UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

¿TWITTER CONSTITUYENTE? MIDIENDO LA INFLUENCIA DE LOS CANDIDATOS A LA CONVENCIÓN CONSTITUCIONAL EN TWITTER

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

JOSÉ MIGUEL CORDERO CARVACHO

PROFESOR GUÍA:
ANDRÉS ABELIUK KIMELMAN

MIEMBROS DE LA COMISIÓN:
NOMBRE COMPLETO UNO
NOMBRE COMPLETO DOS
NOMBRE COMPLETO TRES

SANTIAGO DE CHILE
2022

Resumen

Abstract

"Power does not reside in institutions, not even the state or large corporations. It is located in the networks that structure society."
Manuel Castells

Acknowledgments

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

On May 15 and 16, 2021, the election of the members of the Constitutional Convention, a body whose function is to draft a new constitution, was held in Chile. The Constitutional Convention was approved with more than 78 % of the votes in a national plebiscite on October 25, 2020 as a way out of a political and social crisis experienced in Chile since 2019. In the May election, who will draft the constitution was defined. The results showed the traditional political blocs greatly diminished and an emerging force of independents with progressive ideas ended up occupying a large part of the convention seats.

The growing digital literacy in Chile, accelerated in part by the COVID-19 pandemic, suggests that social networks played a role in the last elections. The microblogging platform Twitter, the network for political discussion par excellence, is particularly noteworthy. The use of Twitter as a political communication tool is not new in our country: since the Chilean presidential election of 2009-2010, this social network has been adopted by parliamentarians, ministers and candidates. However, Twitter is known as an unrepresentative network, dominated by a "young adult population from well-off sectors" [**?**], but with great capacity to install a media agenda.

Analyzing an election on Twitter from the voters' point of view can be complex: it requires a significant and representative amount of conversation to identify positions for or against candidates or pacts (among other things). However, Twitter can also be studied from the point of view of the candidates. Twitter reflects how different political campaign strategies operate. Candidates express positions, are disseminated (retweeted) and validate other actors by disseminating them.

An electoral campaign objective is by definition *to influence* the behavior of voters. Therefore, the *influence* on Twitter is, for a candidate, an indicator of a successful campaign in that medium. A successful campaign mobilizes more votes. Therefore, the hypothesis that there is a relationship between influence on social networks (in this case Twitter) and the number of votes obtained is reasonable. This bachelor thesis aims to measure precisely that: the relationship between the influence on Twitter and the electoral result.

The historical importance of the Constitutional Convention and its electoral result, lead us to search social networks (in this case Twitter) for the reasons for this result. Analyzing the performance on Twitter of the candidates for the convention allows us to understand the communication strategies of the actors present in the current political scene. Twitter offers a continuous record of *what was said*, *when it was said* and it is also possible to reconstruct the dissemination of these messages.

1.2 Research questions

This research aims to answer the following questions:

- RQ1: To what extent does the influence on Twitter correlates with the votes obtained? What variables does this relationship depend on?
  - RQ1-1: Does the relationship between Twitter data and votes increase aggregating data? (i.e. Using list or party results instead of individual candidate data)
  - RQ1-2: Is the last stretch of campaign (e.g. last two weeks) data more close to the electoral results than older data?

1.3 Objectives

1.3.1 General objective

Choose and compute measures of influence of the candidates for constitutional convention on Twitter and contrast them with the electoral results obtained.

1.3.2 Specific objectives

1. Build a network with information downloaded from Twitter where the candidates are included. This network construction must try to be as representative as possible of the real network.
2. Calculate measures of influence of users on the network with information available according to the limitations of the Twitter API.
3. Evaluate the correspondence between votes and influence.

Chapter 2

Background and Related work

2.1 Twitter and elections

Twitter is widely used as a data source to study political phenomena. Problems such as polarization, the detection of misinformation or even the prediction of elections are frequently studied with data from Twitter. Next, we will review the main methodological difficulties when carrying out these analyzes, and then begin to compare the different sampling strategies and close with the state of the art of these methods in Chilean elections.

It is possible to identify three main difficulties when it comes to quantifying the effect of social networks on elections. The first one refers to the representativeness of the data. The survey "The future of the media" carried out by Cadem in September 2020 shows that only 18% of the Chilean population uses Twitter on a daily basis [**?**]. This leads to looking for ways to complement the raw information of social networks with demographic information that allows weighing each message according to the weight it would have in reality [**?**, **?**].

The second difficulty is to properly measure the effect of Twitter. The literature indicates that the effects of social media, when they exist, tend to be small. A study on the 2010 Netherlands elections showed that the variables associated with Twitter explain less than 2% of the variance in the number of votes [**?**]. Another study in the context of the United States congressional elections in 2010 quantified that the propaganda on Facebook calling for participation in the elections achieved an increase in electoral participation of between 0.14% and 0.60% [**?**]. The problem is that measuring a small effect on noisy data requires a very large volume of data to achieve meaningful results (the second study mentioned conducted a randomized trial with 61 million users). However, a small effect can make a decisive difference between candidates in highly contested elections.

The third difficulty lies in the external validity of the results obtained. Both the elections and the use of social networks depend on a historical context, so they are not generalizable a priori. Therefore, it is necessary to be rigorous in stating which variables explain that context (for example, the electoral system) and delimiting in which scenarios the results obtained are valid [**?**]. This also reinforces the need for robust and replicable methodologies that are useful regardless of context.

Sampling a social network

The network construction methodology is critical for the subsequent task, which is the measurement of influence. Some sampling algorithms skew the samples towards the nodes with a higher degree (the degree of a node is the number of connections to other nodes). This changes the total node degree distribution, which is directly related to the centrality of a node [**?**]. Therefore, to obtain good measures of centrality, it is a necessary condition that the graph where these measures are calculated is to some extent representative of the complete network.

Obtaining a representative sample of a graph is an open problem in computing, particularly in the case of online social networks (OSN). In OSN we do not have access to the complete graph, so it is not possible to perform a random sampling of nodes or edges, two forms of conventional sampling.

To obtain a sampling of an OSN, the most suitable methods are those of the *Traversal Based Sampling* (TBS) [**?**] category, also known as exploration sampling, which are based on starting at initial nodes from the which to explore and rebuild the network. This is ideal for social networks like Twitter where I can only go from one node to its neighbors. This way of sampling comes from a type of sampling called *Snowball Sampling*, popular in sociology, where you start at certain seed nodes.

The main sampling methods applicable to OSN are presented below. As part of this memory, it is required to choose and implement one of these algorithms to rebuild a network G_s as close to the network of interest G_{CHI}.

- Breadth First Sampling (BFS): también conocido como búsqueda en profundidad, consiste en tomar un nodo aleatorio, explorar todos sus vecinos, y luego para cada vecino repetir recursivamente la exploración hasta tener la cantidad de nodos requerida. Es un método sesgado hacia los nodos con mayor grado [**?**].
- Random Walk (RW): consiste en tomar un nodo semilla y avanzar de manera aleatoria por alguna de sus aristas, y así avanzar hasta encontrar todos los nodos requeridos. Algunas versiones asignan probabilidad $c = 15\%$ de volver al nodo inicial. RW, al igual que BFS, está sesgado hacia nodos con mayor grado, pero existen dos variaciones insesgadas de RW:
  - Re-Weighted Random Walk (RWRW): consiste en realizar muestreo usando Random Walk y posteriormente corregir el sesgo usando el estimador de Hansen-Hurwitz [**?**].
  - Metropolis-Hastings Random Walk (MHRW): consiste en corregir el sesgo al momento de muestrear, decidiendo si se acepta o rechaza determinado nodo candidato. Para esto existe el algoritmo de Metropolis-Hastings, que se basa en modelar el problema de muestreo como una Cadena de Markov Monte Carlo (*Markov Chain Monte Carlo*, MCMC), donde la probabilidad de samplear un nodo depende exclusivamente de la muestra anterior.
- Frontier Sampling (FS): se basa en tener m random walkers avanzando en simultáneo [**?**]. Ha mostrado tener buena exactitud, pese a ser sesgado. Es importante considerar

que, debido al *trade-off* entre sesgo y varianza, flexibilizar la insesgadez puede llevar a métodos con menos varianza y error [**?**].

- Forest Fire Sampling (FFS): en cada paso se determina la cantidad de nodos por los cuales avanzar, de acuerdo a una distribución geométrica. La idea es que al ir explorando los nodos visitados se "quemen", es decir, evitar volver a explorarlos. Este método ha mostrado buenos resultados en redes reales y sintéticas [**?**].
- Coupling From The Past (CFTP): Es un método MCMC desarrollado por James Propp y David Wilson en 1996. La idea es hacer una cadena de Markov converger a la distribución estacionaria deseada (por ejemplo, uniforme) y así obtener muestras "exactas" [**?**]. Se utiliza este método en conjunto con una condición de independencia condicional (*Conditional Independence Coupler*, [**?**]) para implementar un algoritmo que genere estas muestras aleatorias. Este método es interesante debido a su potencial teórico para generar muestras insesgadas y porque es el único de estos métodos que se ha utilizado en el contexto de Twitter y elecciones [**?**].

2.1.1 Chile

The use of Twitter in Chilean electoral contexts has been studied at least since the 2013 presidential elections. Below is a non-exhaustive review of the literature corresponding to the latest electoral processes.

1. Presidential 2013: Sola-Morales & Flores (2015) [**?**] identified that the number of tweets and retweets of a candidate does not correlate with the votes.
2. Municipal 2016: Jara et al. (2017) [**?**], through a clustering of candidates, conclude that the use of Twitter increases the gap between more and less well-known candidates.
3. Primaries 2017: Santander et al (2017) [**?**] constructed a way to predict electoral results using sentiment analysis of tweets.
4. Presidential / parliamentary 2017: Both Rodríguez et al. (2018) [**?**] and Alegre & Keith (2020) [**?**] report good results using sentiment analysis of tweets and machine learning.

The growing interest in the relationship between social networks and political discussion has also meant the opening in recent years of multiple research spaces in Chilean universities. Among the most prominent are the Political and Social Networks Observatory of the Central University [**?**], the Public Space Electronic Demoscopy group (DEEP) of the Catholic University of Valparaíso [**?**] and more recently the Social Listening Lab from the Catholic University of Chile (SoL-UC) [**?**].

2.2 Measuring influence

There are many ways to measure the influence of a user on Twitter. Each of these measures is a different conception of what it means to be influential on Twitter, and the information

available is limited. Riquelme & González-Cantergiani (2016) [**?**] collected and classified more than 70 influence measures used specifically on Twitter. A proper categorization of measures of influence is presented below. Each category is a way of understanding the meaning of influence.

Volumetría

La influencia en Twitter puede entenderse como un problema de volúmenes: cuentas con más actividad (ej: tweets emitidos) o con mayor cantidad de menciones y retweets se considerarían influyentes. Entran en esta categoría lo que Riquelme & González-Cantergiani denominan "métricas de Twitter", como las siguientes:

- Cantidad de tweets originales del usuario
- Cantidad de retweets de tweets originales
- Cantidad de respuestas

2.3 Social Network Analysis

Centralidad

Tabassum et al. (2018) define centralidad o prestigio como "una medida general de la posición de un actor respecto a la estructura completa de una red social" [**?**] . Las medidas de centralidad clásicas consideran la existencia de múltiples actores interactuando entre sí, pero cuya importancia está dada por el cableado o topología de la red social. A continuación se mencionan las principales medidas de centralidad de un nodo u_i:

- Degree: también conocido como grado, es el tamaño del vecindario de un nodo, o equivalentemente, la cantidad de nodos con los que está conectado. En el caso de grafos dirigidos, es posible separar el in-degree (aristas que llegan a u_i) y el out-degree (aristas que salen de u_i).
- Betweenness: expresa el porcentaje de caminos que pasa por u_i, respecto al total de caminos del grafo. Un nodo con alta betweenness es un nodo que tiene una posición estratégica en el flujo de información.
- Closeness: entrega una medida de qué tan cercano es u_i al resto de nodos del grafo.
- Eigenvector centrality: toma el principio de "nodos importantes se conectan a nodos importantes, no necesariamente a más nodos". Se calcula usando los vectores propios (eigenvectors) de la matriz de adjacencia de la red.
- PageRank: desarrollado originalmente por Page et al. de Google [**?**], puede considerarse una variación de la Eigenvector centrality para el caso de grafos dirigidos.

2.4 Sentiment Analysis

Chapter 3

Methods

RQ1 To answer the main research question (RQ1) we proposed (i) Download Twitter data of each candidate (ii) Build features from Twitter data and (iii) Evaluate features using models. The evaluation of

To answer the secondary research questions, modifications were made to the input of the models in the following way:

- RQ1-1: If using aggregated data (e.g., party-level data), then metrics
- RQ1-2:
- RQ1-3

We will go deeper on what "fits better" means in validation section

3.1 Data collection

Data collection was performed in two steps: first, we downloaded data of candidates for the Constitutional Convention, including electoral data and social media usernames. Secondly, using the obtained Twitter accounts, we download tweets related to them, and used those tweets to build influence features.

3.1.1 Candidates data

The electoral data was obtained from the Electoral Service of Chile (SERVEL). It was released the day after the election and contains information about the candidate and the outcome of election as raw votes and electoral district percentage for each of the 1278 candidates for the Constitutional Convention.

The social media usernames of candidates were obtained from three open websites that gathered information about the candidates and made it public:

- ¿Quiénes son? [*Who they are?*] (`quienesson.cl`)
- Interactivo La Tercera (`interactivo.latercera.com/candidatos-constituyentes`)
- Conoce Tu Candidato - 24 Horas [*Know your candidate*] (`conocetucandidato.24horas.cl`)

Through this web scraping we were able to extract the Twitter username of 832 candidates. All variables collected in this step are shown in Table 3.1.

Table 3.1: List of 14 features of candidates for the Constitutional Convention

Source	Feature	Type
SERVEL	Electoral district	Categorical
	Lista	Categorical
	Party	Categorical
	Order in ballot	Integer
	Order in party list	Integer
	Name	String
	Gender	Categorical
	Votes	Integer
	District votes percentage.	Float
	Elected	Categorical
Web scraping	Age	Integer
	Occupation	String
	Twitter username	String

3.1.2 Twitter data

From 832 Twitter users obtained, only 771 where valid existing non private accounts at the moment of the extraction. Those 771 were consider as the base universe of study of this thesis. Anexo 1 has a summary of the distribution of this sample of 771 candidates.

As stayed in Specific objectives, the data downloaded from Twitter must allow to obtain measures of influence and to rebuild the social network.

The chosen sampling schema is a modified Breadth-first search (BFS) algorithm

1. Download the timeline of all 771 accounts between 01/01/2021 and 14/05/20
2. For each tweet made by a candidate account, download all the . This is similar to the first level of BFS, capturing part of the indegree of the candidates.

3. For each user author of a (i.e. users that made a retweet of a candidate), download the timeline between 01/01/2021 and 14/05/20.

These three steps had limitations by Twitter API.

Because of Twitter API limitations, Table X shows the amount of tweets downloaded in each step. Anexo Y shows the representativity of the retweet sample.

The sampling

3.2 Feature engineering

The Y tweets downloaded in the previous step were processed in order to create features that represented the influence of the users in the platform. These features were divided in three major categories: Twitter metrics, sentiment analysis and network analysis features.

3.2.1 Twitter metrics

3.2.2 Sentiment analysis

3.2.3 Network analysis

Anexo Y shows the detail of all influence features computed

3.3 Validation

The univariate correlation will help us ask which features are closer to the election result. In the other hand, regression analysis allow us to check the gain of information that the twitter influence metrics make comparing to no-twitter features, and measure this gain/loss in terms of percentual variation of R2 and MAE.

3.3.1 Correlation

Spearman correlation

todo lo que correlaciona según pearson correlaciona según spearman pero no viceversa (superconjunto)

3.3.2 Regression

We performed regression with the district percentage as target variable

WE use two models of regression: linear regression and Random Forest Regression

The goals of these models is to compare the twitter influence features to a baseline in terms of R^2 and out-of-sample MAE.

Adjusted R^2

Mean Absolute Error (MAE)

Definition 3.1 (ver [1]) *Definición definitiva*

$$\frac{d}{dx}\int_a^x f(y)dy = f(x).$$

Chapter 4

Results

Chapter 5

Discussion

5.1 Description vs. Explanation vs. Prediction

5.2 Conclusions

Bibliography

[1] Ioannis Karatzas. and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, Berlin, 2nd edition, 2000.

[2] Philip Protter. *Stochastic Integration and Differential Equations*. Springer, 1990.

[3] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*. Number 293 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin [u.a.], 3. ed edition, 1999.

Appendix A

Anexo

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.

Maecenas dui. Aliquam volutpat auctor lorem. Cras placerat est vitae lectus. Curabitur massa lectus, rutrum euismod, dignissim ut, dapibus a, odio. Ut eros erat, vulputate ut, interdum non, porta eu, erat. Cras fermentum, felis in porta congue, velit leo facilisis odio, vitae consectetuer lorem quam vitae orci. Sed ultrices, pede eu placerat auctor, ante ligula rutrum tellus, vel posuere nibh lacus nec nibh. Maecenas laoreet dolor at enim. Donec molestie dolor nec metus. Vestibulum libero. Sed quis erat. Sed tristique. Duis pede leo, fermentum quis, consectetuer eget, vulputate sit amet, erat.

Donec vitae velit. Suspendisse porta fermentum mauris. Ut vel nunc non mauris pharetra varius. Duis consequat libero quis urna. Maecenas at ante. Vivamus varius, wisi sed egestas tristique, odio wisi luctus nulla, lobortis dictum dolor ligula in lacus. Vivamus aliquam, urna sed interdum porttitor, metus orci interdum odio, sit amet euismod lectus felis et leo. Praesent ac wisi. Nam suscipit vestibulum sem. Praesent eu ipsum vitae pede cursus venenatis. Duis sed odio. Vestibulum eleifend. Nulla ut massa. Proin rutrum mattis sapien. Curabitur dictum gravida ante.

Phasellus placerat vulputate quam. Maecenas at tellus. Pellentesque neque diam, dignissim ac, venenatis vitae, consequat ut, lacus. Nam nibh. Vestibulum fringilla arcu mollis arcu. Sed et turpis. Donec sem tellus, volutpat et, varius eu, commodo sed, lectus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Quisque enim arcu, suscipit nec, tempus at, imperdiet vel, metus. Morbi volutpat purus at erat. Donec dignissim, sem id semper tempus, nibh massa eleifend turpis, sed pellentesque wisi purus sed libero. Nullam lobortis tortor vel risus. Pellentesque consequat nulla eu tellus. Donec velit. Aliquam fermentum, wisi ac rhoncus iaculis, tellus nunc malesuada orci, quis volutpat dui magna id mi. Nunc vel ante. Duis vitae lacus. Cras nec ipsum.

Morbi nunc. Aliquam consectetuer varius nulla. Phasellus eros. Cras dapibus porttitor risus. Maecenas ultrices mi sed diam. Praesent gravida velit at elit vehicula porttitor. Phasellus nisl mi, sagittis ac, pulvinar id, gravida sit amet, erat. Vestibulum est. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Curabitur id sem elementum leo rutrum hendrerit. Ut at mi. Donec tincidunt faucibus massa. Sed turpis quam, sollicitudin a, hendrerit eget, pretium ut, nisl. Duis hendrerit ligula. Nunc pulvinar congue urna.

Nunc velit. Nullam elit sapien, eleifend eu, commodo nec, semper sit amet, elit. Nulla lectus risus, condimentum ut, laoreet eget, viverra nec, odio. Proin lobortis. Curabitur dictum arcu vel wisi. Cras id nulla venenatis tortor congue ultrices. Pellentesque eget pede. Sed eleifend sagittis elit. Nam sed tellus sit amet lectus ullamcorper tristique. Mauris enim sem, tristique eu, accumsan at, scelerisque vulputate, neque. Quisque lacus. Donec et ipsum sit amet elit nonummy aliquet. Sed viverra nisl at sem. Nam diam. Mauris ut dolor. Curabitur ornare tortor cursus velit.

Morbi tincidunt posuere arcu. Cras venenatis est vitae dolor. Vivamus scelerisque semper mi. Donec ipsum arcu, consequat scelerisque, viverra id, dictum at, metus. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut pede sem, tempus ut, porttitor bibendum, molestie eu, elit. Suspendisse potenti. Sed id lectus sit amet purus faucibus vehicula. Praesent sed sem non dui pharetra interdum. Nam viverra ultrices magna.

Aenean laoreet aliquam orci. Nunc interdum elementum urna. Quisque erat. Nullam tempor neque. Maecenas velit nibh, scelerisque a, consequat ut, viverra in, enim. Duis magna. Donec odio neque, tristique et, tincidunt eu, rhoncus ac, nunc. Mauris malesuada malesuada elit. Etiam lacus mauris, pretium vel, blandit in, ultricies id, libero. Phasellus bibendum erat ut diam. In congue imperdiet lectus.

Aenean scelerisque. Fusce pretium porttitor lorem. In hac habitasse platea dictumst. Nulla sit amet nisl at sapien egestas pretium. Nunc non tellus. Vivamus aliquet. Nam adipiscing euismod dolor. Aliquam erat volutpat. Nulla ut ipsum. Quisque tincidunt auctor augue. Nunc imperdiet ipsum eget elit. Aliquam quam leo, consectetuer non, ornare sit amet, tristique quis, felis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque interdum quam sit amet mi. Pellentesque mauris dui, dictum a, adipiscing ac, fermentum sit amet, lorem.

Ut quis wisi. Praesent quis massa. Vivamus egestas risus eget lacus. Nunc tincidunt, risus quis bibendum facilisis, lorem purus rutrum neque, nec porta tortor urna quis orci. Aenean aliquet, libero semper volutpat luctus, pede erat lacinia augue, quis rutrum sem ipsum sit amet pede. Vestibulum aliquet, nibh sed iaculis sagittis, odio dolor blandit augue, eget mollis urna tellus id tellus. Aenean aliquet aliquam nunc. Nulla ultricies justo eget orci. Phasellus tristique fermentum leo. Sed massa metus, sagittis ut, semper ut, pharetra vel, erat. Aliquam quam turpis, egestas vel, elementum in, egestas sit amet, lorem. Duis convallis, wisi sit amet mollis molestie, libero mauris porta dui, vitae aliquam arcu turpis ac sem. Aliquam aliquet dapibus metus.

Vivamus commodo eros eleifend dui. Vestibulum in leo eu erat tristique mattis. Cras at elit. Cras pellentesque. Nullam id lacus sit amet libero aliquet hendrerit. Proin placerat, mi non elementum laoreet, eros elit tincidunt magna, a rhoncus sem arcu id odio. Nulla

eget leo a leo egestas facilisis. Curabitur quis velit. Phasellus aliquam, tortor nec ornare rhoncus, purus urna posuere velit, et commodo risus tellus quis tellus. Vivamus leo turpis, tempus sit amet, tristique vitae, laoreet quis, odio. Proin scelerisque bibendum ipsum. Etiam nisl. Praesent vel dolor. Pellentesque vel magna. Curabitur urna. Vivamus congue urna in velit. Etiam ullamcorper elementum dui. Praesent non urna. Sed placerat quam non mi. Pellentesque diam magna, ultricies eget, ultrices placerat, adipiscing rutrum, sem.