO, Daniel. A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. **Social Science Computer Review**, [*s. l.*], vol. 31, no. 6, p. 649–679, 2013. Available at: https://doi.org/10.1177/0894439313493979

GAYO-AVELLO, Daniel. Don't turn social media into another "Literary Digest" poll. **Communications of the ACM**, [*s. l.*], vol. 54, no. 10, p. 121, 2011. Available at: https://doi.org/10.1145/2001269.2001297

GAYO-AVELLO, Daniel; METAXAS, P.T.; MUSTAFARAJ, Eni. **Limits of Electoral Predictions using Social Media Data**. [*S. l.: s. n.*], 2011.

GESCHKE, Daniel; LORENZ, Jan; HOLTZ, Peter. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. **British Journal of Social Psychology**, [*s. l.*], vol. 58, no. 1, p. 129–149, 2019. Available at: https://doi.org/10.1111/bjso.12286

GIL DE ZÚÑIGA, Homero; JUNG, Nakwon; VALENZUELA, Sebastián. Social Media Use for News and Individuals' Social Capital, Civic Engagement and Political Participation. **Journal of Computer-Mediated Communication**, [*s. l.*], vol. 17, no. 3, p. 319–336, 2012. Available at: https://doi.org/10.1111/j.1083-6101.2012.01574.x

GOEBEL, Randy *et al.* Explainable AI: The New 42? *In*: MACHINE LEARNING AND KNOWLEDGE EXTRACTION. [*S. l.*]: Springer, Cham, Switzerland, 2018. p. 295–303. Available at: https://doi.org/10.1007/978-3-319-99740-7_21

GOODWIN, Laura D.; LEECH, Nancy L. Understanding Correlation: Factors That Affect the Size of r. **The Journal of Experimental Education**, [*s. l.*], vol. 74, no. 3, p. 249–266, 2006. Available at: https://doi.org/10.3200/JEXE.74.3.249-266

GOOGLE LLC. **YouTube API**. [*S. l.*], 2020. Available at: https://developers.google.com/youtube/v3. Accessed at: 21 Nov. 2020.

GOTO, Hisaki; GOTO, Yukiko. Regression Analysis of National Elections in Japan Using Social Listening. *In*: , 2019. **12th IADIS International Conference**

Information Systems 2019. [*S. l.*]: IADIS Press, 2019. p. 189–196. Available at: https://doi.org/10.33965/is2019_201905L024

HALL, Mark *et al.* The WEKA Data Mining Software: An Update. **SIGKDD Explorations,** [*s. l.*], vol. 11, no. 1, 2009.

HALL, Wendy; TINATI, Ramine; JENNINGS, Will. From brexit to trump: Social media's role in democracy. **Computer**, [*s. l.*], vol. 51, no. 1, p. 18–27, 2018. Available at: https://doi.org/10.1109/MC.2018.1151005

HASSOUN, Mohamad H. **Fundamentals of Artificial Neural Networks**. [*S. l.*]: MIT Press, 2010.

HAWKINS, Douglas M. On the Investigation of Alternative Regressions by Principal Component Analysis. **Applied Statistics**, [*s. l.*], vol. 22, no. 3, p. 275, 1973. Available at: https://doi.org/10.2307/2346776

HAYKIN, Simon. **Neural Networks: A Comprehensive Foundation**. 2nd editioed. [*S. l.*]: Pearson, 1998.

HE, Xin; ZHAO, Kaiyong; CHU, Xiaowen. AutoML: A survey of the state-of-the-art. **Knowledge-Based Systems**, [*s. l.*], vol. 212, p. 106622, 2021. Available at: https://doi.org/10.1016/j.knosys.2020.106622

HEREDIA, Brian; PRUSA, Joseph D.; KHOSHGOFTAAR, Taghi M. Social media for polling and predicting United States election outcome. **Social Network Analysis and Mining**, [*s. l.*], vol. 8, no. 1, p. 48, 2018. Available at: https://doi.org/10.1007/s13278-018-0525-y

HILLYGUS, D. S. The Evolution of Election Polling in the United States. **Public Opinion Quarterly**, [*s. l.*], vol. 75, no. 5, p. 962–981, 2011. Available at: https://doi.org/10.1093/poq/nfr054

HOLBROOK, Thomas M. Good News for Bush? Economic News, Personal Finances, and the 2004 Presidential Election. **PS: Political Science & Politics**, [*s. l.*], vol. 37, no. 4, p. 759–761, 2004.

HORNIK, Kurt; STINCHCOMBE, Maxwell; WHITE, Halbert. Multilayer feedforward networks are universal approximators. **Neural Networks**, [*s. l.*], vol. 2, no. 5, p. 359–366, 1989. Available at: https://doi.org/10.1016/0893-6080(89)90020-8

INSTITUTO NACIONAL ELECTORAL. **Encuestas Electorales: Estudios entregados a la Secretaría Ejecutiva**. [*S. l.*], 2020. Available at: https://www.ine.mx/voto-y-elecciones/encuestas-electorales/elecciones-federales-ordinarias-2017-2018-estudios-entregados/. Accessed at: 1 Oct. 2020.

ISAAK, Jim; HANNA, Mina J. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. **Computer**, [*s. l.*], vol. 51, no. 8, p. 56–59, 2018. Available at: https://doi.org/10.1109/MC.2018.3191268

ISOTALO, Veikko *et al.* Predicting 2016 US Presidential Election Polls with Online and Media Variables. *In*: SPRINGER PROCEEDINGS IN COMPLEXITY. [*S. l.*]: Springer, Cham, Switzerland, 2016. p. 45–53. Available at: https://doi.org/10.1007/978-3-319-42697-6_5

JACKMAN, Simon. Pooling the polls over an election campaign. **Australian Journal of Political Science**, [*s. l.*], vol. 40, no. 4, p. 499–517, 2005. Available at: https://doi.org/10.1080/10361140500302472

JENNINGS, Will; WLEZIEN, Christopher. Election polling errors across time and space. **Nature Human Behaviour**, [*s. l.*], vol. 2, no. 4, p. 276–283, 2018. Available at: https://doi.org/10.1038/s41562-018-0315-6

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, [*s. l.*], vol. 349, no. 6245, p. 255–260, 2015. Available at: https://doi.org/10.1126/science.aaa8415

JUNGHERR, Andreas *et al.* Digital Trace Data in the Study of Public Opinion. **Social Science Computer Review**, [*s. l.*], vol. 35, no. 3, p. 336–356, 2017. Available at: https://doi.org/10.1177/0894439316631043

JUNGHERR, Andreas; JÜRGENS, Pascal; SCHOEN, Harald. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, A., sprenger, T. O., sander, P. G., & welpe, I. M. “predicting elections with twitter: What 140 characters reveal about political sentiment.” **Social Science Computer Review**, [*s. l.*], 2012. Available at: https://doi.org/10.1177/0894439311404119

JUSTINO GARCIA PRACIANO, Bruno *et al.* Spatio-Temporal trend analysis of the brazilian elections based on twitter data. *In*: , 2019. **IEEE International Conference on Data Mining Workshops, ICDMW**. [*S. l.: s. n.*], 2019. Available at: https://doi.org/10.1109/ICDMW.2018.00192

KALAMPOKIS, Evangelos; TAMBOURIS, Efthimios; TARABANIS, Konstantinos. Understanding the predictive power of social media. **Internet Research**, [*s. l.*], vol. 23, no. 5, p. 544–559, 2013. Available at: https://doi.org/10.1108/IntR-06-2012-0114

KARLIK, Bekir; OLGAC, Av. Performance analysis of various activation

functions in generalized MLP architectures of neural networks. **International Journal of Artificial Intelligence and Expert Systems (IJAE)**, [*s. l.*], vol. 1, no. 4, 2011.

KEMP, Simon; WE ARE SOCIAL; HOOTSUITE. **DIGITAL 2020: Argentina**. [*S. l.*], 2020a. Available at: https://datareportal.com/reports/digital-2020-argentina.

KEMP, Simon; WE ARE SOCIAL; HOOTSUITE. **DIGITAL 2020: Brazil**. [*S. l.*], 2020b. Available at: https://datareportal.com/reports/digital-2020-brazil.

KEMP, Simon; WE ARE SOCIAL; HOOTSUITE. **DIGITAL 2020: Colombia**. [*S. l.*], 2020c. Available at: https://datareportal.com/reports/digital-2020-colombia.

KEMP, Simon; WE ARE SOCIAL; HOOTSUITE. **DIGITAL 2020: Mexico**. [*S. l.*], 2020d. Available at: https://datareportal.com/reports/digital-2020-mexico.

KENT, Michael L.; LI, Chaoyuan. Toward a normative social media theory for public relations. **Public Relations Review**, [*s. l.*], vol. 46, no. 1, p. 101857, 2020. Available at: https://doi.org/10.1016/j.pubrev.2019.101857

KHAMIS, Susie; ANG, Lawrence; WELLING, Raymond. Self-branding, 'micro-celebrity' and the rise of Social Media Influencers. **Celebrity Studies**, [*s. l.*], vol. 8, no. 2, p. 191–208, 2017. Available at: https://doi.org/10.1080/19392397.2016.1218292

KHONDKER, Habibul Haque. Role of the New Media in the Arab Spring. **Globalizations**, [*s. l.*], vol. 8, no. 5, p. 675–679, 2011. Available at: https://doi.org/10.1080/14747731.2011.621287

KITCHENHAM, Barbara; BRERETON, Pearl. A systematic review of systematic review process research in software engineering. **Information and Software Technology**, [*s. l.*], vol. 55, no. 12, p. 2049–2075, 2013. Available at: https://doi.org/10.1016/j.infsof.2013.07.010

KOLI, Abdul Manan; AHMED, Muqeem; MANHAS, Jatinder. An Empirical Study on Potential and Risks of Twitter Data for Predicting Election Outcomes. *In*: RATHORE, V *et al.* (eds.). **Advances in Intelligent Systems and Computing**. [*S. l.*]: Springer, Singapore, 2019. p. 725–731. Available at: https://doi.org/10.1007/978-981-13-2285-3_85

KREHBIEL, Timothy C. Correlation Coefficient Rule of Thumb. **Decision Sciences Journal of Innovative Education**, [*s. l.*], vol. 2, no. 1, p. 97–100, 2004. Available at: https://doi.org/10.1111/j.0011-7315.2004.00025.x

KREISS, Daniel; LAWRENCE, Regina G.; MCGREGOR, Shannon C. In Their Own Words: Political Practitioner Accounts of Candidates, Audiences, Affordances, Genres, and Timing in Strategic Social Media Use. **Political Communication**, [*s. l.*],

2018. Available at: https://doi.org/10.1080/10584609.2017.1334727

KROGH, Anders; VEDELSBY, Jesper. Neural network ensembles, cross validation and active learning. *In*: , 1994. **Proceedings of the 7th International Conference on Neural Information Processing Systems**. [*S. l.: s. n.*], 1994. p. 231–238.

KWAK, Jin-ah; CHO, Sung Kyum. Analyzing Public Opinion with Social Media Data during Election Periods: A Selective Literature Review. **Asian Journal for Public Opinion Research**, [*s. l.*], vol. 5, no. 4, p. 285–301, 2018. Available at: https://doi.org/http://doi.org/10.15206/ajpor.2018.5.4.285

LARS BACKSTROM; THE FACEBOOK. **News Feed FYI: A Window Into News Feed**. [*S. l.*], 2013. Available at: https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed. Accessed at: 19 Nov. 2020.

LEEUW, Edith D De. To Mix or Not to Mix Data Collection Modes in Surveys. **Journal of Official Statistics**, [*s. l.*], vol. 21, no. 2, p. 233–255, 2005.

LERMAN, P. M. Fitting Segmented Regression Models by Grid Search. **Applied Statistics**, [*s. l.*], vol. 29, no. 1, p. 77, 1980. Available at: https://doi.org/10.2307/2346413

LEWIS-BECK, Michael S. Election Forecasting: Principles and Practice. **The British Journal of Politics and International Relations**, [*s. l.*], vol. 7, no. 2, p. 145–164, 2005. Available at: https://doi.org/10.1111/j.1467-856X.2005.00178.x

LEWIS-BECK, Michael S.; RICE, Tom W. Presidential Popularity and Presidential Vote. **Public Opinion Quarterly**, [*s. l.*], vol. 46, no. 4, p. 534, 1982. Available at: https://doi.org/10.1086/268750

LOSING, Viktor; HAMMER, Barbara; WERSING, Heiko. Incremental on-line learning: A review and comparison of state of the art algorithms. **Neurocomputing**, [*s. l.*], vol. 275, p. 1261–1274, 2018. Available at: https://doi.org/10.1016/j.neucom.2017.06.084

LU, Jie *et al.* Learning under Concept Drift: A Review. **IEEE Transactions on Knowledge and Data Engineering**, [*s. l.*], p. 1–1, 2018. Available at: https://doi.org/10.1109/TKDE.2018.2876857

MARTIN, E. A. A Review and Proposal for a New Measure of Poll Accuracy. **Public Opinion Quarterly**, [*s. l.*], vol. 69, no. 3, p. 342–369, 2005. Available at: https://doi.org/10.1093/poq/nfi044

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, [*s. l.*], vol. 5, no. 4, p. 115–133, 1943. Available at: https://doi.org/10.1007/BF02478259

MEDSKER, L R; JAIN, L C. Recurrent Neural Networks: Design and Applications. **Book**, [*s. l.*], 2001. Available at: https://doi.org/10.1201/9781420049176

MEIRA, Silvio R L *et al.* The Emerging Web of Social Machines. *In*: , 2011, Munich. **Computer Software and Applications Conference (COMPSAC)**. Munich: [*s. n.*], 2011. p. 26–27.

MITOFSKY, Warren J. Review: Was 1996 a Worse Year for Polls Than 1948? **Public Opinion Quarterly**, [*s. l.*], vol. 62, no. 2, p. 230–249, 1998. Available at: https://doi.org/10.1086/297842

MONDAK, Jeffery J. Media exposure and political discussion in U.S. elections. **The Journal of Politics**, [*s. l.*], vol. 57, no. 1, 1995. Available at: https://doi.org/10.2307/2960271

MOSTELLER, Frederick *et al.* **The Pre-Election Polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts**. New York, USA: [*s. n.*], 1949.

MURPHY, Sheila T.; ZAJONC, R. B. Affect, Cognition, and Awareness: Affective Priming With Optimal and Suboptimal Stimulus Exposures. **Journal of Personality and Social Psychology**, [*s. l.*], vol. 64, no. 5, p. 723–729, 1993. Available at: https://doi.org/10.1037/0022-3514.64.5.723

MUSTAFARAJ, Eni *et al.* Vocal minority versus silent majority: Discovering the opionions of the long tail. *In*: , 2011. **Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011**. [*S. l.: s. n.*], 2011. Available at: https://doi.org/10.1109/PASSAT/SocialCom.2011.188

MUSTAFARAJ, Eni; METAXAS, Panagiotis Takis. The Fake News Spreading Plague: Was it Preventable? *In*: , 2017. **Proceedings of ACM Web Science Conference**. [*S. l.: s. n.*], 2017. p. 235–239.

NÓBREGA, Jarley P.; OLIVEIRA, Adriano L.I. A sequential learning method with Kalman filter and extreme learning machine for regression and time series forecasting. **Neurocomputing**, [*s. l.*], vol. 337, p. 235–250, 2019. Available at: https://doi.org/10.1016/j.neucom.2019.01.070

O'CONNOR, Brendan *et al.* From tweets to polls: Linking text sentiment to

public opinion time series. *In*: , 2010. **4th International AAAI Conference on Weblogs and Social Media**. [*S. l.: s. n.*], 2010.

O'LEARY, Daniel E. Twitter Mining for Discovery, Prediction and Causality: Applications and Methodologies. **Intelligent Systems in Accounting, Finance and Management**, [*s. l.*], vol. 22, no. 3, p. 227–247, 2015. Available at: https://doi.org/10.1002/isaf.1376

OLIVEIRA, Adriano Lorena Inácio de. **Neural Networks Forecasting and Classification-Based Techniques for Novelty Detection in Time Series**. 180 f. 2004. - Universidade Federal de Pernambuco, [*s. l.*], 2004.

OPPENHEIMER, Bruce I.; STIMSON, James A.; WATERMAN, Richard W. Interpreting U. S. Congressional Elections: The Exposure Thesis. **Legislative Studies Quarterly**, [*s. l.*], vol. 11, no. 2, p. 227–247, 1986. Available at: https://doi.org/10.2307/439877

PEDREGOSA, Fabian *et al.* Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, [*s. l.*], vol. 12, no. 85, p. 2825−2830, 2011.

PERRONE, Michael Peter. **Improving regression estimates: averaging methods for variance reduction with extensions to general convex measure optimization**. 1993. - Brown University, [*s. l.*], 1993.

PEW RESEARCH CENTER. **Sizing Up Twitter Users**. [*S. l.: s. n.*], 2019. Available at: https://www.pewinternet.org/2019/04/24/sizing-up-twitter-users/.

PRADA, Jesus. Predicting With Twitter. *In*: , 2015. **European Conference on Social Media**. [*S. l.: s. n.*], 2015. p. 10.

PRASETYO, Nugroho Dwi; HAUFF, Claudia. Twitter-based Election Prediction in the Developing World. **Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15**, [*s. l.*], 2015. Available at: https://doi.org/10.1145/2700171.2791033

QIU, Junfei *et al.* A survey of machine learning for big data processing. **EURASIP Journal on Advances in Signal Processing**, [*s. l.*], vol. 2016, no. 1, p. 67, 2016. Available at: https://doi.org/10.1186/s13634-016-0355-x

REAL CLEAR POLITICS. **General Election: Trump vs. Clinton**. [*S. l.*], 2016. Available at: https://www.realclearpolitics.com/epolls/2016/president/us/general_election_trump_vs_clinton-5491.html. Accessed at: 1 Nov. 2019.

RHODE, Paul W; STRUMPF, Koleman S. Historical Presidential Betting

Markets. **Journal of Economic Perspectives**, [*s. l.*], vol. 18, no. 2, p. 127–142, 2004. Available at: https://doi.org/10.1257/0895330041371277

RODRÍGUEZ, Sebastián *et al.* Forecasting the Chilean Electoral Year: Using Twitter to Predict the Presidential Elections of 2017. *In*: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS). [*S. l.*]: Springer, Cham, Switzerland, 2018. p. 298–314. Available at: https://doi.org/10.1007/978-3-319-91485-5_23

RODRIGUEZ, J.D.; PEREZ, A.; LOZANO, J.A. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [*s. l.*], vol. 32, no. 3, p. 569–575, 2010. Available at: https://doi.org/10.1109/TPAMI.2009.187

ROGALEWICZ, Michał; SIKA, Robert. Methodologies of Knowledge Discovery from Data and Data Mining Methods in Mechanical Engineering. **Management and Production Engineering Review**, [*s. l.*], vol. 7, no. 4, p. 97–108, 2016. Available at: https://doi.org/10.1515/mper-2016-0040

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, [*s. l.*], vol. 65, no. 6, p. 386–408, 1958. Available at: https://doi.org/10.1037/h0042519

RUDIN, Cynthia; WAGSTAFF, Kiri L. Machine learning for science and society. **Machine Learning**, [*s. l.*], vol. 95, no. 1, p. 1–9, 2014. Available at: https://doi.org/10.1007/s10994-013-5425-9

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, [*s. l.*], vol. 323, no. 6088, p. 533–536, 1986. Available at: https://doi.org/10.1038/323533a0

SALARI, Sasan *et al.* Estimation of 2017 Iran's Presidential Election Using Sentiment Analysis on Social Media. *In*: , 2018. **2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)**. [*S. l.*]: IEEE, 2018. p. 77–82. Available at: https://doi.org/10.1109/ICSPIS.2018.8700529

SANG, Etk; BOS, Johan. Predicting the 2011 dutch senate election results with twitter. *In*: , 2012. **Proceedings of the Workshop on Semantic Analysis in Social Media**. [*S. l.: s. n.*], 2012. p. 53–60.

SCHMIDHUBER, Jürgen. Deep learning in neural networks: An overview. **Neural Networks**, [*s. l.*], vol. 61, p. 85–117, 2015. Available at:

https://doi.org/10.1016/j.neunet.2014.09.003

SEAL, HILARY L. Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. **Biometrika**, [*s. l.*], vol. 54, no. 1–2, p. 1–24, 1967. Available at: https://doi.org/10.1093/biomet/54.1-2.1

SEBER, George A. F.; LEE, Alan J. **Linear Regression Analysis**. [*S. l.*]: Wiley, 2003. (Wiley Series in Probability and Statistics). Available at: https://doi.org/10.1002/9780471722199

SHEARER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing**, [*s. l.*], vol. 5, no. 4, p. 13--22, 2000.

SIGELMAN, Lee. Presidential Popularity and Presidential Elections. **Public Opinion Quarterly**, [*s. l.*], vol. 43, no. 4, p. 532, 1979. Available at: https://doi.org/10.1086/268549

SINGER, Eleanor; YE, Cong. The Use and Effects of Incentives in Surveys. **The ANNALS of the American Academy of Political and Social Science**, [*s. l.*], vol. 645, no. 1, p. 112–141, 2013. Available at: https://doi.org/10.1177/0002716212458082

SINGH, Prabhsimran; SAWHNEY, Ravinder Singh; KAHLON, Karanjeet Singh. Forecasting the 2016 US Presidential Elections Using Sentiment Analysis. *In*: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS). [*S. l.*]: Springer, Cham, Switzerland, 2017. p. 412–423. Available at: https://doi.org/10.1007/978-3-319-68557-1_36

SINGHAL, Kartik; AGRAWAL, Basant; MITTAL, Namita. Modeling Indian General Elections: Sentiment Analysis of Political Twitter Data. *In*: ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING. [*S. l.*]: Springer, New Delhi, India, 2015. p. 469–477. Available at: https://doi.org/10.1007/978-81-322-2250-7_46

SPECHT, D.F. A general regression neural network. **IEEE Transactions on Neural Networks**, [*s. l.*], vol. 2, no. 6, p. 568–576, 1991. Available at: https://doi.org/10.1109/72.97934

SPECHT, Donald F. Probabilistic neural networks. **Neural Networks**, [*s. l.*], vol. 3, no. 1, p. 109–118, 1990. Available at: https://doi.org/10.1016/0893-6080(90)90049-Q

SPOHR, Dominic. Fake news and ideological polarization. **Business Information Review**, [*s. l.*], vol. 34, no. 3, p. 150–160, 2017. Available at: https://doi.org/10.1177/0266382117722446

STANLEY, K.O.; MIIKKULAINEN, R. Efficient evolution of neural network topologies. *In*: , 2002. **Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)**. [*S. l.*]: IEEE, 2002. p. 1757–1762. Available at: https://doi.org/10.1109/CEC.2002.1004508

STUDENT. The Probable Error of a Mean. **Biometrika**, [*s. l.*], vol. 6, no. 1, p. 1, 1908. Available at: https://doi.org/10.2307/2331554

SWAP, Walter C. Interpersonal Attraction and Repeated Exposure to Rewarders and Punishers. **Personality and Social Psychology Bulletin**, [*s. l.*], 1977. Available at: https://doi.org/10.1177/014616727700300219

THE HUFFINGTON POST. **2016 General Election: Trump vs. Clinton**. [*S. l.*], 2016. Available at: https://elections.huffingtonpost.com/pollster/2016-general-election-trump-vs-clinton. Accessed at: 1 Nov. 2019.

THE NEW YORK TIMES. **The NYT Last Elections Polls 2016**. [*S. l.*], 2016. Available at: https://www.nytimes.com/interactive/2016/us/elections/polls.html. Accessed at: 1 Nov. 2019.

TIBSHIRANI, Robert. Regression Shrinkage and Selection Via the Lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, [*s. l.*], vol. 58, no. 1, p. 267–288, 1996. Available at: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

TILTON, Shane. Virtual polling data: A social network analysis on a student government election. **Webology**, [*s. l.*], 2008.

TRESP, Volker. Committee Machines. *In*: HU, Yu Hen; HWANG, Jenq-Neng (eds.). **Handbook of Neural Network Signal Processing**. [*S. l.*]: CRC Press, 2001. Available at: https://doi.org/10.1201/9781315220413-5

TRUNK, G. V. A Problem of Dimensionality: A Simple Example. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [*s. l.*], vol. PAMI-1, no. 3, p. 306–307, 1979. Available at: https://doi.org/10.1109/TPAMI.1979.4766926

TSAKALIDIS, Adam *et al.* Predicting Elections for Multiple Countries Using Twitter and Polls. **IEEE Intelligent Systems**, [*s. l.*], 2015. Available at: https://doi.org/10.1109/MIS.2015.17

TUMASJAN, Andranik *et al.* Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *In*: , 2010. **Fourth International AAAI Conference on Weblogs and Social Media**. [*S. l.: s. n.*], 2010.

TWITTER INC. **Twitter API**. [*S. l.*], 2020. Available at:

https://developer.twitter.com/en/docs. Accessed at: 21 Nov. 2020.

UNITED NATIONS. **World Population Prospects 2019**. [*S. l.*], 2019. Available at: https://population.un.org/wpp/Download/Files/1_Indicators (Standard)/EXCEL_FILES/1_Population/WPP2019_POP_F01_1_TOTAL_POPULATION_BOTH_SEXES.xlsx. Accessed at: 5 Jan. 2021.

VAPNIK, Vladimir; LEVIN, Esther; CUN, Yann Le. Measuring the VC-Dimension of a Learning Machine. **Neural Computation**, [*s. l.*], vol. 6, no. 5, p. 851–876, 1994. Available at: https://doi.org/10.1162/neco.1994.6.5.851

WANG, Yaqing *et al.* Generalizing from a Few Examples: A Survey on Few-shot learning. **ACM Computing Surveys**, [*s. l.*], vol. 53, no. 3, p. 1–34, 2020. Available at: https://doi.org/10.1145/3386252

WE ARE SOCIAL; HOOTSUITE. **DIGITAL 2020: JULY GLOBAL STATSHOT**. [*S. l.*], 2020. Available at: https://datareportal.com/reports/digital-2020-july-global-statshot. Accessed at: 18 Nov. 2020.

WEI, William W.S. **Time Series Analysis**. [*S. l.*]: Oxford University Press, 2013. Available at: https://doi.org/10.1093/oxfordhb/9780199934898.013.0022

WILCOXON, Frank. Individual Comparisons by Ranking Methods. **Biometrics Bulletin**, [*s. l.*], 1945. Available at: https://doi.org/10.2307/3001968

YAN, Xin; GANG SU, Xiao. **Linear regression analysis: Theory and computing**. [*S. l.: s. n.*], 2009. Available at: https://doi.org/10.1142/6986

YOUTUBE INC. **How YouTube's Home Sceen Works**. [*S. l.*], 2017. Available at: https://www.youtube.com/watch?v=69tpVNunQEU. Accessed at: 19 Nov. 2020.

ZAJONC, ROBERT B. Attitudinal Effects Of Mere Exposure. **Journal of Personality and Social Psychology**, [*s. l.*], vol. 9, no. 2, 1968. Available at: https://doi.org/10.1037/h0025848

ZAJONC, R. B. Feeling and thinking: Preferences need no inferences. **American Psychologist**, [*s. l.*], vol. 35, no. 2, p. 151–175, 1980. Available at: https://doi.org/10.1037/0003-066X.35.2.151

ZAJONC, R. B. Mere exposure: A gateway to the subliminal. **Current Directions in Psychological Science**, [*s. l.*], vol. 10, no. 6, p. 224–228, 2001. Available at: https://doi.org/10.1111/1467-8721.00154

ZHANG, He; BABAR, Muhammad Ali; TELL, Paolo. Identifying relevant studies in software engineering. **Information and Software Technology**, [*s. l.*], vol. 53, no. 6, p. 625–637, 2011. Available at: https://doi.org/10.1016/j.infsof.2010.12.010

ZHANG, Xiaodong. Social media popularity and election results: A study of the 2016 Taiwanese general election. **PLoS ONE**, [*s. l.*], 2018. Available at: https://doi.org/10.1371/journal.pone.0208190