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

MEASURING THE INFLUENCE OF CANDIDATES TO THE CONSTITUTIONAL CONVENTION IN TWITTER

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

JOSÉ MIGUEL CORDERO CARVACHO

PROFESOR GUÍA:
ANDRÉS ABELIUK KIMELMAN

MIEMBROS DE LA COMISIÓN:
BENJAMÍN BUSTOS CÁRDENAS
MARÍA CECILIA RIVARA ZÚÑIGA

SANTIAGO DE CHILE
2022

Midiendo la influencia de los candidatos a la Convención Constitucional en Twitter

Las redes sociales almacenan trazas digitales de los humanos y tienen el potencial de explicar fenómenos sociales e inferir cantidades de interés. Un uso recurrente de la red social Twitter es para la predicción de resultados electorales, dado la extensión de su uso para comunicación política. Esta memoria de título busca comprobar que existe una relación entre la influencia de un candidato en Twitter y su resultado electoral. Las preguntas de investigación fueron ¿En qué medida la influencia en Twitter se correlaciona con los votos obtenidos? ¿De qué variables depende esta relación? (RQ1) y ¿La afiliación política de un candidato (pertenecer a un partido político, coalición política o lista electoral) es relevante para la influencia en Twitter? ¿Cómo interactúa la afiliación con Twitter? (RQ2).

Para responder estas preguntas, se construyó una bases de datos de Twitter usando las cuentas de 771 candidatos a la Convención Constitucional elegida en 2021 en Chile. A partir de estos datos, se construyó una red de retweets y diversas variables de influencia a ser evaluadas.

Para responder la pregunta RQ1 se evaluó cada variable usando coeficientes de correlación y modelos de regresión y clasificación respecto al porcentaje de votos obtenido. Los mejores modelos de regresión fueron los que incluyeron las variables cantidad de favoritos (*likes*) por tweet ($R^2 = 0.582$) y cantidad de retweets por tweet ($R^2 = 0.573$). Tanto para regresión como clasificación, los modelos con variables de Twitter superaron el baseline de variables político demográficas. Para responder la pregunta RQ2 se modificó el algoritmo del PageRank para incluir información de afiliación (lista, partido o coalición) y se evaluaron esas variables modificadas respecto al desempeño del PageRank original.

Concluimos que (i) las características calculadas desde Twitter añaden información relevante para inferir el resultado de esta elección multipartido, aunque de forma acotada (ii) Existe una brecha entre mujeres y hombres y entre capital/regiones en el efecto que tienen, detectada a través de variables de interacción en la regresión lineal. La relación entre las variables de influencia en Twitter y los votos es más fuerte en el caso de hombres que de mujeres, y en la Región Metropolitana que en otras regiones (iii) El PageRank modificado no mejoró el desempeño de las tareas de clasificación, lo que podría indicar que la información inyectada ya está contenida en el grafo de retweets (iv) Varias limitaciones deben superarse para obtener modelos más robustos y estables. Se proponen caminos a seguir y problemas a resolver para construir modelos que puedan explicar y predecir resultados electorales.

Measuring the influence of candidates to the Constitutional Convention in Twitter

Social networks store digital traces of humans and have the potential to explain social phenomena and infer quantities of interest. A recurring use of the social network Twitter is to predict election results, given the extent of its use for political communication. This thesis seeks to verify that there is a relationship between the influence of a candidate on Twitter and her electoral result. The research questions were: To what extent is influence on Twitter correlated with votes obtained? What variables does this relationship depend on? (RQ1) y Is the political affiliation of a candidate (belonging to a political party, political coalition or electoral list) relevant to influence on Twitter? How does affiliation interact with Twitter? (RQ2).

To answer these questions, a Twitter database was built using the accounts of 771 candidates for the Constitutional Convention elected in 2021 in Chile. From these data, a network of retweets and various influence variables to be evaluated were built.

To answer question RQ1, each variable was evaluated using correlation coefficients and regression and classification models with respect to the percentage of votes obtained. The best regression models were those that included the variables number of favorites (*likes*) per tweet ($R^2 = 0.582$) and number of retweets per tweet ($R^2 = 0.573$). For both regression and classification, the models with Twitter variables exceeded the baseline of political demographic variables. To answer question RQ2, the PageRank algorithm was modified to include affiliation information (list, party, or coalition) and these modified variables were evaluated against the performance of the original PageRank.

We conclude that (i) the features computed from Twitter add relevant information to infer the result of this multiparty election, although in a limited way (ii) There is a gap between women and men and between capital/regions in the effect they have, detected through of interaction variables in linear regression. The relationship between the variables of influence on Twitter and votes is stronger in the case of men than women, and in the Metropolitan Region than in other regions (iii) The modified PageRank did not improve the performance of the classification tasks, which could indicate that the injected information is already contained in the retweets network (iv) Several limitations must be overcome to obtain more robust and stable models. Paths to follow and problems to solve are proposed to build models that can explain and predict electoral results.

"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

As it is impossible to thank all the people who deserve it, I will be brief.

I thank my parents Ana María and Mauricio and my sisters Francisca and Javiera for their support and care in this long process.

I thank my friends Doris, Vicente, Victor, Alfredo and many others I've walked with and they gave me a hand.

I thank the University of Chile and their workers for making my stay possible. I thank the DCC professors for their commitment and human quality. I thank specially my advisor Andrés Abeliuk for his support and for letting me keep on with this research.

Finally, I thank the citizens of Chile as the historical subject of the transformations that we are experiencing and that motivate this thesis.

Table of Content

1 Introduction **1**

1.1 Motivation 1

1.2 Research questions 2

1.3 Objectives 2

1.3.1 General objective 2

1.3.2 Specific objectives 2

2 Background and Related work **3**

2.1 Twitter and elections 3

2.1.1 Chile 5

2.2 Measuring influence 5

2.2.1 Volumetry 6

2.2.2 Centrality in Social Network Analysis 6

3 Methods **7**

3.1 Data collection 7

3.1.1 Candidates data 7

3.1.2 Twitter data 8

3.2 Feature engineering 9

3.2.1 Twitter metrics 10

3.2.2 Sentiment analysis 12

3.2.3 Network analysis 13

3.2.4 Custom political PageRank 14
3.3 Validation 17
3.3.1 Preprocessing 17
3.3.2 Base features 19
3.3.3 Correlation 20
3.3.4 Regression 21
3.3.5 Classification 23
3.3.6 Political PageRank 28

4 Results **29**
4.1 RQ1: Twitter influence features assessment 29
4.1.1 Descriptive analysis 29
4.1.2 Correlation 32
4.1.3 Regression 34
4.1.4 Classification 38
4.2 RQ2: Modified PageRank 45

5 Conclusions **52**
5.1 Description, Explanation or Prediction? 53

Bibliography **55**

Annexed **59**

List of Tables

3.1 List of 14 features obtained from web sources of candidates for the Constitutional Convention 8

3.2 Twitter metrics for a candidate u_i 11

3.3 Sentiment Analysis features for a candidate u_i 13

3.4 Network Analysis features for a candidate u_i 14

3.5 Preprocessing of Twitter influence features 19

3.6 Base features for a candidate u_i 20

4.1 Spearman correlation between Twitter features and votes percentage of candidates. 32

List of Figures

3.1 Leave-one-district-out Cross Validation 24

3.2 Definition of Precision and Recall 25

3.3 Example of Precision Recall Curves 26

3.4 Example of ΔAP 27

4.1 Comparison between the vote percentage distribution of candidates by Twitter account owning 29

4.2 Correlation between raw Twitter influence features 30

4.3 Correlation between raw Network analysis features 31

4.4 Spearman correlation between Twitter features and votes with different pre-processing 33

4.5 Standardized regression coefficient for Twitter influence features 34

4.6 ΔR^2 for Twitter influence features 35

4.7 Interactions terms 36

4.8 Interaction terms interpretation examples 37

4.9 Regression analysis of the best linear model (base features + favorite count feature) 38

4.10 PR-AUC score for different threshold quantiles and features 39

4.11 Precision Recall curves for different class thresholds quantiles q 40

4.12 ΔAP score for each Twitter influence feature and threshold quantiles 0.5 and 0.9 42

4.13 Average of ΔAP of all quantiles for each Twitter influence feature 43

4.14 Regression performance vs. Classifier performance for each Twitter feature . 44

4.15 Distribution of Custom political PageRanks for selected γ values 45
4.16 Correlation between PageRank and Custom political PageRanks 46
4.17 Correlation between electoral outcome and Custom political PageRanks . . . 47
4.18 ΔR^2 for each Custom political PageRank 48
4.19 Classifier performance of each Custom political PageRank feature 49
4.20 Comparison of raw PageRank and Political Party PageRank 51

Chapter 1

Introduction

1.1 Motivation

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

The growing digital literacy in Chile, accelerated partly by the COVID-19 pandemic, suggests that social networks played a crucial 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 used by parliamentarians, ministers and candidates. However, Twitter is known as an unrepresentative network, dominated by "a young adult population from well-off sectors" [6], but with great capacity to install a media agenda.

Analyzing an election on Twitter from the voter's point of view can be complex: it requires a significant and representative amount of conversation to identify positions for candidates or pacts. However, Twitter can also be studied from the point of view of the candidates. Twitter reflects how different political campaign operate. Candidates express positions; those positions are disseminated (retweeted) and validated by others,

An electoral campaign's objective is *to influence* the behavior of voters. Therefore, the *influence* on Twitter is, for a candidate, an indicator of a successful campaign. A successful campaign mobilizes more votes. Thus, the hypothesis that a relationship exists between influence on social networks and the number of votes obtained is reasonable. This thesis aims to measure 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?
- RQ2: Is a candidate's political affiliation (belonging to a political party, political coalition, or electoral list) relevant for the Twitter influence? How does affiliation interact with Twitter?

1.3 Objectives

1.3.1 General objective

Choose and compute measures of the 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 social network with information downloaded from Twitter where the candidates are included. This network construction must try to be as representative as possible of the existing 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 using several metrics.

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 looked at with data from Twitter. Next, we will review the main methodological difficulties when carrying out these analyzes, 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 in quantifying the effect of social networks on elections. The first one refers to the representativeness of the data. For example, the survey "The future of the media" carried out by Cadem in September 2020 shows that only 18% of the Chilean population uses Twitter daily [5]. This leads to looking for ways to complement the basic information on social networks with demographic information that allows weighing each message according to the weight it would have in reality [15, 43].

The second difficulty is to measure the effect of Twitter correctly. The literature indicates that the effects of social media when they exist tend to be small. For example, 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 [23]. 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% [4]. The problem is that measuring a small effect on noisy data requires enormous 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. The elections and 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 scenarios the results obtained are valid [37]. This also reinforces the need for robust and replicable methodologies.

Sampling a social network

The network construction methodology is critical for the subsequent task of measuring 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 [18]. Therefore, to obtain reasonable measures of centrality, it is necessary that the graph where these measures are calculated is to some extent representative of the complete network.

Getting 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 performing a random sampling of nodes or edges is impossible, two forms of conventional sampling.

To obtain a sampling of an OSN, the most suitable methods are those of the *Traversal Based Sampling* (TBS) [18] category, also known as exploration sampling, which is based on starting at initial nodes from which to explore and rebuild the network. This is ideal for social networks like Twitter where from one node we can only explore its neighbors. This way of sampling comes from a type of sampling called *Snowball Sampling*, popular in sociology, where you start at specific seed nodes.

The primary sampling methods applicable to OSN are presented below. As part of this thesis, choosing and implementing one of these algorithms is required to rebuild a network as close to the network of interest.

- Breadth First Sampling (BFS) consists of taking a random node, exploring all its neighbors, and then recursively repeating the exploration for each neighbor until the required number of nodes is found. It is a biased method towards nodes with higher degree [24].
- Random Walk (RW) consists of taking a seed node and advancing randomly along with one of its edges and thus advancing until all the required nodes are found. Some versions assign probability $c = 15\%$ of returning to the initial node. RW, like BFS, is biased to high degree nodes, but there are two unbiased variations of RW:
  - Re-Weighted Random Walk (RWRW) consists of sampling using the Random Walk and then correcting the bias using the Hansen-Hurwitz estimator [18].
  - Metropolis-Hastings Random Walk (MHRW) consists of correcting the bias at the sampling time and deciding whether to accept or reject a certain candidate node. The Metropolis-Hastings algorithm is used, which is based on modeling the sampling problem as a Markov Chain Monte Carlo (MCMC), where the probability of sampling a node depends exclusively on the previous sample [18].
- Frontier Sampling (FS) is based on having m random walkers moving forward simultaneously [34]. It has shown good accuracy, despite being biased. It is important to consider that, due to the trade-off between bias and variance, relaxing unbiasedness can lead to methods with less variance and error [40].

- Forest Fire Sampling (FFS): The number of nodes through which to advance is determined according to a geometric distribution in each step. As As the visited nodes are explored, they "burn" to avoid exploring them again. This method has shown promising results in real and synthetic networks. [25].
- Coupling From The Past (CFTP) is a MCMC method developed by James Propp and David Wilson in 1996. The idea is to make a Markov chain converge to the desired stationary distribution (e.g. uniform) and thus obtain "exact" samples [31]. This method is used in conjunction with a conditional independence condition (*Conditional Independence Coupler*, [26]) to implement an algorithm that generates these random samples. This method is interesting because of its theoretical potential to generate unbiased samples and because it has been used in the context of Twitter and elections [12].

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) [39] identified that the number of tweets and retweets of a candidate does not correlate with the votes.
2. Municipal 2016: Jara et al. (2017) [19], 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) [38] constructed a way to predict electoral results using sentiment analysis of tweets.
4. Presidential / parliamentary 2017: Both Rodríguez et al. (2018) [36] and Alegre & Keith (2020) [2] 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 [42], 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) [35] collected and classified more

than 70 influence measures used on Twitter. Proper categorization of measures of influence is presented below. Each category is a way of understanding the meaning of influence.

2.2.1 Volumetry

Influence on Twitter can be understood as a problem of volume: a user may be considered influential with more activity (i.e., tweets made) or with higher mentions and retweets. In this category are those features denominated by Riquelme & González-Cantergiani as a Twitter metric, for example:

- Number of tweets made by a user
- Number of retweets of user tweets
- Number of replies

2.2.2 Centrality in Social Network Analysis

Tabassum et al. (2018) defines centrality or prestige as "a general measure of the position of an actor with respect to the entire structure of a social network" [41] . The classic centrality measures consider the existence of multiple actors interacting with each other, but whose importance is given by the wiring or topology of the social network. The main measures of centrality of a node u_i are mentioned below:

- Degree: is the size of a node's neighborhood, or equivalently, the number of nodes it is connected to. In the case of directed graphs, it is possible to separate the in-degree (edges that reach u_i) and the out-degree (edges that leave u_i).
- Betweenness: expresses the percentage of paths that pass through u_i, concerning the full paths of the graph. A node with high betweenness is a node that has a strategic position in the information flow.
- Closeness: returns a measure of how close u_i is to the rest of the nodes in the graph.
- Eigenvector centrality: it takes the principle of "important nodes connect to important nodes, not necessarily more nodes". It is calculated using the eigenvectors of the adjacency matrix of the network.
- PageRank: originally developed by Page et al. from Google [29], can be considered a variation of Eigenvector centrality for the case of directed graphs.

Chapter 3

Methods

To answer the first research question (RQ1) we propose (i) downloading the Twitter data of each candidate, (ii) build influence features from Twitter data and (iii) Evaluating those features using models. The evaluation of the features is detailed in validation section of this chapter.

To answer the second research question (RQ2), we proposed a modified version of the PageRank algorithm, then evaluated the modified feature to the original PageRank feature built for RQ1. The idea is to inject the political affiliation information into the algorithm, then assess if this new information leads to a better influence metric to estimate the electoral outcome of each candidate.

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, we downloaded tweets related to them using the acquired Twitter accounts 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 contained information about the candidate and the election outcome 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 extracted 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 obtained from web sources of candidates for the Constitutional Convention

Source	Feature	Type
SERVEL	Electoral district	Categorical
	Electoral list	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, only 771 users were valid existing non-private accounts at the moment of the extraction. Those 771 were considered as this thesis's universe of study.

As stated in the specific objectives, the data downloaded from Twitter should allow us to obtain influence measures and rebuild the social network. Therefore, the chosen sampling schema is a modified Breadth-first search (BFS) algorithm with the following steps:

1. Download the timeline (i.e. the tweets posted) of all 771 accounts between 01/01/2021 and 14/05/2021.
2. For each tweet made by a candidate account, download all the retweets. This is similar to the first level of BFS, capturing part of the indegree of the candidates.
3. For each user who is the author of a retweet (users that made a retweet of a candidate), download their timeline between 01/01/2021 and 14/05/2021.

The described algorithm is shown as pseudocode in Algorithm 1.

Algorithm 1 Modified Breadth-first search for tweet scrapping

Require: TWEETS(u_i, start, end): retrieve tweets from Twitter user u_i created between time range start and end
Require: RETWEETS(t_j): retrieve retweets from tweet t_j
Require: USER(t_i): get user of tweet t_j
Ensure: $U_{candidates} \leftarrow$ Twitter users of candidates
  for $u_i \in U_{candidates}$ **do** ▷ u_i is a candidate Twitter user
    $T_i \leftarrow$ TWEETS(u_i, 01/01/2021, 14/05/2021)
    for $t_j \in T_i$ **do** ▷ t_j is a tweet
      $RT_j \leftarrow$ RETWEETS(t_j)
      for $t_k \in RT_j$ **do** ▷ t_k is a tweet
        $u_{rt} \leftarrow$ USER(t_k)
        $RTT_{jk} \leftarrow$ TWEETS(u_{rt}, 01/01/2021, 14/05/2021)
      end for
    end for
  end for

Twitter API limitations

In Algorithm 1, we defined functions TWEETS, RETWEETS and USER to collect data from the Twitter world. To implement this download, access to Twitter API was needed. Twitter offers a REST API to obtain its data through HTTP requests. Not every Twitter data is reachable, and defined policies exist to access it.

Using a Python library wrapper of Twitter API called `tweepy`, we implemented the download and found the following limitations:

- For implementation of TWEETS we used the `GET statuses/user_timeline` Twitter API endpoint, wrapped in the `user_timeline` method of `tweepy`. This endpoint is limited only to 3.200 most recent tweets per account.
- For implementation of RETWEETS we used the `GET statuses/retweets/:id` Twitter endpoint, wrapped in the `retweets` method of `tweepy`. This endpoint is limited to the 100 most recent retweets for each tweet.

3.2 Feature engineering

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

3.2.1 Twitter metrics

Tweets contain metadata related to creator user, location, datetime and other features captured by Twitter. Appendix has an example of a complete tweet object. We built several features aggregating these metadata for each candidate. The goal is to obtain features that capture the activity of the candidates on the platform, e.g., how many tweets they posted, how popular those tweets were, how many videos they posted.

A full list of these created features can be found in Table 3.2. The `tweets_made` feature was computed as a per-week average, i.e., count of tweets made by a candidate divided by number of weeks. All other features were computed as a per-tweet average, i.e., a ratio of a total value related to the total number of tweets made. For example, a per-tweet video average of 0.1 means that, for that candidate, 10% of tweets contains videos.

Table 3.2: Twitter metrics for a candidate u_i

Feature name	Description	Formula
tweets_made	Activity of user in Twitter, measured as the average of tweets made per week.	$\frac{\#\text{ Tweets of user } u_i}{18}$
retweet_count	Average retweets received per tweet	$\frac{\#\text{ Retweets received by user } u_i}{\#\text{ Tweets of user } u_i}$
favorite_count	Average favorites (likes) received per tweet	$\frac{\#\text{ Favorites received by user } u_i}{\#\text{ Tweets of user } u_i}$
user_mentions	Average number of mentions in tweets per tweet. A mention is naming a username (like example) in a tweet, a tweet may have from zero to multiple mentions.	$\frac{\#\text{ Users menctioned by user } u_i}{\#\text{ Tweets of user } u_i}$
photos	Average photos uploaded per tweet	$\frac{\#\text{ Retweets received by user } u_i}{\#\text{ Tweets of user } u_i}$
retweets_made	Average retweets made per tweet. Is equivalent to the percentage of retweets related to total tweets made (retweets counts as regular tweets)	$\frac{\#\text{ Retweets received by user } u_i}{\#\text{ Tweets of user } u_i}$
replies_made	Average replies made per tweet.	$\frac{\#\text{ Replies of user } u_i}{\#\text{ Tweets of user } u_i}$
quotes_made	Average quotes made per tweet. A quote is similar to a retweet, but it allows the user to add a comment to the referenced tweet.	$\frac{\#\text{ Quotes of user } u_i}{\#\text{ Tweets of user } u_i}$
videos	Average videos uploaded per tweet	$\frac{\#\text{ Videos uploaded by user } u_i}{\#\text{ Tweets of user } u_i}$
hashtags	Average hashtags used per tweet	$\frac{\#\text{ Hashtags used by user } u_i}{\#\text{ Tweets of user } u_i}$

3.2.2 Sentiment analysis

The sentiment of the tweets related to candidates has been used as a predictor in several works on electoral forecasting [38] [2] [36]. Therefore, it was interesting to understand the relationship between sentiments and their electoral outcome.

The mentioned studies select a sample of tweets that are user opinions in the social network. This sample may be representative enough of the population, and with text classification techniques, the average sentiment of the sample can be inferred and the number of votes can be predicted.

Our sampling strategy does not work with that kind of method, because BFS sampling is biased to the seed nodes, so cannot be used as a representative sample of the voting population. So, rather than inferring some public opinion trends with social media sentiment analysis as a proxy, we are assessing the political communication strategy of each candidate in terms of the positive or negative (or other sentiments) tweets they chose to post.

To build these candidate sentiment features we defined (i) a way to attach each tweet in our database to one or multiple sentiments, at least considering the categories positive and negative, and (ii) an aggregation transformation that maps from the labeled tweets to a real number for each candidate.

To accomplish task (i), sentiment analysis is an advanced field in Natural Language Processing. We researched the state of the art in terms of text sentiment classification and selected a neural network model named BERT (Bidirectional Encoder Representations from Transformers) [13]. We used a Python library called `pysentimiento` [32].

The output of the BERT classifier is the outcome of a softmax function and represents the probability of the observation belong to each class. This library has two different types of classifiers: first, a sentiment classifier with classes positive, negative and neutral. Second is an emotion classifier with classes anger, surprise, fear, disgust, joy, sadness and others. Both classifiers contain pre-trained weights from a model called BETO [7], which is trained in a massive Spanish corpus.

We applied both classifiers and obtained a probability for each candidate tweet, excluding retweets, for all the classes mentioned above. We chose a threshold of 90%, so if the probability for some class was higher than 90%, we labeled the tweet with that class.

For task (ii), we defined the proposed sentiment features of a candidate as the percentage of tweet labeled as a sentiment. The proposed sentiment features can be found at Table 3.3. Classes neutral and others were not included because they were not interpretable. Fear and surprise classes were not included because they were absent in some districts. The selected classes are present in every electoral district for at least a single candidate.

Table 3.3: Sentiment Analysis features for a candidate u_i

Feature name	Description	Formula
pos	Percentage of tweets classified as positive, according to BERT sentiment classifier	$\frac{\#\text{ Positive tweets made by } u_i}{\#\text{ Tweets of user } u_i}$
neg	Percentage of tweets classified as negative, according to BERT sentiment classifier	$\frac{\#\text{ Negative tweets made by } u_i}{\#\text{ Tweets of user } u_i}$
sadness	Percentage of tweets classified as sad, according to BERT emotion classifier	$\frac{\#\text{ Sad tweets made by } u_i}{\#\text{ Tweets of user } u_i}$
anger	Percentage of tweets classified as angry, according to BERT emotion classifier	$\frac{\#\text{ Angry tweets made by } u_i}{\#\text{ Tweets of user } u_i}$
joy	Percentage of tweets classified as joy, according to BERT emotion classifier	$\frac{\#\text{ Joy tweets made by } u_i}{\#\text{ Tweets of user } u_i}$

3.2.3 Network analysis

The Twitter metrics and the Sentiment analysis features are computed for each candidate independently. This is, the feature engineering required to compute it can be done with the data of a single candidate u_i. However, network centrality metrics, as introduced in Section 2.2.2, cannot be computed for a single candidate. Instead, they must be computed for multiple candidates linked in a graph.

Using social network analysis algorithms, we can take advantage of the Twitter users' links in terms of retweets. This information is not contained in the Twitter metrics or the Sentiment analysis features.

To compute the network features, we need to build a network. The network was constructed in the following way:

1. We selected a time window of seven days between dates a and b.
2. Using the collected tweets created between dates a, b, we built a directed graph in which each Twitter user in the database becomes a node. a points to $b \iff a$ retweeted b. We called this graph as $G_{a,b}$.

3. Several network centrality metrics were computed from graph $G_{a,b}$

The result of this process is a time series of each centrality measure. To aggregate this vector and obtain a single value for each candidate, we computed the average of the candidate time series for each measure. The chosen network centrality metrics algorithms to compute are contained in Table 3.4.

For example, for a candidate u_i the value of degree feature $d(u_i)$ is the mean of the degrees $d_t(u_i)$ for t going from day 07/01/2021 to day 14/05/2021, the last day. Likewise, the degree value for a day t is the degree of the graph of retweets between days $t-6$ and t.

Table 3.4: Network Analysis features for a candidate u_i

Feature name	Description
degree	Number of edges linked to node u_i
out_degree	Number of nodes pointed by u_i
in_degree	Number of nodes that points to u_i
eigenvector_centrality	$p = A'p$, with A' the trasposed adjacency matrix of the graph and p the eigenvector
harmonic_centrality	Sum of the reciprocal of the shortest path distances from all other nodes to u_i
pagerank	eigenvector centrality with random jump between nodes

3.2.4 Custom political PageRank

To answer RQ2, we proposed the hypothesis that retweeting a candidate of a party helps the electoral result for that candidate and all party candidates. The logic behind this hypothesis is that there is a relationship between the ideas of candidates of the same party, so if one of them has many retweets, all may get benefit of that exposure. The same hypothesis can be raised with electoral list or political coalitions instead of parties.

To test this hypothesis, we proposed a modified version of PageRank centrality measure to inject the political information needed to complete the missing links and consider the spillover that is not contained in the Twitter data. Then, we tested if the modification increased the performance metrics for classification, relative to the base, not modified PageRank. If the modified PageRank achieves better performance metrics than original PageRank, this would mean that the party information added to the graph (or the electoral list, or the political coalition) is meaningful information to infer electoral outcomes.

How do we inject the affiliation (party/list/political coalition) links into the graph? First, let's look deeper into the PageRank calculation.

PageRank as Markov chain

PageRank algorithm's goals is to rank the web pages in terms of the links between them. It assumes there is a random surfer that jump from page to page with some probability. Modelling the problem as a stochastic process, the importance of each webpage is measured as the probability of being in that page. The modelling is based on Markov chains, where each node represents a state, and the directed edges represent state transition. In the case of our graph, those retweet edges represent the information flow.

Definition 3.1 ([20]) *Markov Matrix*

*A **Markov matrix** (or **stochastic matrix**) is a square matrix M whose columns are probability vectors (i.e., non-negative and sum to 1)*

Definition 3.2 ([20]) *Markov Chain*

*A **Markov chain** is a sequence of probability vectors $\vec{x_0}$, $\vec{x_1}$, $\vec{x_2}$, ... such that $\vec{x_{k+1}} = M\vec{x_k}$ for some Markov matrix M*

As defined in Definition 3.2, a Markov Chain has two components: a stochastic matrix and probability vectors $\vec{x_k}$. The question is whether this sequence of vectors converges to a constant probability vector x that satisfies $\vec{x} = M\vec{x}$. This is known as the stationary distribution of the Markov chain, and it is also an eigenvector of matrix M with eigenvalue $\lambda = 1$.

Markov matrix may be built directly from a graph. It is similar to a adjacency matrix, where the position i, j is 1 if $i \implies j$ in the graph. The difference is that in Markov matrix columns should be normalized to sum 1, and each row contains the probabilities to arrive at that node.

PageRank is just the stationary distribution of a specific Markov matrix called Google Matrix, defined in Definition 3.3. This matrix contains two terms: a matrix A that contains the information of the links between the pages, and a matrix B with links between all pages. The matrix B represents the random jump from a webpage to any existing webpage. The probability of a random jump is given by $(1 - \alpha)$, with $\alpha = 0.85$ as standard.

Definition 3.3 ([20]) *Google Matrix*

*A **Google Matrix** is a Markov matrix built from the Web graph of hyperlinks,*

$$G_M = \alpha A + (1 - \alpha)B$$

$$A_{i,j} = \begin{cases} \frac{1}{n_j} & \text{if webpage } j \text{ has a link to } i \\ 0 & \text{otherwise.} \end{cases}$$

$$B_{ij} = \frac{1}{n}$$

Definition 3.4 ([20]) *PageRank*

The ***PageRank*** x_{pr} *of a graph is the steady state of the Markov chain defined by its Google Matrix.*

$$x_{pr} = G_M x_{pr}$$

Modified PageRank

The proposed way to modify PageRank is to add the party, electoral list or political coalition information in the Markov matrix and then find the eigenvector as the traditional PageRank.

We introduced the political information through a political matrix P as defined in Definition 3.5. The idea is to enforce a clique into the Google Matrix. A clique is a subgraph where every node point to all nodes of the subgraph. $P_{\text{electoral list}}$ and $P_{\text{political coalition}}$ were defined in a similar way.

Definition 3.5 *Political Google Matrix*

$$PG_M = (\alpha - \gamma)A + (1 - \alpha)B + \gamma P$$

With A, B the same that in Definition 3.3. P is the political matrix with can be computed

$$P_{party} = \begin{cases} \frac{1}{n_l} & \textit{if i and j have the same party} \\ 0 & \textit{otherwise.} \end{cases}$$

$$P_{coalition} = \begin{cases} \frac{1}{n_l} & \textit{if i and j have the same political coalition} \\ 0 & \textit{otherwise.} \end{cases}$$

$$P_{electoral\ list} = \begin{cases} \frac{1}{n_l} & \textit{if i and j have the same electoral list (same district)} \\ 0 & \textit{otherwise.} \end{cases}$$

Definition 3.6 *Political PageRank*

$$x_{pr} = PG_M x_{pr}$$

We computed Custom political PageRank using P_{party}, $P_{\text{coalition}}$ and $P_{\text{electoral list}}$. Several γ in the range $[0, .85]$ were used, and α was kept constant as 0.85. Notice that original Google Matrix in Definition 3.3 is an special case of our Political Google Matrix in Definition 3.5 when $\gamma = 0$.

3.3 Validation

Both research questions RQ1 and RQ2 were answered evaluating each Twitter influence feature association with votes using several tools:

- Correlation: we computed the Spearman correlation coefficient between each influence feature and the votes obtained
- Regression: we model the problem as a regression task, in which each observation is a candidate, the target variable is votes obtained, and the predictor variables were base features and Twitter influence features. This evaluation was performed in-sample
- Classification: we model the problem as a classification task. Instead of a continuous target variable like votes, we defined positive class as the top p% candidates with more votes. This is equivalent to defining class 1 as those candidates with more votes than percentile p. We used several p values from 50% to 90%.

The univariate correlation will help us ask which features are closer to the election result. In the other hand, regression analysis allows 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 variation of R^2.

As target feature for this validation, we used the district votes percentage of candidates. We prefer the percentage against the raw vote count, because the vote count hides more information related to turnout and number of districts votes. More votes can mean a better performance or just competing in a bigger district.

3.3.1 Preprocessing

Several preprocessing for Twitter features were performed. Three reasons motivate preprocessing: (i) Target variable and features should be in the same scale, (ii) We wanted features as comparable as possible (iii) Specifically for linear regression, we needed transformations to achieve more interpretable results.

Normalization

We selected vote percentage of electoral district of each candidate as target variable for analysis below. This is a local measure of relevance by district. Predictors should also be measured as local (by district) measures of influence instead a raw global number to be consistent with that choice. This would address reasons (i) and (ii), because influence features and electoral result would be measured the same way: per district.

Definition 3.7 *District normalization for candidate i*

$$district\_norm(y_i) = \frac{y_i}{\sum_{j \in district(i)} y_j}$$

We decided to implement a district normalization, defined in Definition 3.7. It represents the mass percentage of a candidate feature relative to their electoral district. We applied this normalization to all influence features (Twitter metrics, Sentiment analysis features and Network features).

This decision has advantages and disadvantages. It advantage is that features can be compared between them and with target variable in a common scale (district based). On the other hand, a disadvantage is that we need data from all the candidates on a district, because the input is a relative metric (percentage). Another disadvantage is that all features keep in the range of $[0, 1]$. This is a disadvantage for our regression validation because this constrain is not assumed in the model.

Improving normality

We decided to transform all features to a normal distribution: This is useful to achieve better results in linear regression and to accomplishing the assumption of the normality of the residuals (the difference between estimated by the regression and the observed values) and homoscedasticity (residuals with constant variance). To achieve normality, we used the Yeo-Johnson transformation [44]. The λ parameter is obtained via Max Likelihood Estimator, implemented in sklearn.

Definition 3.8 ([44]) *Yeo-Johnson transformation*

$$y\text{-}j(y_i, \lambda) = \begin{cases} ((y_i + 1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y_i + 1) & \text{if } \lambda = 0, y \geq 0 \\ -((-y_i + 1)^{(2-\lambda)} - 1)/(2 - \lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y_i + 1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

Standardization

For linear regression, we performed a standardization: scaling to mean 0 and standard deviation 1. This was accomplished computing the Z-score for all our features.

Definition 3.9 *Z-score for variable X*

$$Z = \frac{X - \bar{X}}{s_X}$$

Chaining transformations

The transformations previously listed were applied in a specific order. Log transform was also applied because influence features like PageRank and degree follow exponential distributions [27]. Table 3.5 contains the preprocessing used.

Table 3.5: Preprocessing of Twitter influence features

Preprocessing name	Formula	Used in
District normalization (`dn`)	district_norm(y_i)	Correlation
Log district normalization (`logdn`)	district_norm(log(y_i))	Correlation
District normalization and Yeo-Johnson power transform (`dnpt`)	y-j(district_norm(y_i))	Used only to compare distributions
Log district normalization and Yeo-Johnson power transform (`logdnpt`)	y-j(district_norm(log(y_i)))	Regression, Classification

3.3.2 Base features

For classification and regression tasks, we defined a set of features related to personal, electoral and political characteristics of a candidate. These features can be understanded as control variables for our experiments. Table 3.6 contains all the base features used.

Note that, in total, we have 30 base features. As none of these features are related to Twitter, we use the base features as a baseline for the estimation made with the base features and the Twitter influence features.

Table 3.6: Base features for a candidate u_i

Feature name(s)	Description
gender	Gender of candidate: 1 for woman, 0 for man
rm	Indicates if candidate competes in a district in Metropolitan Region (*Región Metropolitana*, RM), the capital region of Chile. 1 if candidate belong to RM, 0 if belong to other region
n_candidates	Number of candidates competing in the district of the candidate u_i (included). With more candidates the votes keep more sparse, so adding this control variable was neccesary
1, 2, 3, 4	Place of the candidate in their list in ballot as a One Hot Encoding (4 features). Each electoral list in a ballot has an specific order of candidates. If candidate u_i is the first candidate of the list (list head) it will have a value of 1 for feature 1 and 0 for 2, 3 and 4.
CIUDADANOS, COMUNES, CONVER., EVOPOLI, FREVS, IGUALDAD, PCC, PCCH, PDC, PEV, PH, PL, PNC, PPD, PR, PRO, PS, PTR RD, REPUBL. RN, UDI, UPA	Party of the candidate as a One Hot Encoding (23 features)

3.3.3 Correlation

Correlation is a measure of statistical dependency of two variables (i.e., changes in one variable implies changes in the other). For example, assessing the correlation strength between the electoral outcome of candidates and each Twitter influence features gives an idea of which Twitter features hold more information related to votes.

A standard measure of the correlation between two random variables is Pearson correlation coefficient r, defined in Definition 3.10. Pearson's r is a measure of a linear correlation, and it is sensitive to nonlinear transformation, like log. This was a problem, because we preferred not to make assumptions about the data distribution.

Definition 3.10 ([8]) *Pearson correlation*

$$r(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \hat{=} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman correlation

Instead, we seek a correlation metric invariant to monotonic transformations that measure a monotonic (not necessarily linear) correlation. We chose the Spearman correlation coefficient ρ, in Definition 3.11, as the correlation metric, and computed the Spearman ρ between the influence features and the votes percentage of each candidate.

Definition 3.11 ([8]) *Spearman correlation*

$$\rho(X, Y) = r(Rank(X), Rank(Y))$$

Where function $Rank(Z) : 2^{\mathbb{R}} \implies 2^{\mathbb{N}}$ *maps any set of numbers to its ranking, mean that the max value will turn to 1 and the lower value to len(Z). Tied values (repeated values in Z) are mapped to the same index.*

3.3.4 Regression

Correlation is a good first approximation to understand the dependency between Twitter influence and electoral results, but its main pitfall is that does not consider other control variables.

To include more information, we modeled the problem as linear regression. Each candidate is an observation; the target variable is the district votes percentage of each candidate. As predictors, we used the base features defined in Section 3.3.2.

We did not use all the influence features at the same regression, but instead, we computed a single regression for each feature because we wanted to compare features with features in terms of how they determine the electoral outcome.

Definition 3.12 *Linear regression model to evaluate a feature F*

$$y_{(Log\ District\ Vote\ Percentage)} = \beta_0 + \beta_F x_F + \sum_{i \in Base\ Features} \beta_i x_i$$

$$\forall F \in Twitter\ metrics \cup Sentiment\ analysis\ features$$
$$\cup\ Network\ analysis\ features \cup Custom\ PageRank\ features$$

For each feature F named in Section 3.2, we performed a Log District Normalization Yeo-Johnson and then a Standardization, as explained in Section 3.3.1. After that, the regression model defined in Definition 3.12 was used and two evaluation metrics were obtained.

The first metric is coefficient β_F. The idea behind the normalization and standardization is to be able to compare the different beta coefficients in terms of their standard deviation [1]. There are some criticisms to this approach because standardized coefficients are not numerically interpretable [21], but we use their value relative to standardized coefficients of other features and keep the same base predictors.

Definition 3.13 ([16]) R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

Where SS_{res} is the residual sum of squares, SS_{reg} is the regression sum of squares, and $SS_{tot} = SS_{reg} + SS_{res}$.

Definition 3.14 ΔR^2 *of a feature F*

$$\Delta R^2_F = R^2_{\textit{base features} \cup \{F\}} - R^2_{\textit{base features}}$$

The second metric is the R^2 score, as defined in Definition 3.13. The interpretation of this score is the percentage of variance of dependent variable explained by the independent variables. This is the same that $1-$ the percentage of variance captured by the residuals, which by definition is not captured by the regression. R^2 is a metric of goodness of fit related to linear regression.

We compared the difference between the R^2 of the regression with the F feature and a base regression without the F features (only Base features and intercept). For each F, we measured that difference as ΔR^2, defined in Definition 3.14. The interpretation for this delta is the percentage of the electoral result variance explained by feature F.

Interaction term

The linear regression model assumes that the effect of each predictor in the target variable is independent. *Does the relation between the retweet count and percentage of votes of a candidate depend on gender? Does it depend on the location of the candidate?* We need to model the relationship between these categorical features and the influence features to answer questions like this. We included an interaction term in each feature's regression to achieve this, as shown in Definition 3.15.

Definition 3.15 *Linear regression model with interaction term $\beta_{I(F,c)}$ between F and c*

$$y_{\textit{(Log District Vote Percentage)}} = \beta_0 + \beta_F x_F + \beta_{I(F,c)} x_F x_c + \sum_{i \in \textit{Base Features}} \beta_i x_i$$

$$\forall F \in \textit{Twitter metrics} \cup \textit{Sentiment analysis features} \cup \textit{Network analysis features} \cup \textit{Custom Pagerank features}$$
$$c \in \{\textit{gender, rm}\}$$

Variable c must be binary (0 or 1). When $c = 1$, we can refactor the coefficient of x_F in the regression as $\beta_{F'}$ using $\beta_{F'} x_F = (\beta_F + \beta_{I(F,c)}) x_f$. When $\beta_{I(F,c)}$ is significant, we can conclude that variable c is a moderator [9] in the relationship between variables F and y. A moderator is a variable *"that affects the direction and/or strength of the relation"* of a

predictor and the dependent variable [3]. The value of $\beta_{I(F,c)}$ shows the difference in the strength of this relationship when $c = 0$ vs $c = 1$.

For each Twitter influence feature F, we computed the moderation effect with the binary variables `gender` (1=woman, 0=man) and `rm` (1=Metropolitan Region, 0=Another region) presented in Table 3.6.

3.3.5 Classification

Linear regression allowed us to assess the strength of the linear relationship between Twitter influence features and electoral outcome of each candidate, using base features as control variables. That relationship is measured for all the candidates in all the range of vote percentage for range $[0, 1]$.

The OLS considers all observation equally. Although in electoral processes, the number of candidates may be considerably higher than the number of seats. For example, in the election we are studying (2021 Constitutional Convention election) there were 1279 candidates for 138 seats, so only $\sim 10\%$ of the candidates were elected. This means the relevant candidates (those we want to infer from data) are a small fraction of the sample. Therefore, we used a model to capture the performance of the Twitter influence features in estimating the top-n% candidates with more votes. Instead of a regression task, this transforms the problem into a classification task of top-n% candidates.

Labelling strategies

The target variable must be discretized in two or more classes to transform the regression problem into a classification problem. We defined two classes: a candidate belonging to positive class (1) when it is part of the top-n% candidates with more votes in their district. Else, they belong to the negative class (0). To separate the positive and negative classes, we defined a vote percentage threshold, where values lower (or equal) are set to class 0 and values higher are set to class 1. To define the threshold, district quantiles were used. For each district, we computed the q quantile of the district percentage of the candidates.

The following are examples of what happens with different values of q. A quantile $q = .5$ split the district in two sets of $\sim 50\%$ (both set is a different class). A quantile $q = .9$ splits the district in a negative class set with $\sim 90\%$ of candidates and a positive class set with the other $\sim 10\%$. Values are approximate because the exact cut depends on the number of candidates of each district.

For example, in a district with an odd number of candidates (e.g. 7), taking the quantile $q = .5$ (median) would not lend to an exact 50%-50% balanced positive and negative classes distribution. Because positive class is defined as the candidates with vote percentage **higher** than median, only 3 of 7 (42.85%) would be positive class and 4 of 7 negative class.

Instead of this district quantiles approach, we could have used if the candidates got a seat as the target class (1 if winner, 0 if loser). The seats are a function of the election outcome

and the electoral system in place. For this election case, D'Hondt method [14] was applied to assign seats to electoral lists and parties, but also gender correction took place. As the seat winning depends of other factor rather than the candidate itself, we keep the votes-based class.

In summary, we set the class of each candidate based on the q quantile of votes in their districts. We experimented with several q values from 0.5 ($\sim$ 50% True class) to 0.9 ($\sim$ 10% True class). The input for the classifiers was built with the same methodology of Linear regression: a model was fitted for each feature and base features were added.

Cross validation

The performance of the classifiers was evaluated out-of-sample (i.e., the data used in training the classifier is different from the data used for the evaluation). We defined a specific way of evaluating named Leave-one-district-out Cross Validation. It is based on a common evaluation scheme called leave-one-out cross validation, when for n observations, n models are trained using $n-1$ samples and evaluated using the remaining sample.

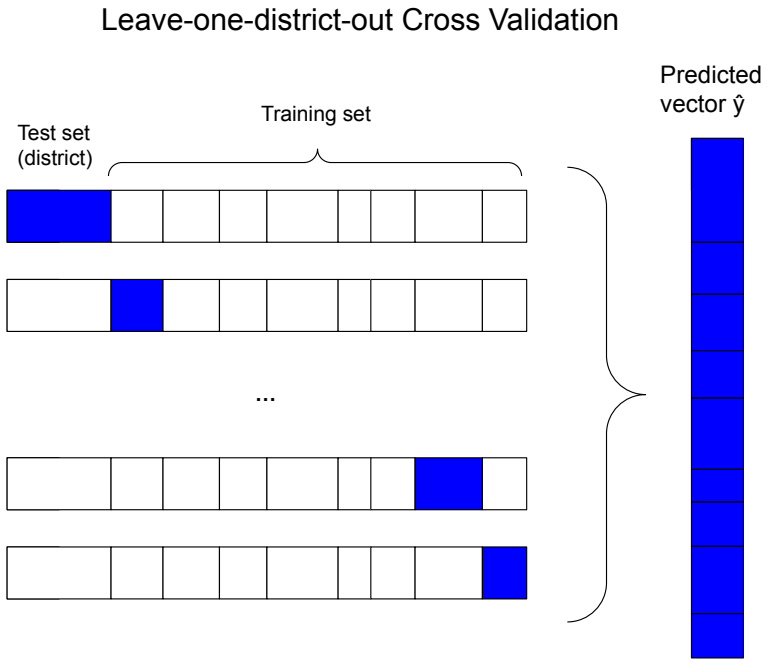


Figure 3.1: Leave-one-district-out Cross Validation. Each fold (chunk of candidates) is a district. Folds have different sizes to illustrate that districts have different number of candidates. Predicted vector $\hat{y}$ is used to compute all evaluation metrics

We defined the folds (packages of observations) of the cross validation as the electoral districts. This divided our data in 28 districts with a variable number of candidates in each one. We used each district as test (predicted) set and the remaining 27 districts to train a classifier. Rather than computing the average of the district metrics, we concatenated

these predictions to have a single vector $\hat{y}$ that contains out-of-sample predictions for each candidate. Comparing the estimation with the real values y, we obtained several performance metrics (defined below) for the classifiers. The full process is shown in Figure 3.1.

Performance metrics

As the classes may be highly imbalanced (e.g. in the case of using quantile $q = 0.9$ for classes), the performance metrics should capture both if the classifier is sensitive to the positive class (recall) and if it raises precise predictions (precision). Both concepts are defined in Figure 3.2.

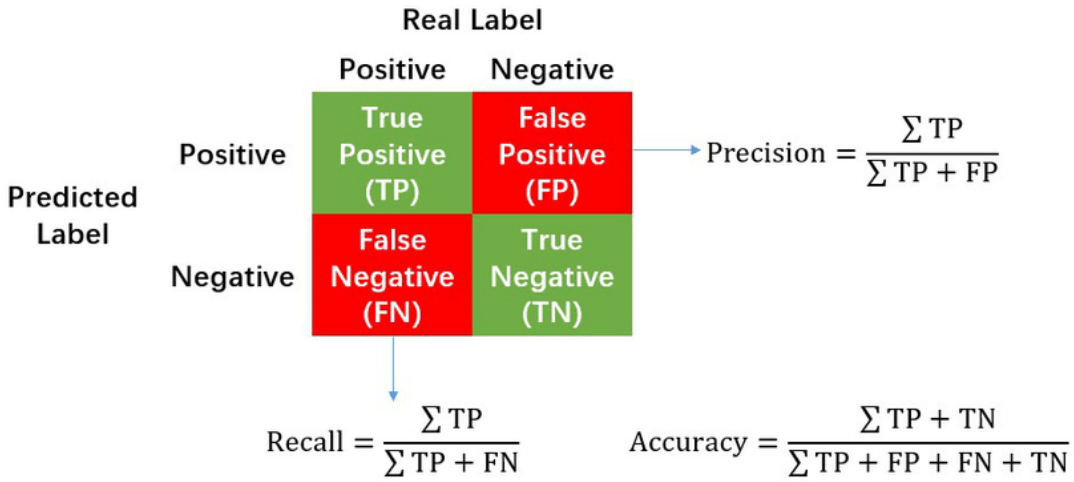


Figure 3.2: Definition of Precision and Recall in a confusion matrix of a binary classification (Figure from [28])

There is a tradeoff between precision and recall. A classifier may be conservative to assign a True class, that would lead to a high precision but low recall situation. On the other hand, a generous classifier assigning lots of True classes may achieve a higher recall but with a cost in the precision of their predictions. Good validation metric for classifiers need to consider both components of the classification.

We decided to use a metric that combines precision and recall. We selected Precision Recall Curves (PR). Precision Recall Curves are a way to summarize the tradeoff between precision and recall within a single classifier, so it's more robust that a raw F1 score. To do that, we need the probability of class that the model assigns to each observation and compute precision and recall for each probability threshold. This means that precision recall curves capture the notion of a conservative classifier (a higher probability threshold, assigns fewer positive class, but more accurate) and a generous classifier (a lower probability threshold, has higher recall, bus less accurate) that we mention earlier. Figure 3.3 contains an example of PR curves.

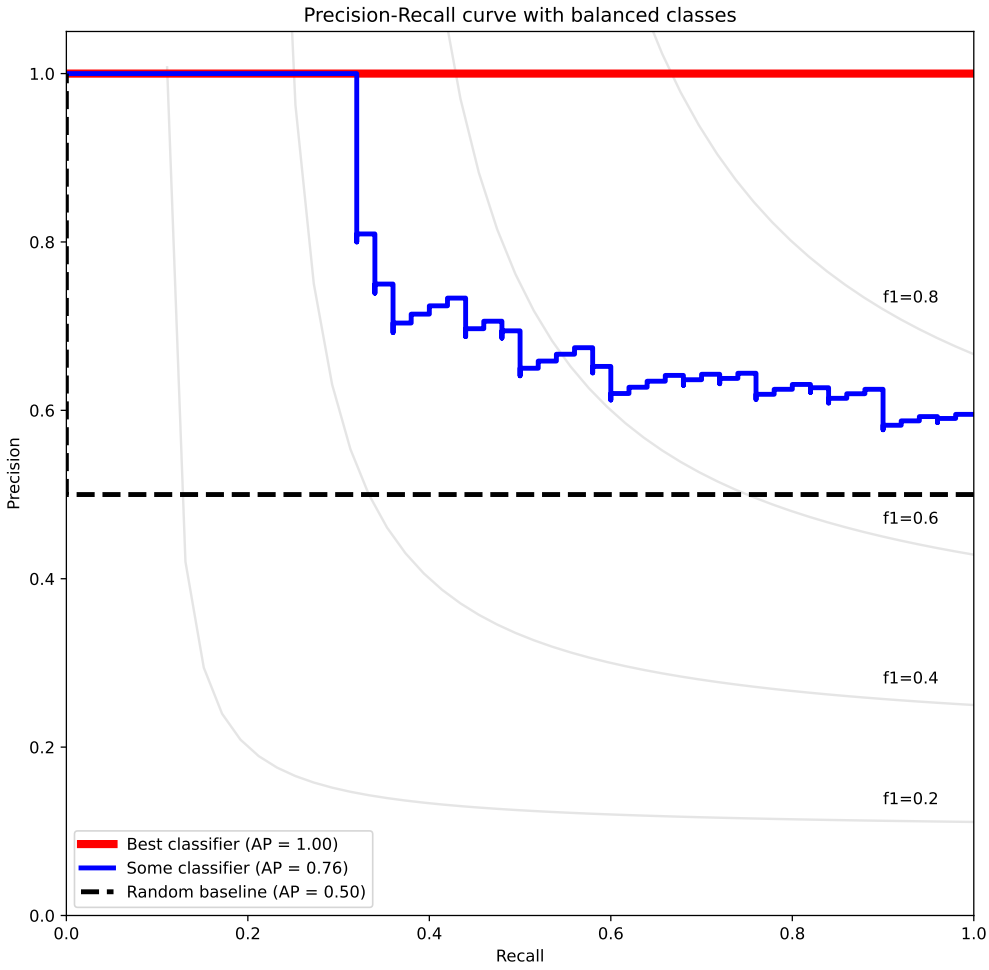


Figure 3.3: Example of Precision Recall Curves. Red line represents the perfect classifier, with 100% precision for all recalls. Black dashed line represents a random classifier, with average precision equivalent to the percentage of positive class observation (in this case, as classes are balanced, baseline is 50% precision). Blue curve represents a classifier better than random. Precision-recall trade off appears when Precision is lower for higher recalls.

A way to summarize a Precision Recall curve is the Area Under the Curve (AUC), also known as Average Precision, as defined in Definition 3.16. This metric is more robust that a raw F1 score, because it accounts for multiple possible thresholds chosen by an specific classifier, rather that choosing a single one, that is the case of F1.

Definition 3.16 ([30]) *Average Precision (AP), equivalent to Precision Recall Area Under the Curve (PR-AUC)*

$$AP = \sum_n (Recall_n - Recall_{n-1}) Precision_n$$

Definition 3.17 ΔAP *of a feature* F

$$\Delta AP_F = AP_{base\ features \cup \{F\}} - AP_{base\ features}$$

To evaluate the performance of each Twitter feature, we compared the performance of the classifier using only the base features with the classifier adding a Twitter feature F, as defined in Definition 3.17. This is similar to what we did for regression in Definition 3.14. This metric can be interpreted as the amount of precision that is due to each feature F. A graphic interpretation can be found in Figure 3.4. If adding a feature F worsens the average precision, ΔAP will be negative.

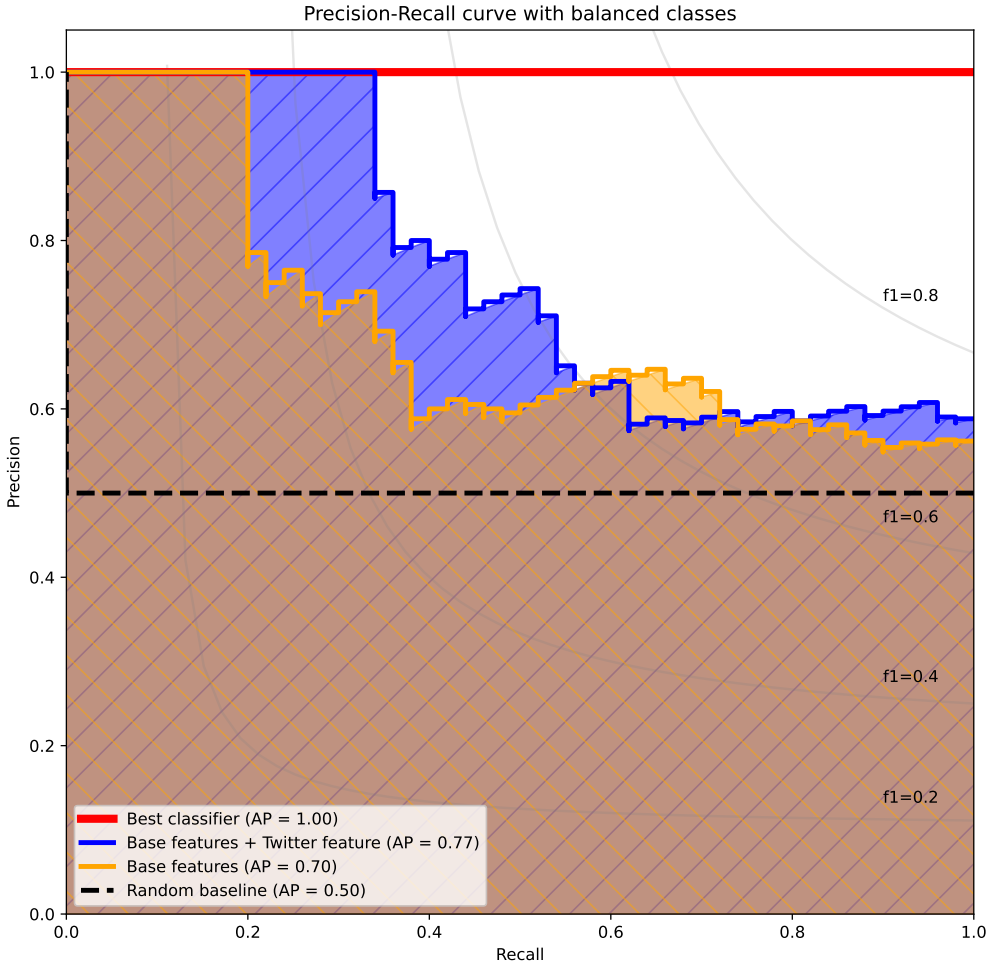


Figure 3.4: Example of ΔAP. Orange curve represents the classifier performance using only base features. In that case, Area Under the Curve is $AP = 0.70$. Blue curve represents the classifier performance using base features and a Twitter feature F. We expect performance to increase when adding a Twitter feature, like in this example, with an area $AP = 0.77$. In this case $\Delta AP_F = 0.77 - 0.70 = 0.07$, and that value is equivalent to the area between the two curves (blue and light orange), including possible negative values.

Like the regression model, for each feature F named in Section 3.2 (Feature Engineering), we performed a Log District Normalization Yeo-Johnson, as explained in Section 3.3.1 (Preprocessing). A Random Forest classifier with 100 predictors and undersampling was used. Standardization was not need because the Decision Trees are invariant to the scale of the inputs.

3.3.6 Political PageRank

We evaluated the Political PageRank features using classification the same way we evaluate the other features. The procedures named above were repeated to these features. Instead of comparing the Modified PageRank to the base model, it is compared with the regular PageRank to measure if the modification had any effect in that metric performance.

Chapter 4

Results

4.1 RQ1: Twitter influence features assessment

4.1.1 Descriptive analysis

As an exploratory analysis, we checked the distribution of vote percentage depending on whether the candidate had a Twitter account. We considered as candidates with Twitter only the candidates whose account was scraped, if no account was found, it was assumed the candidate did not have an account. The boxplot in Figure 4.1 confirms the first intuition that candidates with Twitter have more votes that candidates without Twitter. The quartiles of Twitter candidates' sample are higher than for No Twitter candidates.

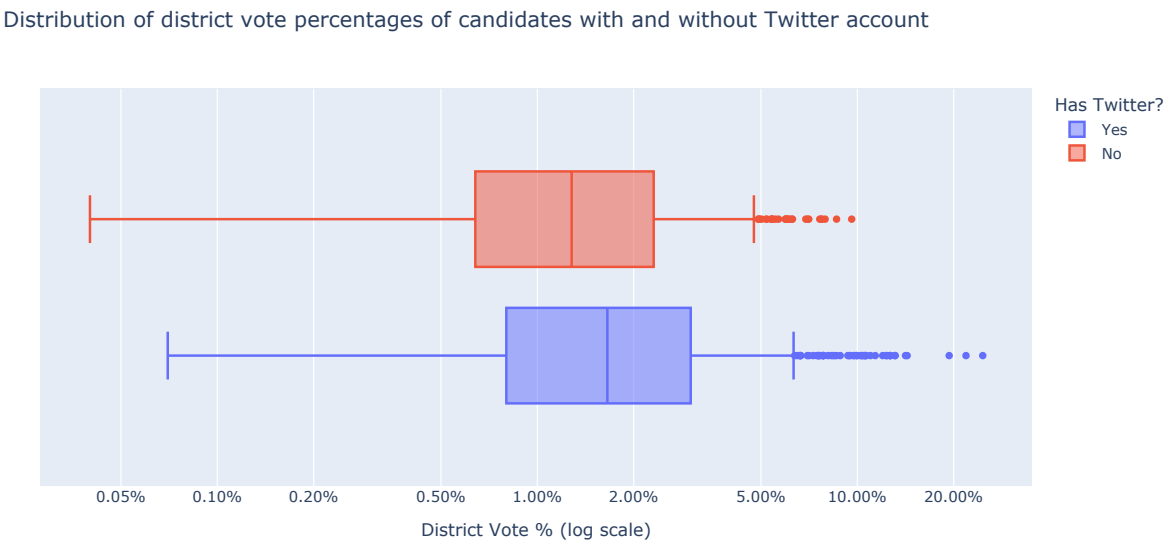


Figure 4.1: Comparison between the vote percentage distribution of candidates depending on if they have a Twitter accounts, based in the web scraping.

We confirmed this difference doing a T-test of difference of means. We used the log-district percentage, as votes percentage is similar to a log-normal distribution. The null hypothesis is $H_0 : \mu_{\text{Twitter}} \leq \mu_{\text{No Twitter}}$ and the alternative hypothesis $H_a : \mu_{\text{Twitter}} > \mu_{\text{No Twitter}}$, where μ is the vote percentage mean of each sample. We used a significance level of 5%. Running the test with our candidate data returned a T-statistic value of -5.311 and a p-value $< 10^{-8}$. Null hyphothesis was rejected, socandidates for this election with Twitter had significantly a higher percentage of votes that candidates without Twitter accounts in average.

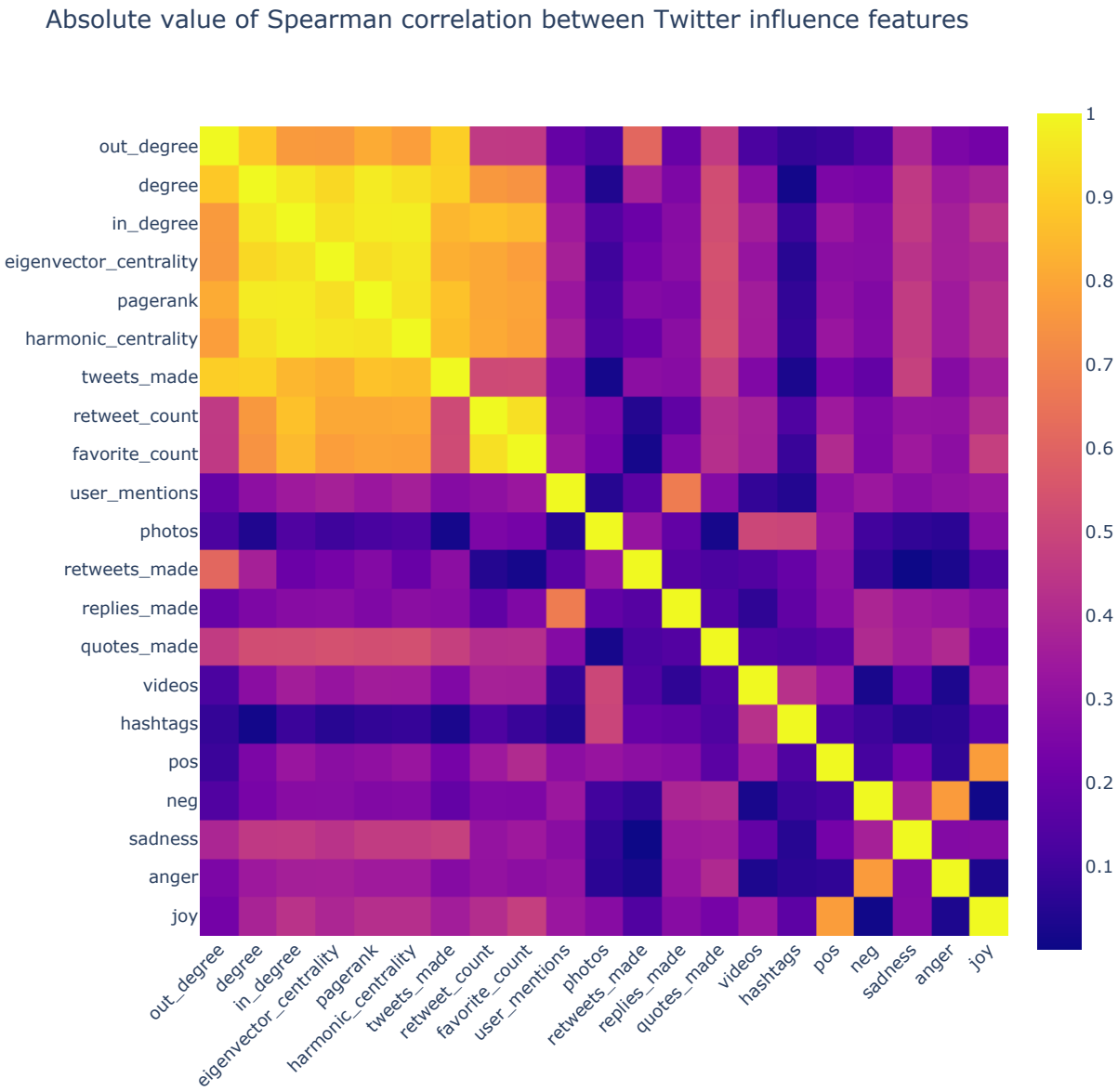


Figure 4.2: Correlation between raw Twitter influence features

Figure 4.2 shows the correlation between all the influence features computed from Twitter. We used the absolute value of Spearman to visualize the magnitude of the correlation, not it positive or negative sign. As shown in the figure, all network centrality metrics (`out_degree`,

`degree`, `in_degree`, `eigenvector_centrality`, `pagerank`, `harmonic_centrality`) correlate with Spearman values > 0.8. Furthermore, those network features also correlate with other tree features: `tweet_made`, `retweet_count` and `favorite_count`. This make sense: with more tweets, there are more retweet possibilities, and also more retweets mean more exposition and more favorites. Finally, all network features were built from the retweet graph, so, consistently, retweet count correlates with all these metrics.

The correlation between the network centrality features can be observed in more detail in Figure 4.3. Spearman correlation captured the correlation of `harmonic_centrality` with other features regardless the non-linearity of that correlation.

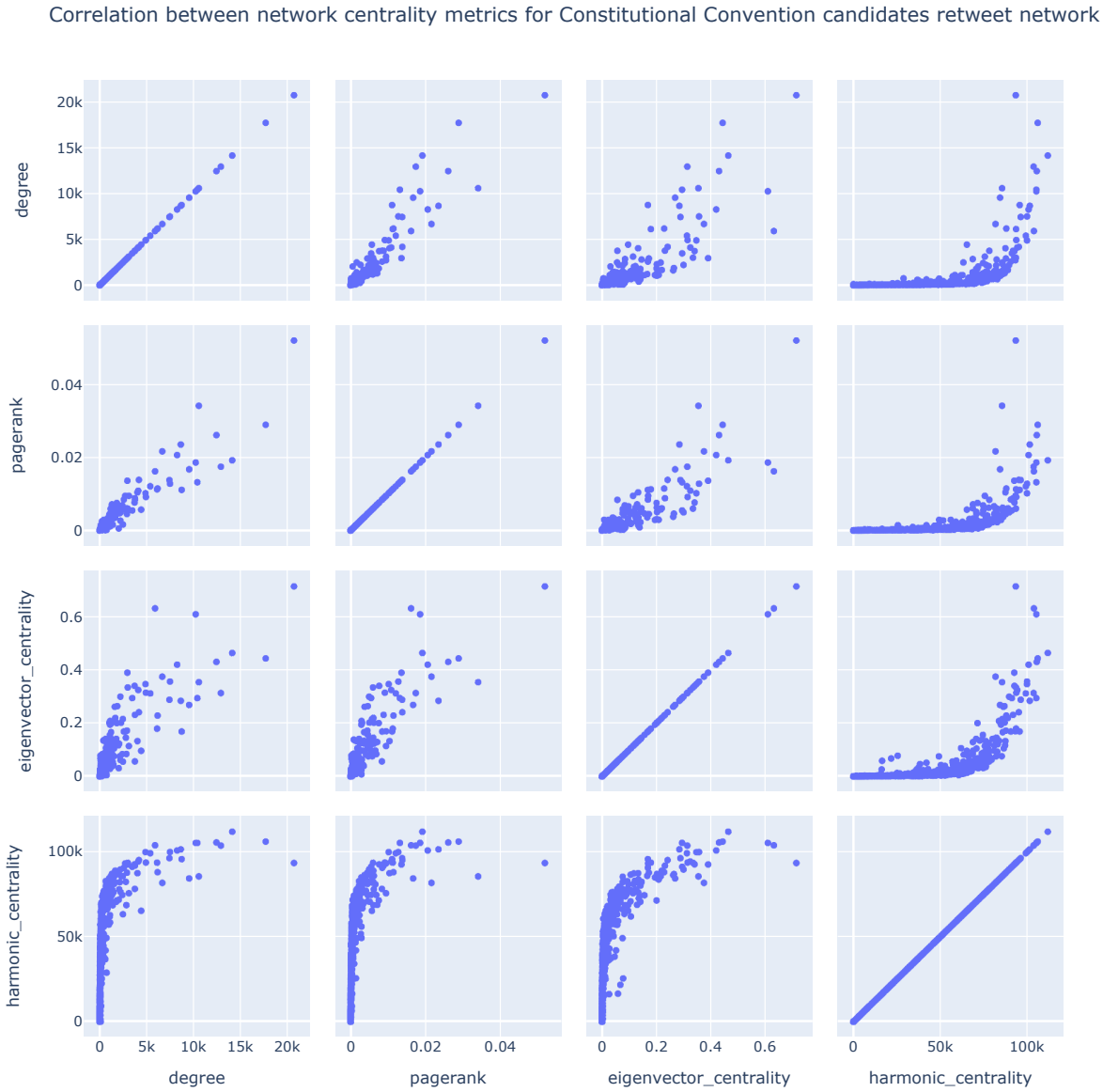


Figure 4.3: Correlation between raw Network analysis features

4.1.2 Correlation

Figure 4.4 (data in Table 4.1) shows the Spearman correlation between the vote percentage of each candidate and each feature proposed. Several aspects can be observed from this plot.

First thing we can observe is related to preprocessing: District normalization increased the correlation with electoral outcome compared to the raw (without preprocessing) feature. This was an expected behavior of the preprocessing, because, as we said, the district normalization sets both dependent and independent variable on the same scale.

The best features in terms of correlation with votes are `favorite_count` and `degree`, with Spearman > 0.5 and variables `pagerank`, `retweet_count`, `in_degree`,`tweets_made`, `harmonic_centrality`, `out_degree` achieve Spearman > 0.4. Those eight variables have a moderate correlation with electoral result, according to Dancey & Reidy (2007) [11]. The rest of the variables have a weak correlation with votes.

Another interesting fact is that features related to positive sentiments had better correlation with votes (`pos`: $\rho = 0.334$, `joy`: $\rho = 0.345$) that negative sentiments (`neg`: $\rho = 0.092$, `anger`: $\rho = 0.047$).

Table 4.1: Spearman correlation between Twitter features and votes percentage of candidates. *** $p < .001$, ** $p < .01$, * $p < .1$

	Preprocessing		
feature	raw	district normalize	log district normalize
anger	-0.074*	0.044	0.047
neg	-0.119**	0.082*	0.092*
hashtags	-0.013	0.084*	0.118**
sadness	0.092*	0.15***	0.151***
videos	0.075*	0.162***	0.165***
photos	0.041	0.153***	0.167***
quotes_made	0.045	0.166***	0.17***
replies_made	0.006	0.187***	0.197***
user_mentions	-0.045	0.207***	0.238***
eigenvector_centrality	0.199***	0.334***	0.325***
pos	0.167***	0.327***	0.334***
retweets_made	0.151***	0.32***	0.335***
joy	0.219***	0.338***	0.341***
out_degree	0.183***	0.335***	0.422***
harmonic_centrality	0.193***	0.382***	0.426***
tweets_made	0.162***	0.374***	0.444***
in_degree	0.231***	0.449***	0.463***
retweet_count	0.251***	0.485***	0.478***
pagerank	0.235***	0.481***	0.48***
degree	0.242***	0.477***	0.517***
favorite_count	0.27***	0.535***	0.537***

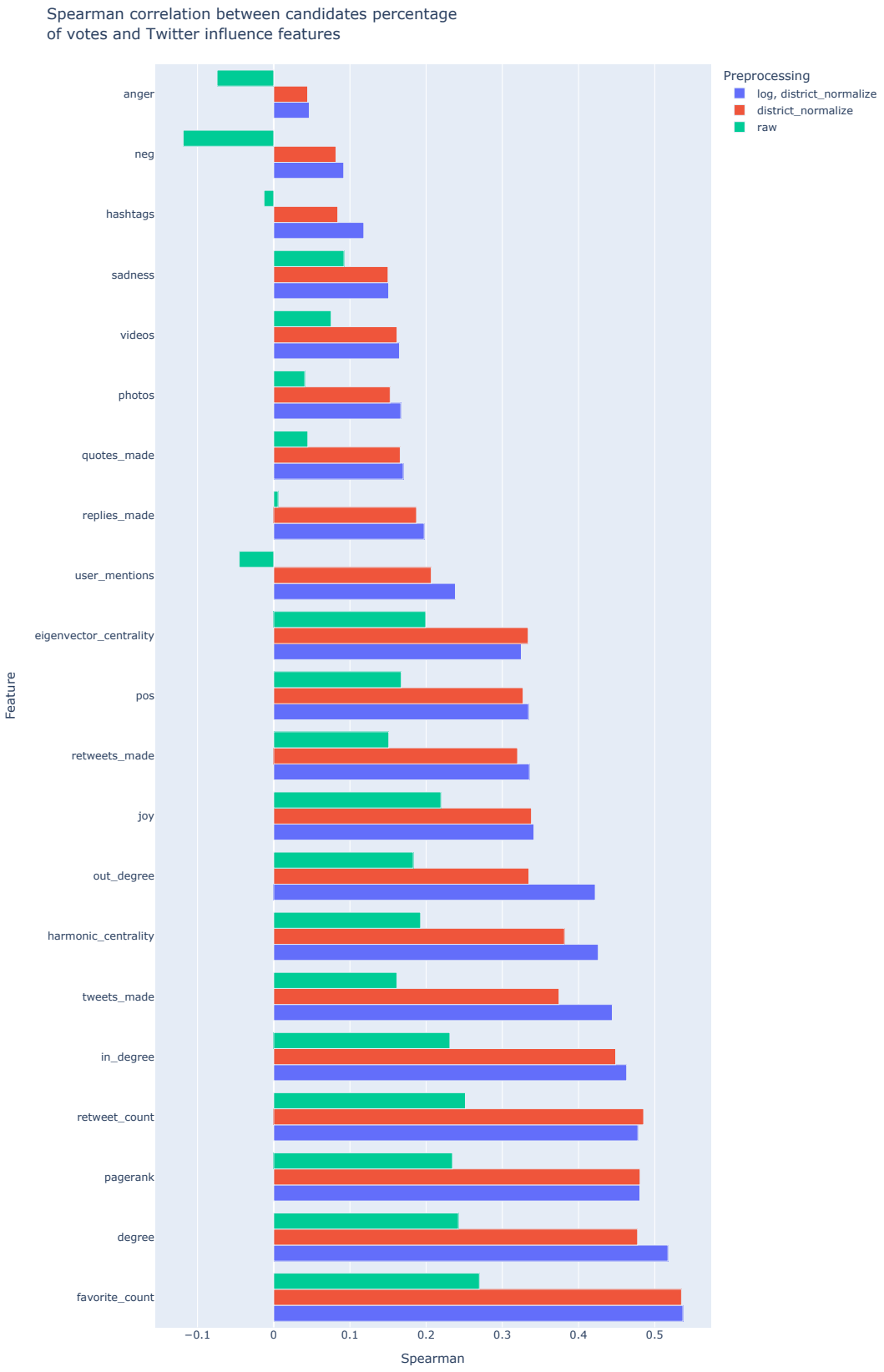


Figure 4.4: Spearman correlation between Twitter features and votes with different preprocessing

4.1.3 Regression

Standardized coefficients

We performed regressions and computed standardized coefficients β_F and explained variance ΔR^2 for each influence feature, as detailed in Section 3.3.4. Figure 4.5 shows the standardized coefficient of influence features using the same base features in each regression. Using a significance level of 5%, we can notice that positive sentiment features (`joy` and `pos`) have significant coefficients, but negative sentiment features (`neg` and `anger`) does not. In that order, the features more closely related to the vote percentage are `favorite_count`, `retweet_count`, `pagerank`, `degree`, `in-degree` and `eigenvector_centrality`. These features are also correlated between them, as we saw in Figure 4.2.

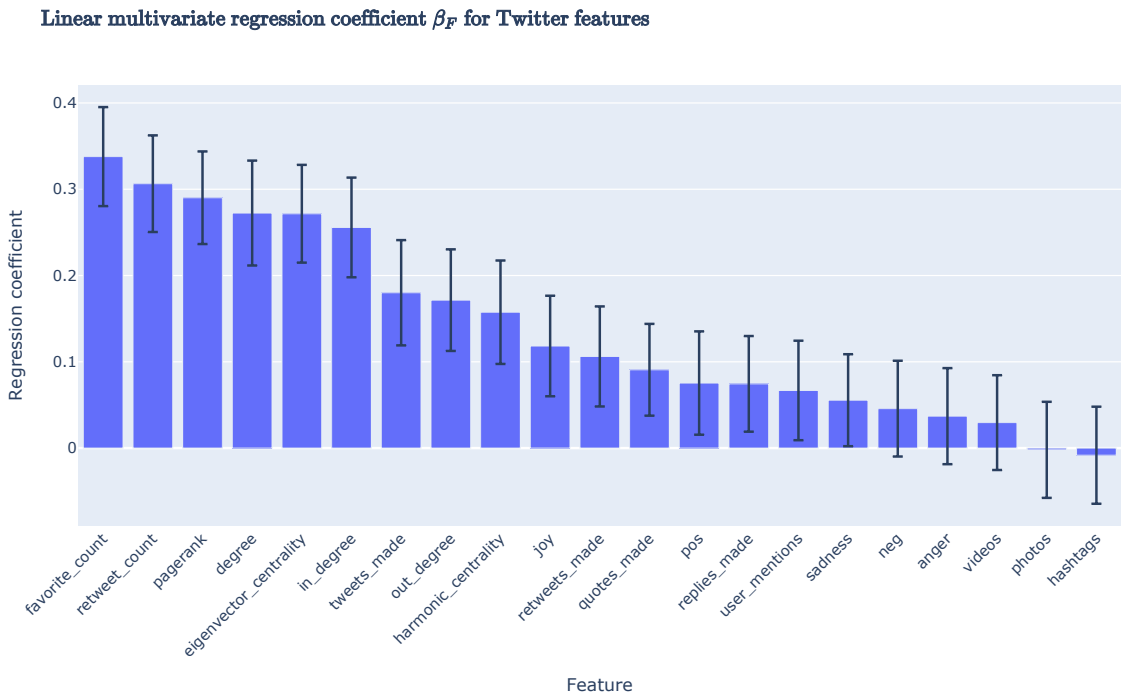


Figure 4.5: Standardized regression coefficient for Twitter influence features. Error bars represent the 95% confidence interval for each estimated value β_F

Explained variance

Figure 4.6 shows the percentage of variance explained by each feature. The value of ΔR^2 is consistent with β_F, as the top-6 features that explain more variance of the votes percentage (`favorite_count`, `retweet_count`, `pagerank`, `degree`, `in-degree` and `eigenvector_centrality`, all with $\Delta R^2 > 4\%$) are the same with higher standardized coefficients. The best result is obtained by `favorite_count`, which explains 7.7% of variance of votes percentage.

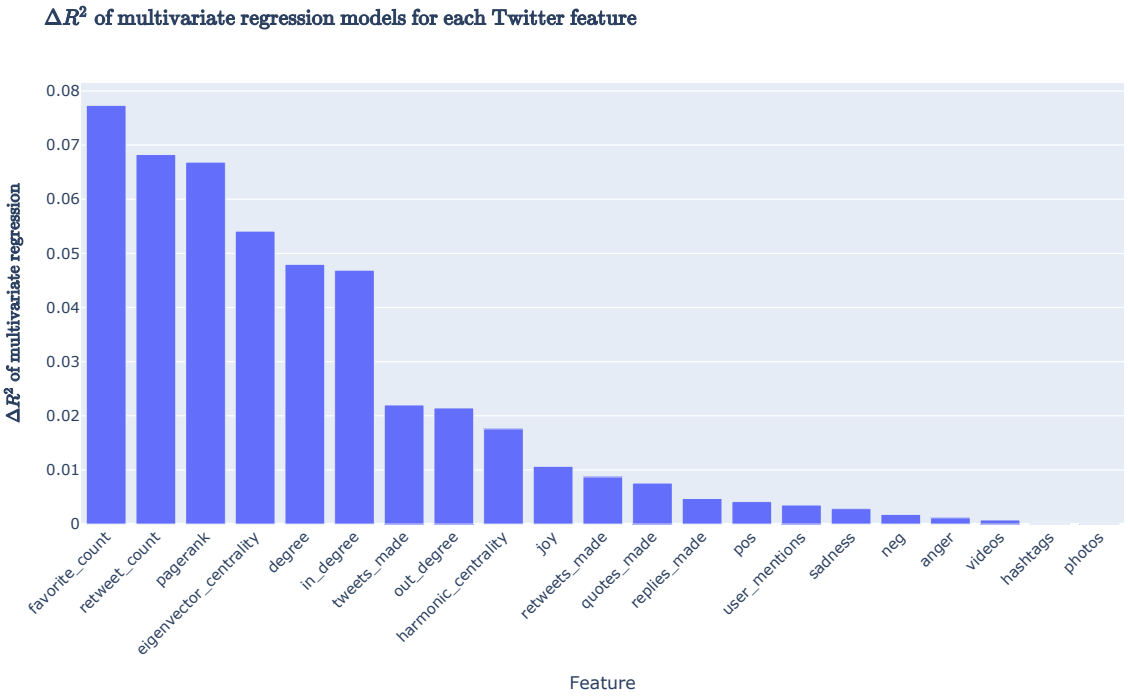
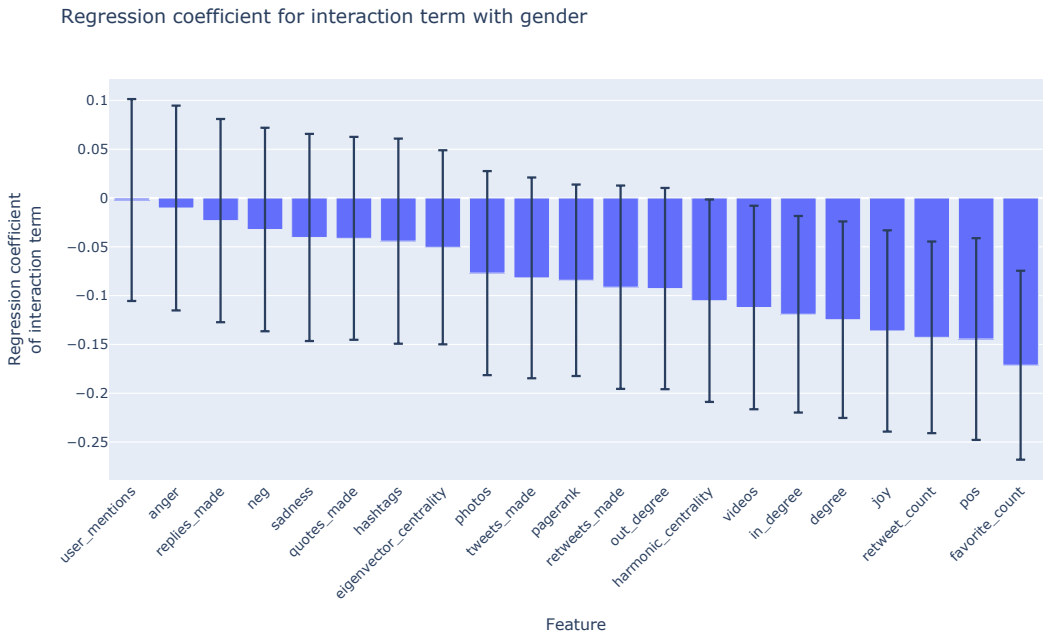


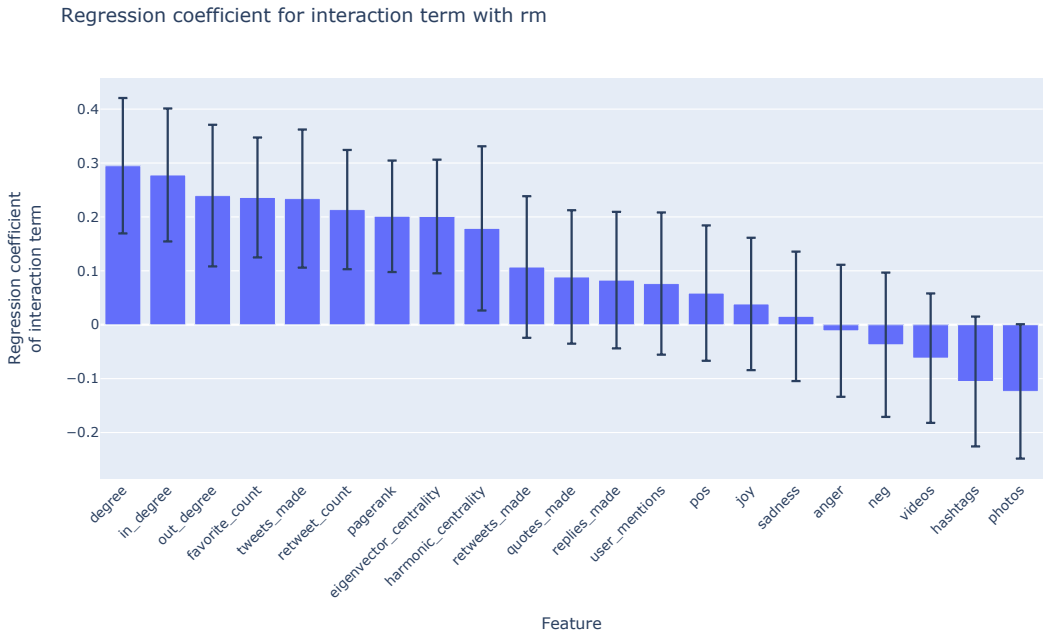
Figure 4.6: ΔR^2 for Twitter influence features

Interaction term

We performed regressions and computed interaction term coefficients $\beta_I(F, c)$ for each influence feature and for categorical variables `gender` and `rm`, as detailed in Section 3.3.4.



(a) Interaction betweeen gender and Twitter features

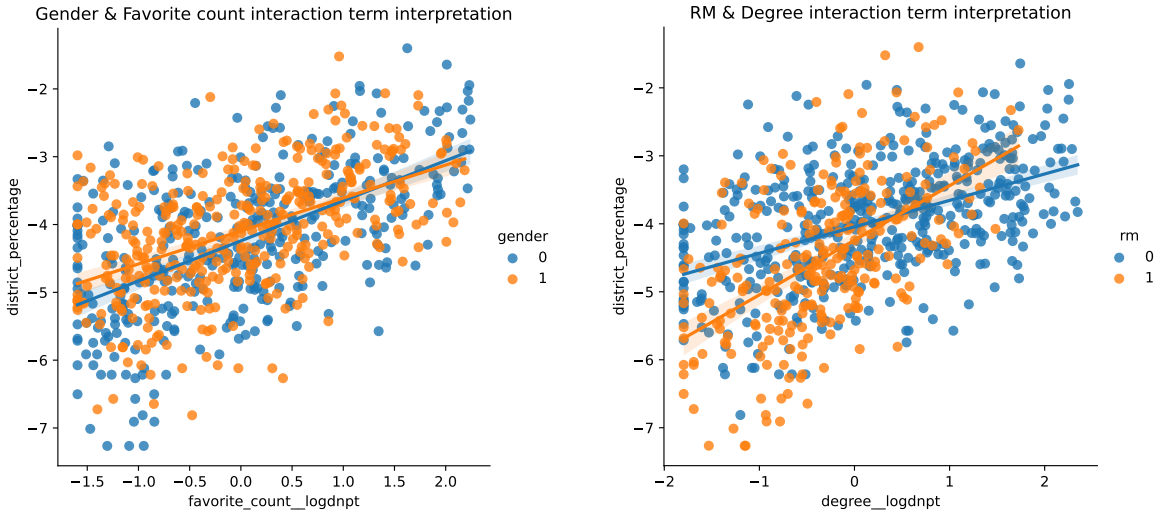


(b) Interaction between rm (Metropolitan Region) and Twitter features

Figure 4.7: Interactions terms

Figure 4.7a shows the interaction coefficients with variable gender. The coefficient is significant for variables `favorite_count`, `pos`, `retweet_count`, `joy`, `degree`, `in_degree`, `videos`. Gender moderates the relationship between the votes percentage of a candidate and the variables listed. All the values are negative, that means that the relationship with votes is weaker for gender = 1 (women) and stronger for gender = 0 (man).

Figure 4.7b shows the interaction coefficients with variable rm. The coefficient is significant for variables `degree`, `in_degree`, `out_degree`, `favorite_count`, `tweet_made`, `retweet_count`, `pagerank`, `eigenvector_centrality`, `harmonic_centrality` (all correlated). As the value is positive, this means that relationship with votes is stronger for rm=1, candidates from Metropolitan Region, and weaker for candidates from other regions.



(a) Interaction betweeen gender and Twitter features

(b) Interaction between rm (Metropolitan Region) and Twitter features

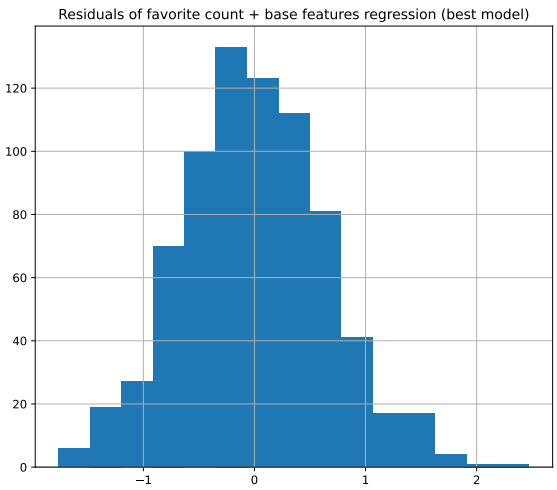
Figure 4.8: Interaction terms interpretation examples

A graphic interpretation of interaction term is shown in Figure 4.8. The value of interaction coefficient $\beta_I(F, c)$ is the angle between the regression lines for each category. In Figure 4.8b the value of $\beta_I(F, c)$ should be positive, because the slope (coefficient) of `degree` for rm=1 is greater than for rm=0. On the other hand, in Figure 4.8a, the interaction term should be negative, as slope for gender=1 (woman) is lower than for gender=0 (man). Both assertions were confirmed in Figure 4.7.

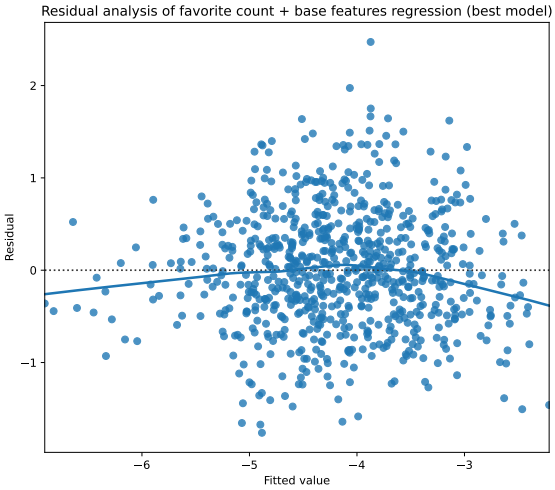
Best regression model

Figure 4.9 shows the regression results for base features plus feature `favorite_count`, which achieve the best results in terms of Pearson correlation and R^2.

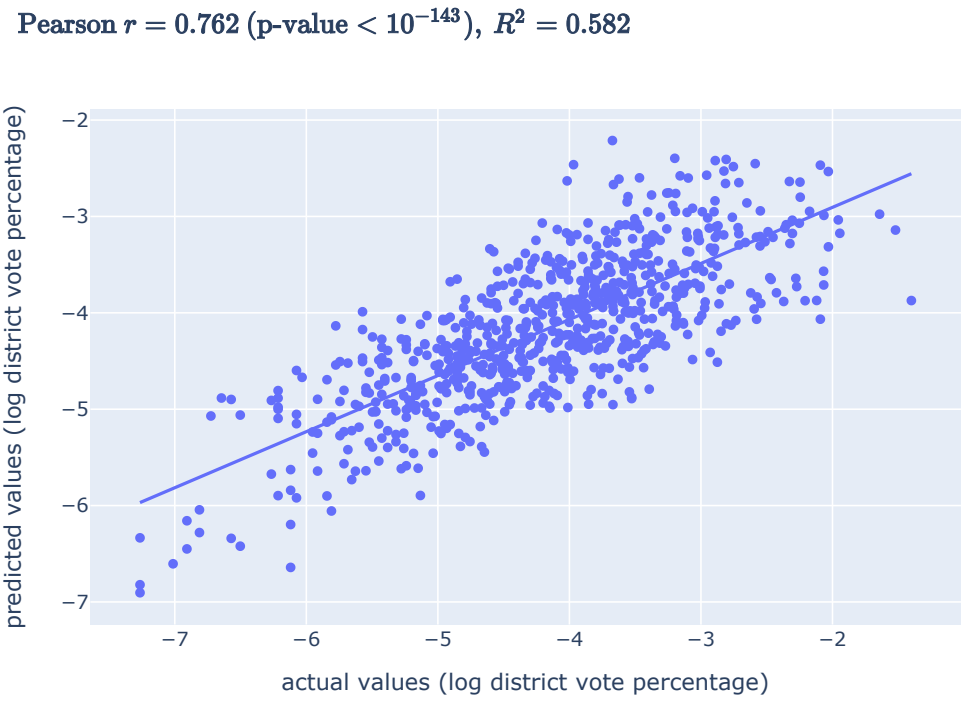
Figure 4.9a shows that the residual of the model has a normal-like distribution. Plot 4.9b shows a subtle heteroscedasticity in the trend line. Finally, Figure 4.9c show the fitted model values compared to the real values.



(a) Residual distribution



(b) Residual of fitted values



(c) Actual vs. predicted electoral results

Figure 4.9: Regression analysis of the best linear model (base features + favorite count feature)

4.1.4 Classification

We performed classification for each Twitter influence feature using a labeling strategy based on the quantile of district votes, as detailed in Section 3.3.5.

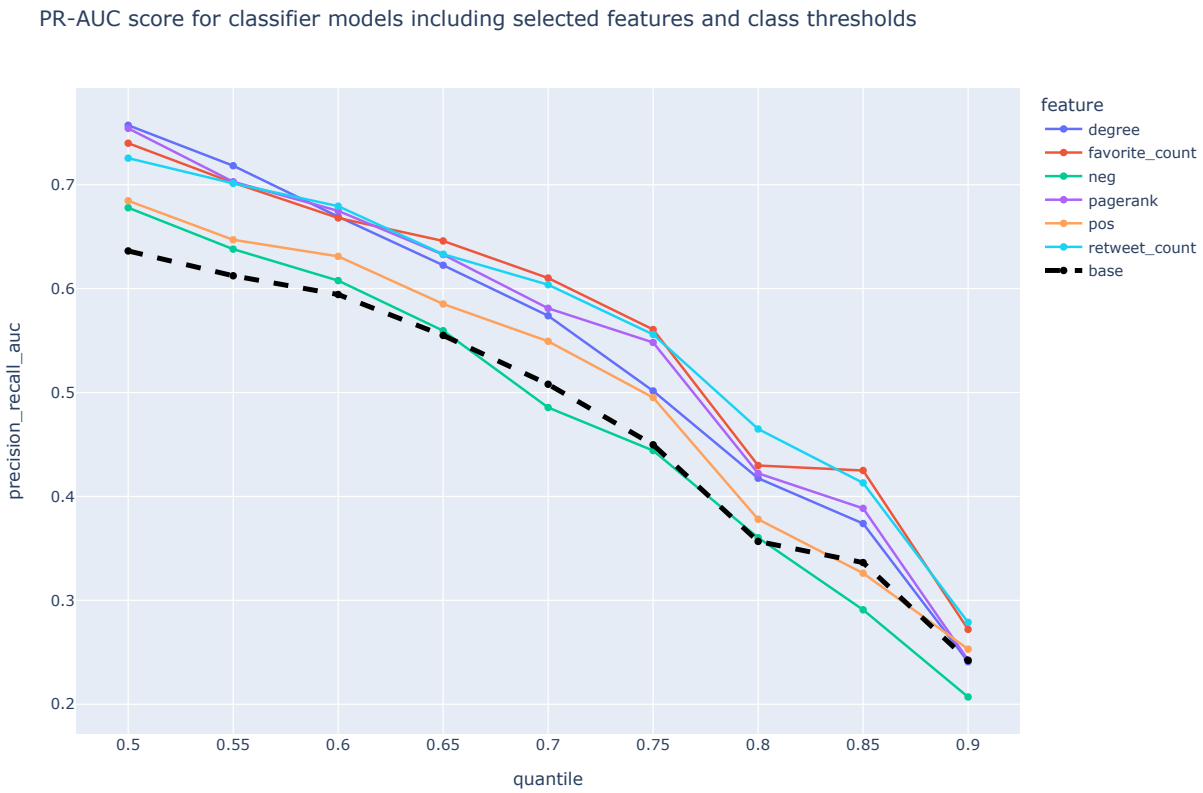
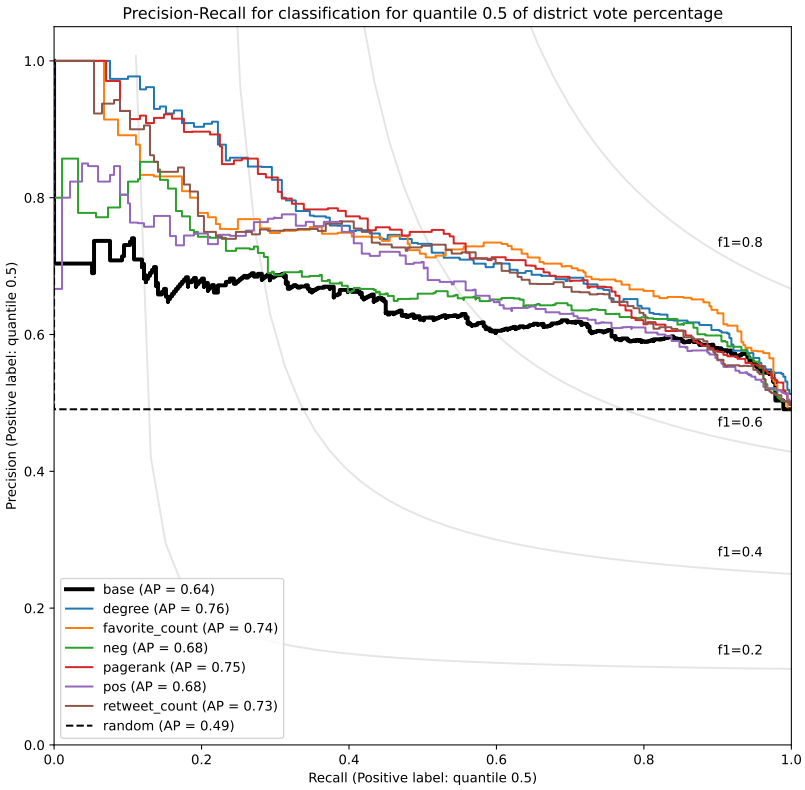


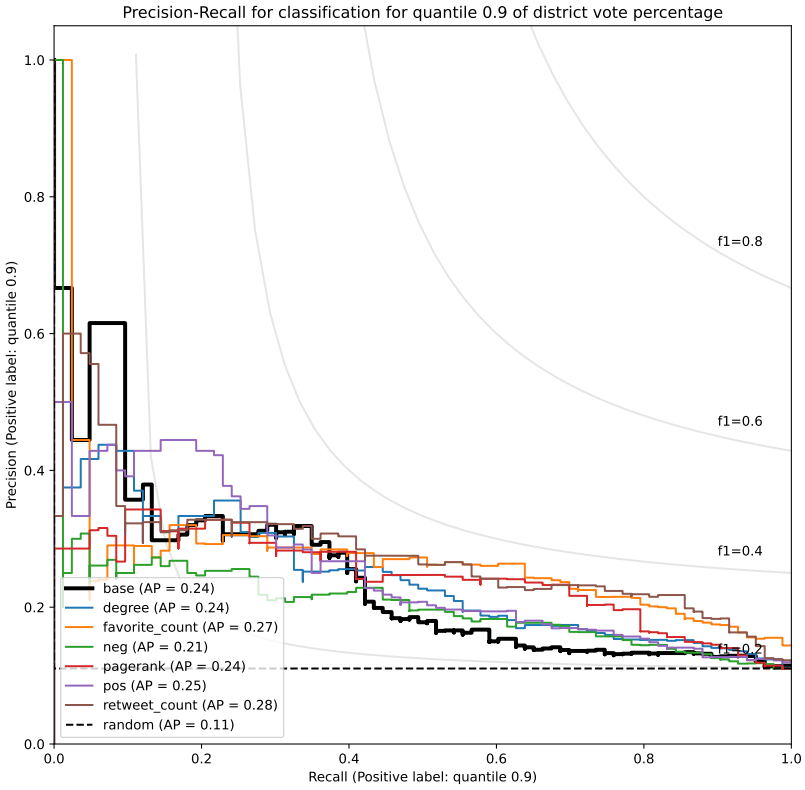
Figure 4.10: PR-AUC score for different threshold quantiles and features

Figure 4.10 shows the classification performance results using different class thresholds (quantiles). Each dot in the figure represents a different classifier. For example, the blue point in (0.75, 0.5) represents a classifier trained using base features ∪ {`degree`}, where positive class is a candidate with more votes than 75% of candidates, in each district. In that case, 0.5 is the Precision Recall Area Under the Curve (or Average Precision) value from cross-validation estimation using this classifier.

The dashed line is the classifier performance using only base features (no Twitter features). The expected result when including a Twitter influence feature is to have a better PR-AUC performance, as more information is available to infer the electoral outcome. This is true for features like `degree`, `favorite_count` and `pagerank`, which consistently achieve higher values that the baseline. In the best case, the PR-AUC gain is about ~ 0.1. Notice that this gain (the difference between the baseline and the model with Twitter feature) is exactly ΔAP_F (having a Δ for each quantile and feature). On the other hand, for feature `neg`, we do not have a consistent gain of performance. Other issue to notice is that higher values of quantile q lead to lower values of PR-AUC. This make sense, because is harder to find the top 10% candidates with more votes than the half with move votes.



(a) $q = 0.5$



(b) $q = 0.9$

Figure 4.11: Precision Recall curves for different class thresholds quantiles q

We focused in quantiles 0.5 and 0.9, because they represent the most general case (predicting the half with more votes in each district) and the most specific case (predicting the top 10% most voted for each district) respectively. Figure 4.11 shows the Precision Recall curves for selected features and quantiles 0.5 and 0.9.

All the Twitter features shown in Figure 4.11a improved the classifier's precision. The best case is `degree` ($AP = 0.76$) followed by `pagerank` ($AP = 0.75$) and `favorite_count` ($AP = 0.74$). Even the worst case showed in the figure improves (`neg`,$AP = 0.68$). The gain of precision is higher when recall is lower. For high recall values, (close to 1), all curves tend to converge to the random classifier.

The case of Figure 4.11b is different. As it was shown in Figure 4.10, precision values for $q = 0.9$ are lower than for $q = 0.5$. For example, PR-AUC for baseline using $q = 0.5$ is $AP = 0.64$, and using $q = 0.9$ that value drops to $AP = 0.24$. A narrower positive class makes precision drop. Also, it is interesting that the gain of precision is higher in the range of recall (0.4, 0.9). Keep in mind that recall in both figures is different. A recall of 1 in figure A means identifying 50% of the sample, and the same recall in figure B means identifying 10% of the sample. Using the same reasoning, a recall of 0.1 in figure A mean identifying 5% of the sample, and the same recall in figure B means identifying 1% of the sample. As 1% of sample is smaller than 10 candidates, it seems that Twitter features do not have the power to identify so finely the top candidates.

Explained precision

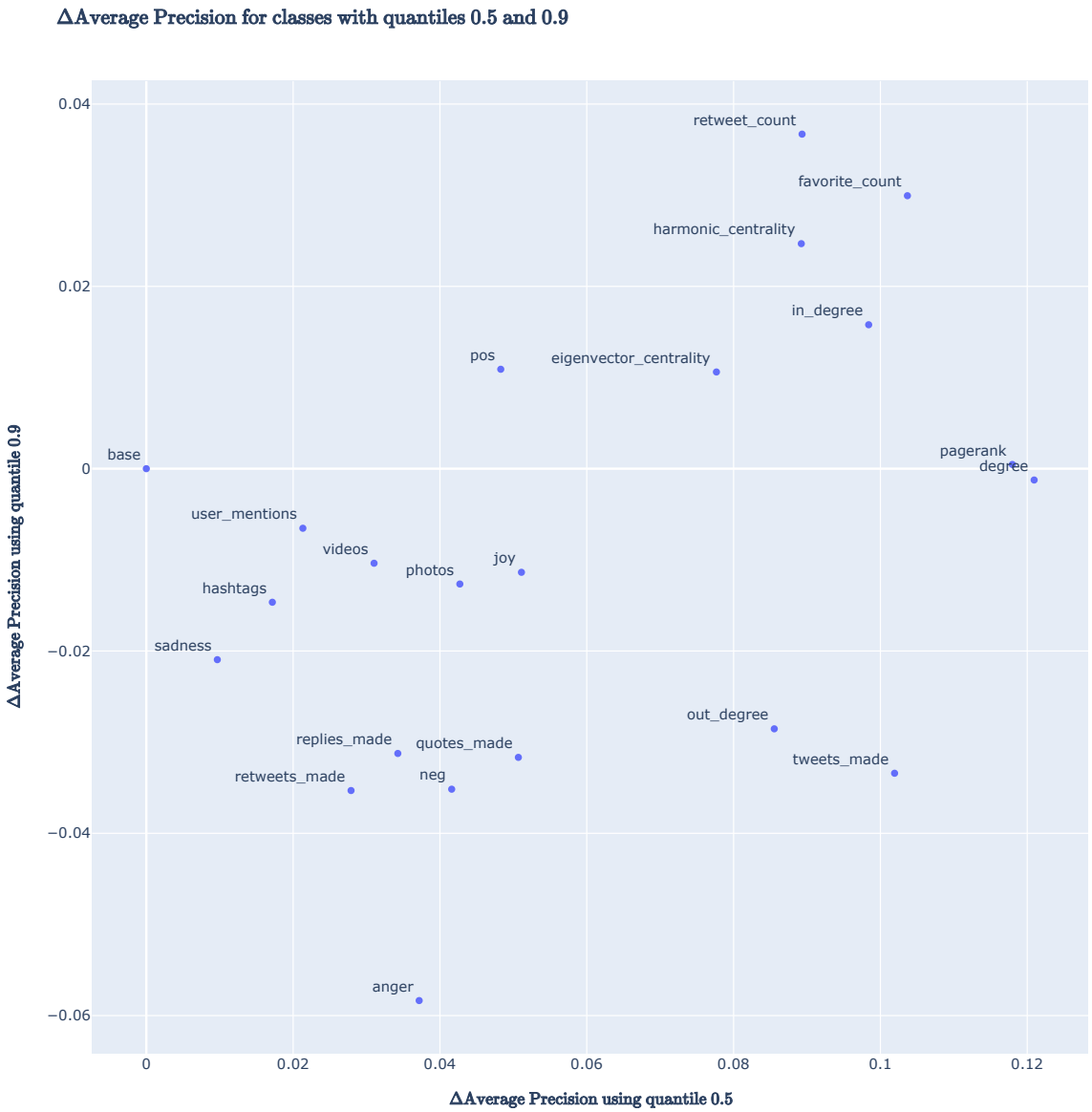


Figure 4.12: ΔAP score for each Twitter influence feature and threshold quantiles 0.5 and 0.9

Figure 4.12 shows the ΔAP for both thresholds quantiles 0.5 and 0.9. As we said earlier, ΔAP can take negative values in the case that adding the Twitter feature makes the model worse than the baseline. We see that for quantile 0.5, values for all features are positive, so all the features add precision to the classifier. The worst case is feature `sadness` with $\Delta AP = 0.01$, and the best case is feature `degree` with $\Delta AP = 0.12$. In contrast, for quantile 0.9 the majority of Twitter features decrease the classifier's precision. The best case is feature `retweet_count` with $\Delta AP = 0.04$.

Average precision of quantiles

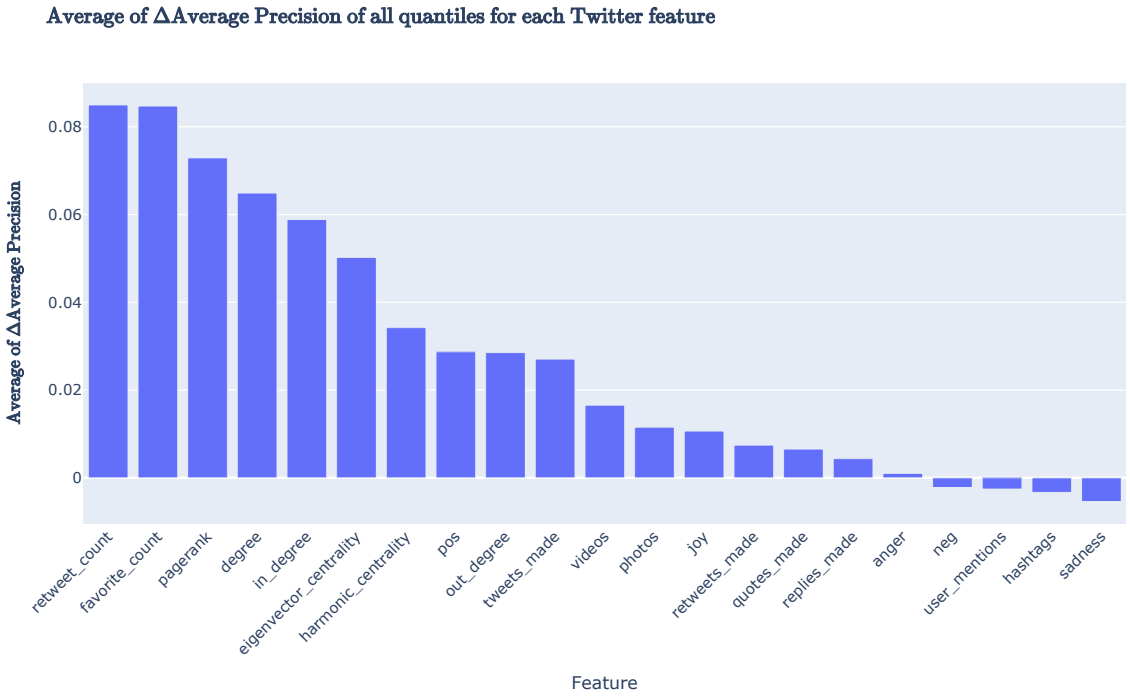


Figure 4.13: Average of ΔAP of all quantiles for each Twitter influence feature

Instead of taking each quantile individually, we average the results of all quantiles to have a single ΔAP for each feature. This is equivalent to taking the average of each curve in Figure 4.10 for the nine quantiles values in the Y axis. Figure 4.13 shows the value of quantile average ΔAP of each Twitter influence feature. The interpretation of this value is the precision that is explained by a single feature in average. For example, a value of 0.5 (or 50%) would mean that feature is responsible for increasing the precision of the classifier in 0.5 in average, regardless of the specific target class label chosen to make the classification.

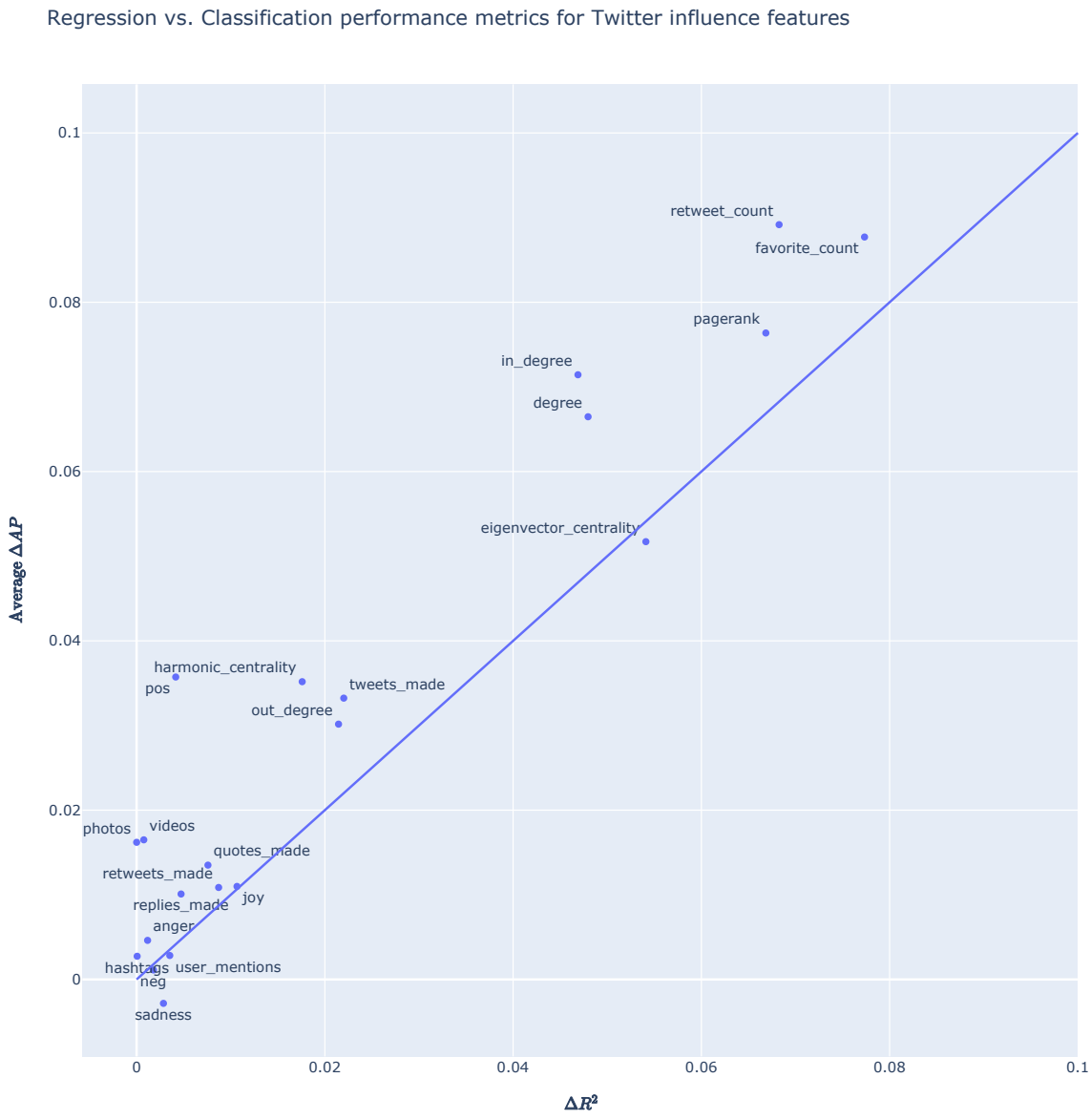


Figure 4.14: Regression performance (ΔR^2) vs. Classifier performance (Average of ΔAP) for each Twitter feature. Blue line is the identity function

ΔR^2 and ΔAP are similar metrics: both represent the percentage explained by a feature but in different ways. ΔR^2 is related to the variance of the votes that is captured by the feature. It cannot take negative values, because R^2 does not decrease with a bad feature (in the worst case, it remains constant).

ΔAP is the precision associated to a feature. It may take negative values for models worse than baseline. We would expect this metrics to be correlated, because a feature that explains more variance should lead to more precise predictions. To check this relation, Figure 4.14 shows the ΔR^2 and ΔAP for each Twitter feature. The two created metrics are consistent, correlated and have similar values, as shown by the identity function.

4.2 RQ2: Modified PageRank

We computed three modifications of PageRank as defined in Section 3.2.4. Several γ values were used, from 0 to 0.85. Figure 4.15 shows the distribution of the PageRank values for each modification and selected γ values. As expected, $\gamma = 0$ returned the same values for all the PageRank types, because that specific case is equivalent to the original PageRank formula.

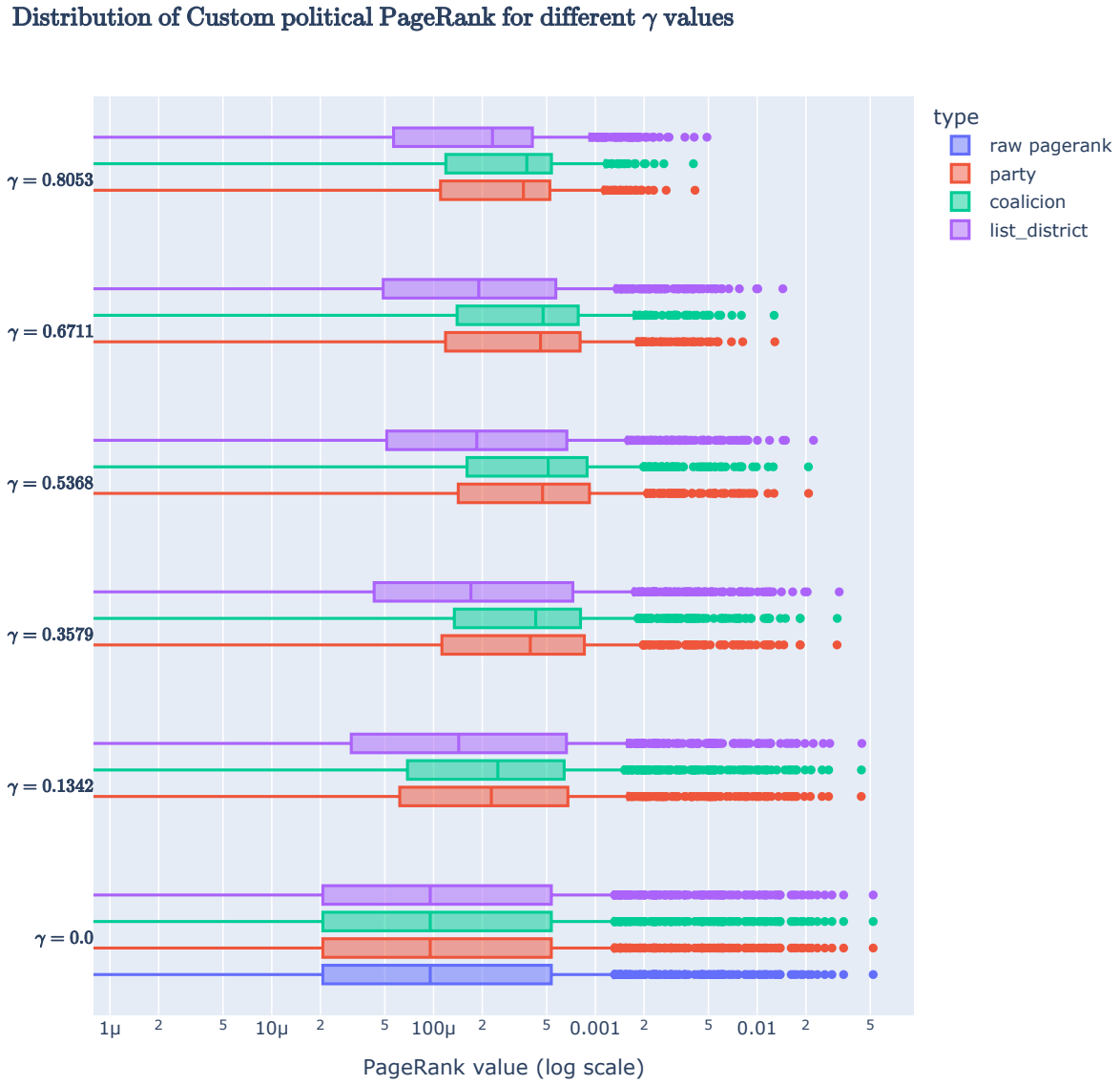


Figure 4.15: Distribution of Custom political PageRanks for selected γ values

It can be noticed that with a higher γ the PageRank has fewer outliers (less values to the right of of the upper fence of the boxplot). Also, when $\gamma > 0$ the PageRank median is higher comparing to $\gamma = 0$.

Both consequences of the PageRank modification (fewer outliers and higher median) were expected, as the modification was to transfer PageRank *mass* from the more influential candidates to candidates less known but politically connected (by party, district electoral list or political coalition) to relevant candidates.

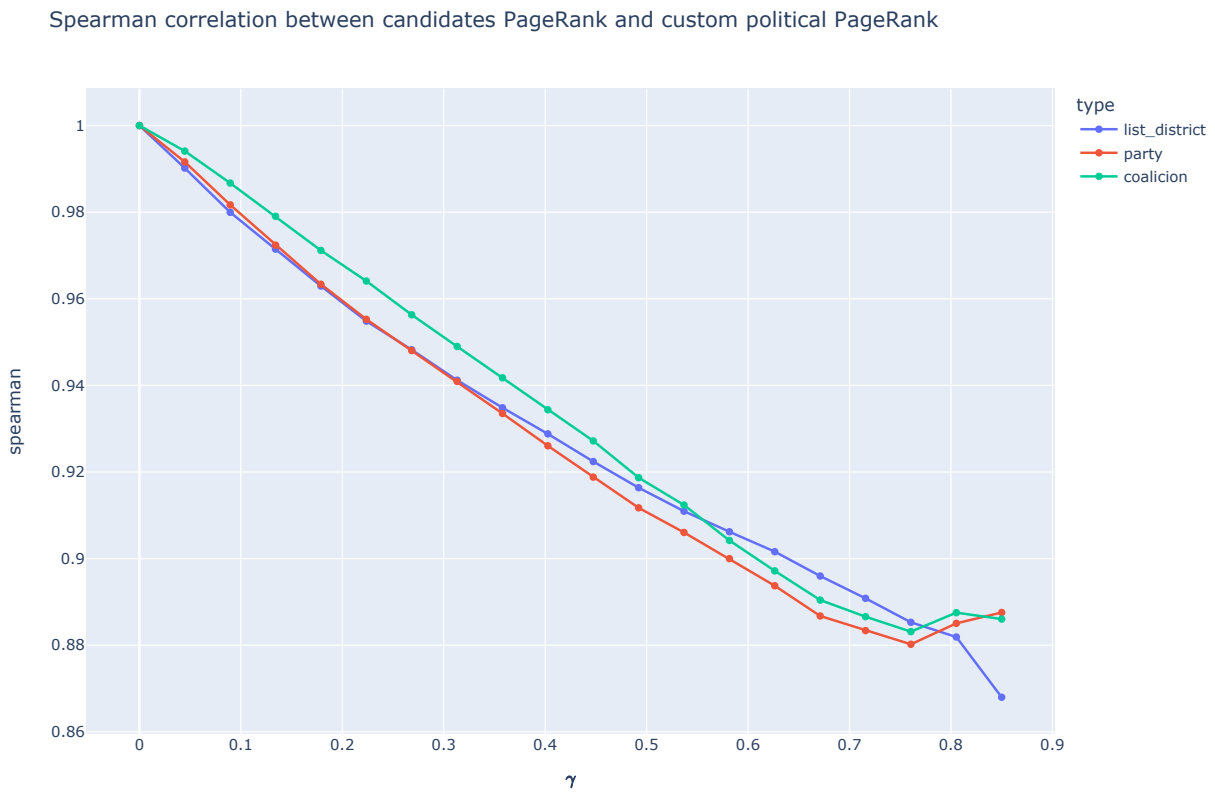


Figure 4.16: Correlation between PageRank and Custom political PageRanks

Figure 4.16 shows how correlated are the new modified PageRank features with the original PageRank. As can be observed, for all γ values the Spearman ρ is higher than 0.86. The data shows that the distribution shift (shown in figure 4.15) generated by the PageRank modification does not affect the order of the candidates, only the value, as the Spearman coefficient is related to the rank of the observations. Also, we observe that using higher values of γ generates lower correlation with the original PageRank, and that makes sense because with higher γ values, the original network connections are less important than the political connections.

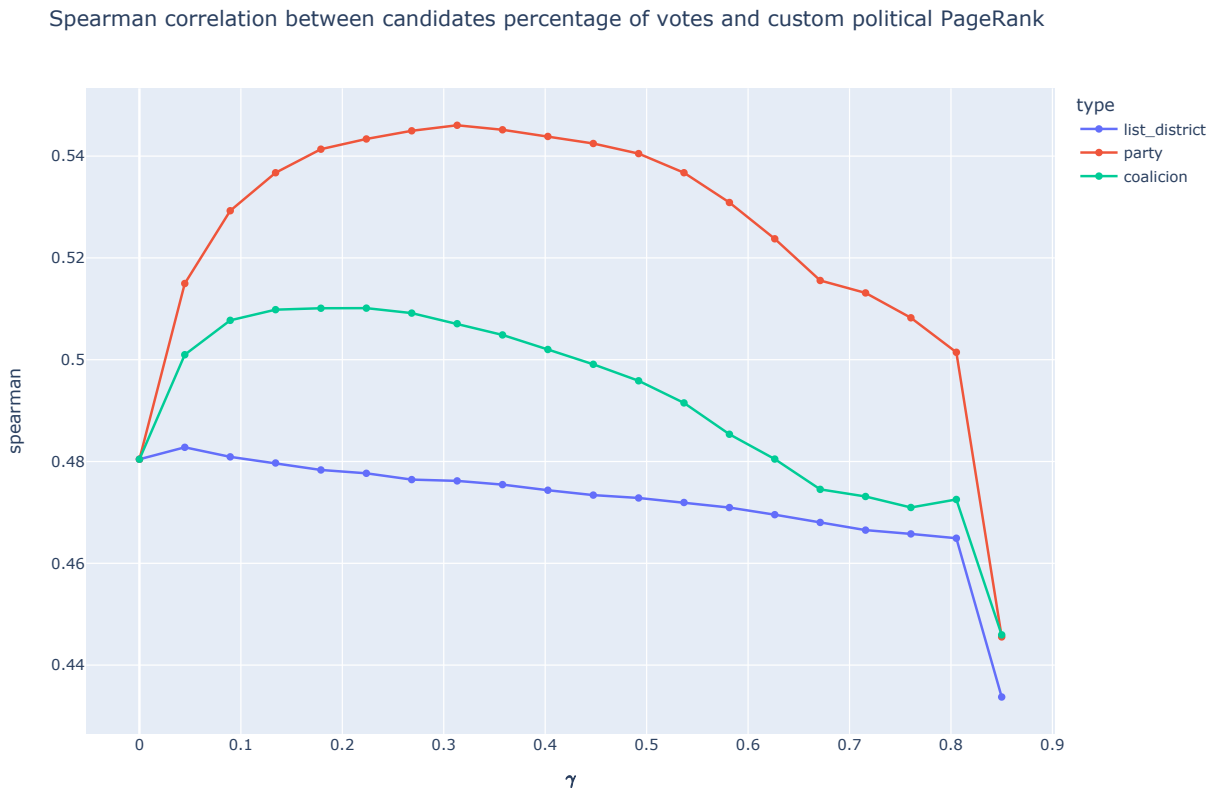


Figure 4.17: Correlation between electoral outcome and Custom political PageRanks

Figure 4.17 shows the correlation between the modified PageRank features and the vote district percentage of candidates. The point (0, 0.48) is equivalent to the original PageRank and match the Spearman value presented in Table 4.1. We observe that the difference with the original PageRank depend on the type of modification applied. The party-based PageRank had the best performance in terms of correlation, with a best case of Spearman $\rho = 0.546$ using $\gamma = 0.313$, outperforming the raw PageRank for every $\gamma < 0.85$. The political coalition-based PageRank also outperforms the original PageRank for every $\gamma < 0.65$, and a best case of Spearman $\rho = 0.510$ with $\gamma = 0.223$. The district list PageRank was the exception because for $\gamma > 0.10$ the correlation with votes decreases comparing to the original PageRank.

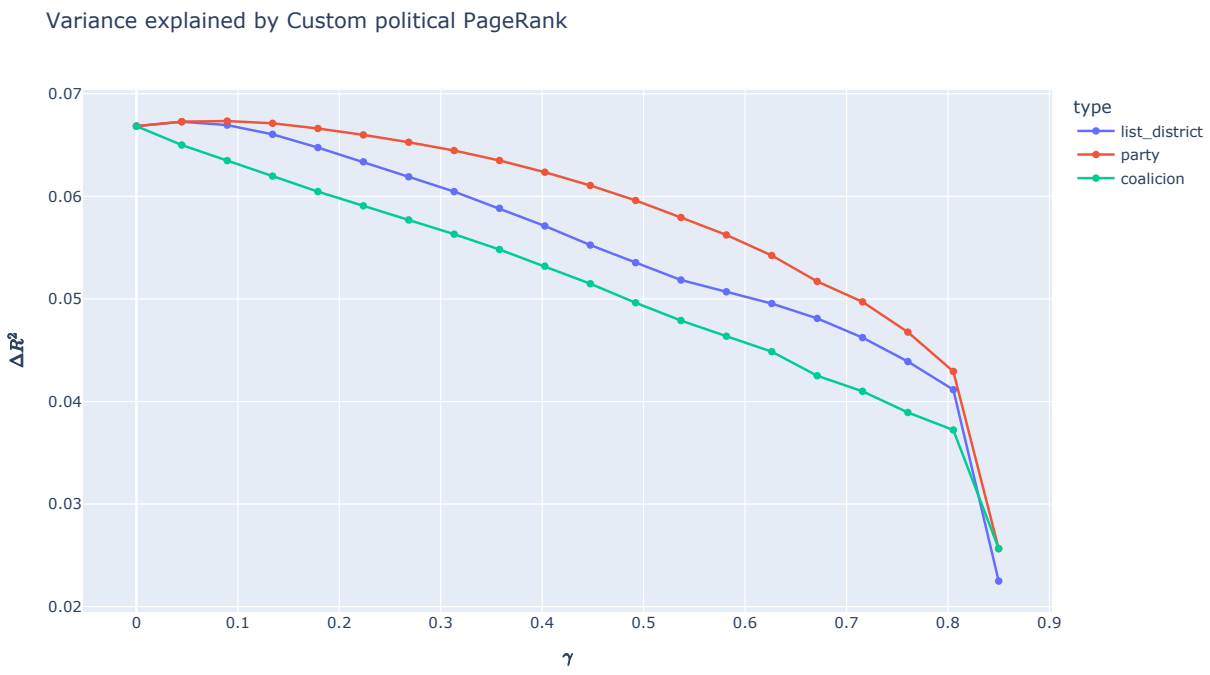
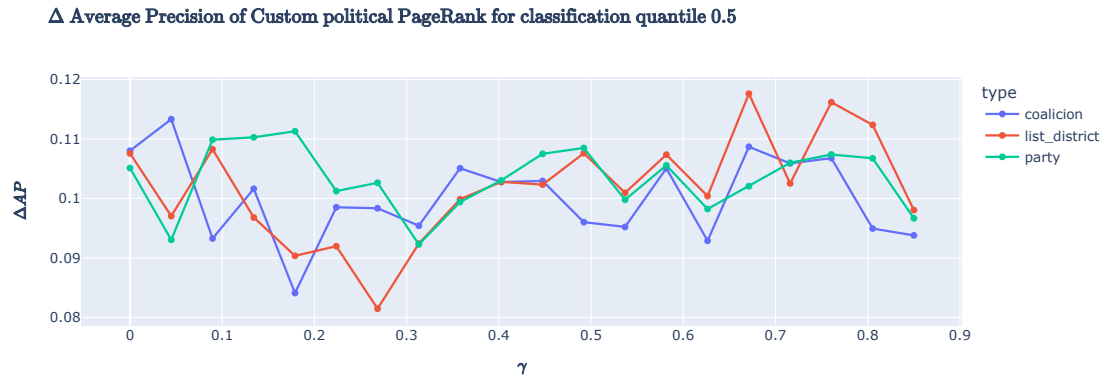
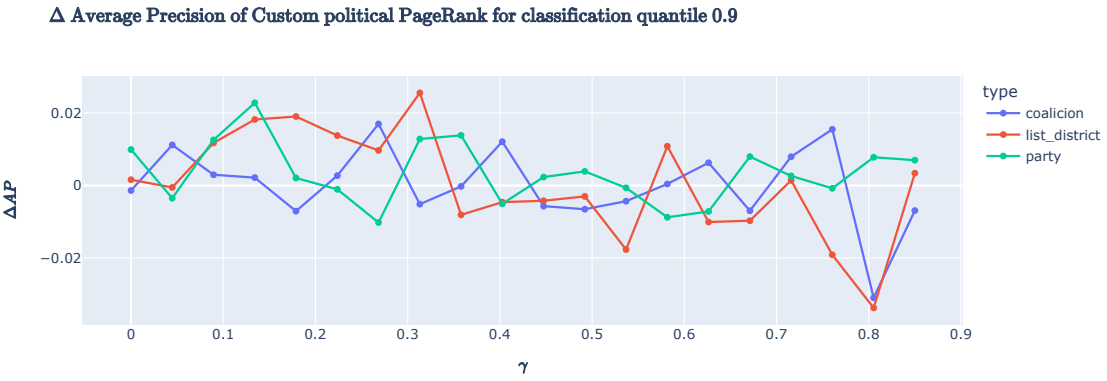


Figure 4.18: ΔR^2 for each Custom political PageRank

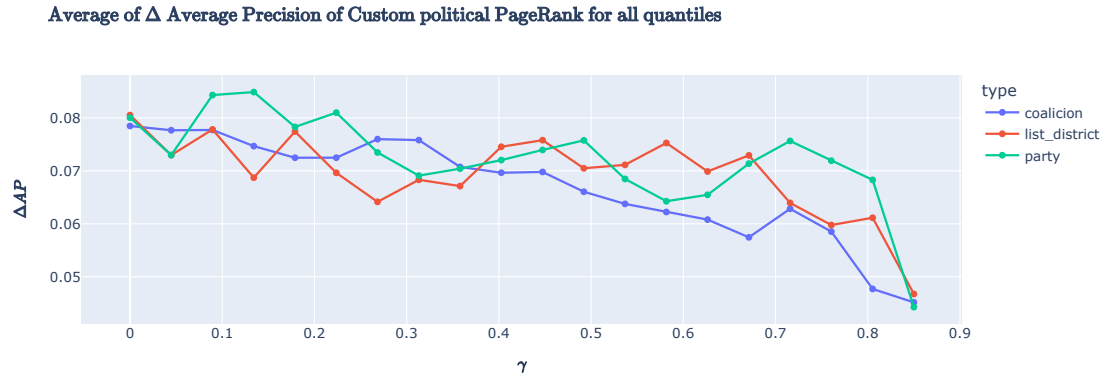
Using regressions, we computed the ΔR^2 for the modified PageRank to assess if this customization increases the explained variance of the features. Figure 4.18 shows the results. For all the γ values, the value of ΔR^2 achieved is worse than the original PageRank. The custom political PageRank explains less variance than the raw PageRank for all the modifications. The decrease of ΔR^2 is smaller for the party PageRank, followed by the list district PageRank, and finally the coalition PageRank had the most significant drop compared to the raw PageRank.



(a) ΔAP for $q = 0.5$



(b) ΔAP for $q = 0.9$



(c) Average ΔAP

Figure 4.19: Classifier performance of each Custom political PageRank feature

Finally, we used classification to evaluate the custom PageRank features, computing the ΔAP for each one and comparing to the original PageRank. Figure 4.19 shows ΔAP for several γ values. In Figure 4.19a we observe the case of classification quantile $q = 0.5$ and in Figure 4.19b the case of quantile 0.9. Both have in common that there is not a clear pattern

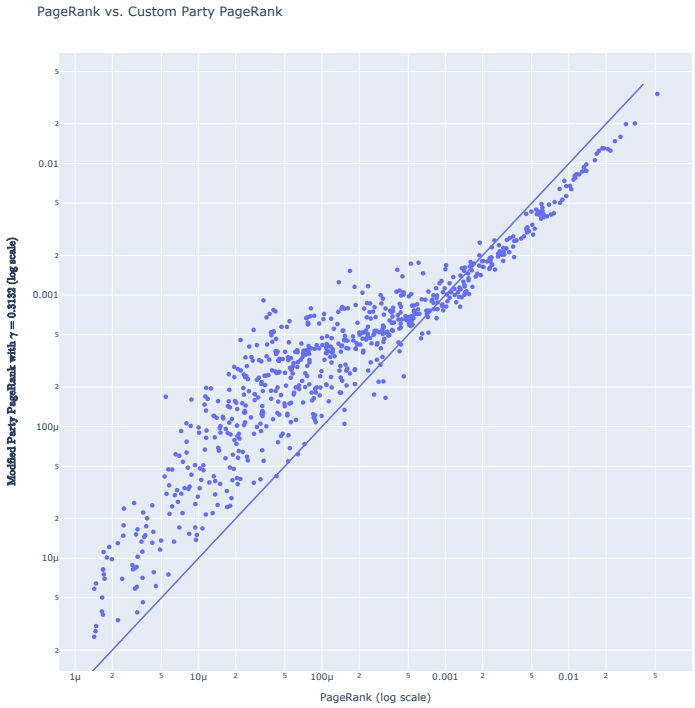
of an increase or decrease of ΔAP

Figure 4.19c shows the average of ΔAP for all quantiles. We noticed that, on average, ΔAP decreases when γ increases. Only three measurements from the party PageRank outperforms the raw PageRank, but the gain of performance is low (< 0.01) precision, so we did not consider it relevant.

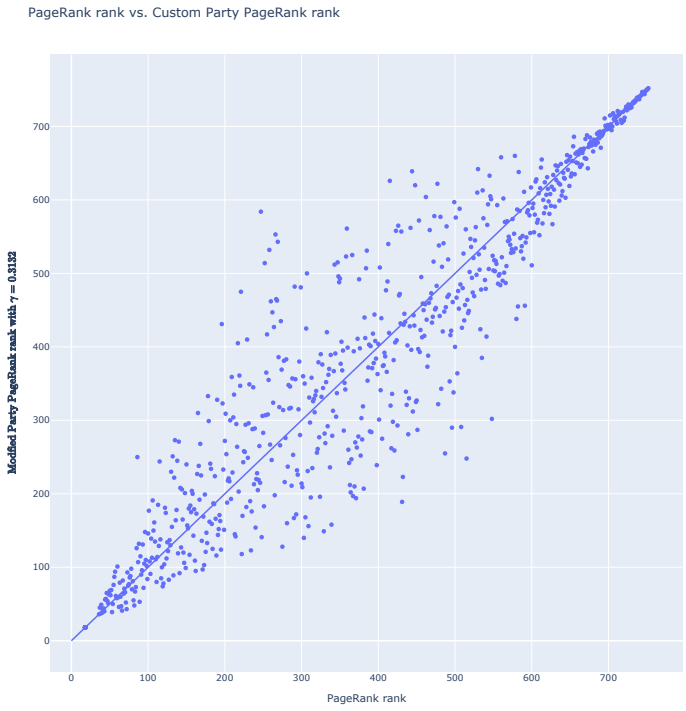
Also, notice that for $\gamma = 0$ in all subfigures from Figure 4.19, all PageRanks should be equal (and equal to the raw PageRank), like in the previous figures. That is not the case because of the stochastic component of the Random Forest classifier

Although it may seem contradictory, modified party and coalition PageRanks have higher correlation with votes but lower performance on regression and classification. Figure 4.20 contains a possible explanation. In Figure 4.20a, the x-axis is the raw PageRank and the y-axis is the modified party PageRank. The same effect noticed in Figure 4.15 is present here: modified PageRank takes *mass* from the highest values and transfer it to lower ones. This generates an increase in the sample median and the reduction in the amount of outliers.

Figure 4.20b tells a different history. It contains the same information that Figure 4.20a, but using the rank of the feature instead of the PageRank value. The lowest and highest values tend to mantain their positions, and the variation happens with the values in the middle. This variation in the rank is producing the little increase (< 0.07) in the Spearman coefficient shown in Figure 4.17, as Spearman is computed with the rank of the features. Nevertheless, regression and classification use the numeric value, so the correlation improvement does not translate into regression or classification.



(a) Raw features



(b) Rank of features

Figure 4.20: Comparison of raw PageRank and Political Party PageRank

Chapter 5

Conclusions

Using the result of the previous section, we propose the following conclusions to answer the research questions.

RQ1: To what extent does the influence on Twitter correlates with the votes obtained? What variables does this relationship depend on?

1. Twitter's explanatory power on elections is minimal but exists. Only 6 of the 21 Twitter influence feature achieved results higher than 5% in either ΔR^2 or the average of ΔAP for all quantiles:
   (a) `favorite_count`: $\Delta R^2 = 0.0773$, Average $\Delta AP = 0.0847$
   (b) `retweet_count`: $\Delta R^2 = 0.0682$, Average $\Delta AP = 0.0849$
   (c) `pagerank`: $\Delta R^2 = 0.0668$, Average $\Delta AP = 0.0729$
   (d) `degree`: $\Delta R^2 = 0.0479$, Average $\Delta AP = 0.0649$
   (e) `in_degree`: $\Delta R^2 = 0.0468$, Average $\Delta AP = 0.0588$
   (f) `eigenvector_centrality`: $\Delta R^2 = 0.0541$, Average $\Delta AP = 0.0502$

   Based on these results, an optimistic interpretation would be that Twitter's influence can explain itself $\sim$ 8% of the electoral results by itself. That is the case of `favorite_count`. A more pessimistic interpretation would be that Twitter influence explains at least $\sim$ 4% of the electoral results by itself, as is shown by the 6 features listed above.

   As all other features capture less than that, we conclude Twitter influence features capture a small fraction of the electoral phenomenon. But, again, this is consistent with other studies on elections [23] [4].

2. Positive tweets have a higher correlation with election results than negative tweets, as shown consistently by the correlation coefficient, regression and classification. As was seen in Results:
   (a) `pos`: Spearman $\rho = 0.3343$, $\Delta R^2 = 0.0041$, Average $\Delta AP = 0.0287$

(b) `joy`: Spearman $\rho = 0.3414$, $\Delta R^2 = 0.0106$, Average $\Delta AP = 0.0106$

(c) `sadness`: Spearman $\rho = 0.1508$, $\Delta R^2 = 0.0028$, Average $\Delta AP = -0.0054$

(d) `neg`: Spearman $\rho = 0.0918$, $\Delta R^2 = 0.0017$, Average $\Delta AP = -0.0022$

(e) `anger`: Spearman $\rho = 0.0466$, $\Delta R^2 = 0.0011$, Average $\Delta AP = -0.0010$

Positive and joy tweets have a stronger relationship with the electoral outcome than sad, negative or angry tweets for all the metrics proposed. For the classification, negative features even worsen the precision of the baseline.

3. The location of the candidates affects how much is the influence in Twitter related to the electoral outcome. Candidates from Metropolitan Region (capital) of Chile had a stronger relationship between centrality measures and vote percentage than the other candidates, as showed by the regression interaction term. The interaction term analysis also showed a gap between women and men in the correlation between the electoral outcome and Twitter influence, not only in centrality but also in positive tweets. The relationship between positive/joy tweets and electoral outcome is moderated by gender, with men having a higher correlation than woman.

RQ2: Is the political affiliation of a candidate (belonging to a political party, political coalition or electoral list) relevant for the Twitter influence? How does affiliation interact with Twitter?

1. The data did not show a significant increase on the performance of inference of electoral outcome (measured as Average Precision score) using the proposed modified PageRank features, even with an increase of correlation with votes in some cases, like party PageRank. This suggests we cannot reject the hypothesis that the political spillover effect does not exist. All modified PageRank features had equal or worse performance than the original PageRank, which led to the hypothesis that the affiliation information we injected into the network was already contained. Therefore, no new information was added, and there was no performance gain. This hypothesis is interesting and should be tested.

5.1 Description, Explanation or Prediction?

Hofman et al [17] proposes an scheme of different levels of empirical modelling with two main axes. The first axis is the focus, *Focus on specific features or effects* (to explain) or *Focus on predicting outcomes* (to predict). Second axis relates to intervention, *No intervention or distributional changes* or *Under interventions or distributional changes*.

Using that framework, this research belongs to the Descriptive modelling quadrant, this is *Focus on specific features or effects* and *No intervention or distributional changes*. This means our results are neither causal nor predictive but rather a measure of correlation, valid only for our sample.

How can we integrate explanation and prediction in this problem to achieve more robust results? Predictive and causal approaches need to be included.

To have a predictive model, we need to validate the model results using data from a similar future election (e.g. the Chilean Parliament elections of 2021, 6 month after the Constitutional Convention election). Validate the result of a regression or classifier using data from the same election counts as an inference, but it is not a prediction.

To achieve a causal explanation, a counterfactual (e.g. What happens if Twitter does not exist?) must be set to assess the causal effect of an intervention, e.g., having Twitter in an election. This is hard, because it requires imagining a world where Twitter does not exist and trying to estimate the results of the election in that world.

Correlation does not imply causation. This is a statistical mantra used to remember that causal effects cannot be inferred from the facts two variables grow together. However, causation may imply correlation. As stayed in the Reichenbach's Common Cause Principle: *if two events are correlated, then either there is a causal connection between the correlated events that is responsible for the correlation or there is a third event, a so called (Reichenbachian) common cause, which brings about the correlation* [33].

We proved that there is a statistical relation between Twitter features (the strength of the relation depends on the feature) and vote percentage of candidates. We could assume a latent variable such as the public knowledge (PK) of a candidate, and attribute both the election outcome and the Twitter metrics as an effect of PK. Structural Causal Models, Graphical model or other tools to model these causal relationships may be helpful for such modeling in future research.

Bibliography

[1] Andrés Abeliuk, Daniel M Benjamin, Fred Morstatter, and Aram Galstyan. Quantifying machine influence over human forecasters. *Scientific reports*, 10(1):1–14, 2020.

[2] Tomás Alegre Sepúlveda and Brian Keith Norambuena. Twitter sentiment analysis for the estimation of voting intention in the 2017 chilean elections. *Intelligent Data Analysis*, 24(5):1141–1160, 2020.

[3] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[4] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[5] Cadem. El futuro de los medios. September 2020.

[6] Alejandro Cárdenas, Carlos Ballesteros, and René Jara. Redes sociales y campañas electorales en Iberoamérica. Un análisis comparativo de los casos de España, México y Chile. *Cuadernos. info*, (41):19–40, 2017.

[7] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*, 2020.

[8] Nian Shong Chok. *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. PhD thesis, University of Pittsburgh, 2010.

[9] Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology press, 2014.

[10] Hercules Dalianis. *Evaluation Metrics and Evaluation*, pages 45–53. 05 2018.

[11] Christine P Dancey and John Reidy. *Statistics without maths for psychology*. Pearson education, 2007.

[12] Shainen M Davidson and Kenton White. Forecasting Seat Counts in the 2019 Canadian Federal Election Using Twitter. In *Canadian Conference on Artificial Intelligence*, pages 151–162. Springer, 2020.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Michael Gallagher. Proportionality, disproportionality and electoral systems. *Electoral studies*, 10(1):33–51, 1991.

[15] Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. How representative is an abortion debate on Twitter? In *Proceedings of the 10th ACM Conference on Web Science*, pages 133–134, 2019.

[16] Inge S Helland. On the interpretation and use of R2 in regression analysis. *Biometrics*, pages 61–69, 1987.

[17] Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, et al. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188, 2021.

[18] Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*, 2013.

[19] René Jara, Antoine Faure, Jarnishs Beltrán, and Gonzalo Castro. The Political Awareness in the candidates using Twitter. A clusterization exercise for the municipal elections in Chile (2016). *Revista Latina de Comunicación Social*, 72:803, 2017.

[20] Jeff Jauregui. Math 312: Markov chains, Google's PageRank algorithm. `https://www2.math.upenn.edu/~kazdan/312F12/JJ/MarkovChains/markov_google.pdf`, 2012. Accessed: 2022–03-26.

[21] Gary King. How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, pages 666–687, 1986.

[22] Jasmin Komić. *Harmonic Mean*, pages 622–624. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[23] Sanne Kruikemeier. How political candidates use Twitter and the impact on votes. *Computers in human behavior*, 34:131–139, 2014.

[24] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. Towards unbiased BFS sampling. *IEEE Journal on Selected Areas in Communications*, 29(9):1799–1809, 2011.

[25] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636, 2006.

[26] Guichong Li. Sampling graphical networks via conditional independence coupling of Markov chains. In *Canadian Conference on Artificial Intelligence*, pages 298–303. Springer, 2016.

[27] Nelly Litvak, Werner RW Scheinhardt, and Yana Volkovich. In-degree and PageRank: why do they follow similar power laws? *Internet mathematics*, 4(2-3):175–198, 2007.

[28] Jun MA, Yuexiong Ding, Jack Cheng, Yi Tan, Vincent Gan, and Jingcheng ZHANG. Analyzing the leading causes of traffic fatalities using xgboost and grid-based analysis: A city management perspective. *IEEE Access*, PP:1–1, 10 2019.

[29] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252, 1996.

[32] Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, 2021.

[33] Miklós Rédei. Reichenbach's common cause principle and quantum correlations. In *Non-locality and Modality*, pages 259–270. Springer, 2002.

[34] Bruno Ribeiro and Don Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 390–403, 2010.

[35] Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on Twitter: A survey. *Information processing & management*, 52(5):949–975, 2016.

[36] Sebastián Rodríguez, Héctor Allende-Cid, Wenceslao Palma, Rodrigo Alfaro, Cristian Gonzalez, Claudio Elortegui, and Pedro Santander. Forecasting the Chilean electoral year: Using Twitter to predict the presidential elections of 2017. In *International Conference on Social Computing and Social Media*, pages 298–314. Springer, 2018.

[37] Hernando Rojas and Sebastián Valenzuela. A call to contextualize public opinion-based research in political communication. *Political Communication*, 36(4):652–659, 2019.

[38] Pedro Santander, Claudio Elórtegui, Cristián González, Héctor Allende-Cid, and Wenceslao Palma. Redes sociales, inteligencia computacional y predicción electoral: el caso de las primarias presidenciales de Chile 2017. *Cuadernos.info*, pages 41 – 56, 12 2017.

[39] Salomé Sola-Morales and Paula Flores. Twitter y las Elecciones Presidenciales 2013 en Chile. *Diálogos de la comunicación*, (91):12, 2015.

[40] Leah F South, Marina Riabiz, Onur Teymur, Chris Oates, et al. Post-Processing of MCMC. *arXiv preprint arXiv:2103.16048*, 2021.

[41] Shazia Tabassum, Fabiola SF Pereira, Sofia Fernandes, and João Gama. Social network analysis: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5):e1256, 2018.

[42] Universidad Central de Chile (UCEN). Presentación Observatorio de Política y Redes Sociales. `https://www.ucentral.cl/observatorio-politica-y-rrss/presentacion-observatorio-de-politica-y-redes-sociales` (visited: 2022-05-17).

[43] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067, 2019.

[44] Lakhana Watthanacheewakul. Transformations for Left Skewed Data. In *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2021*, pages 101–106, 2021.

ANNEXED

Example tweet from Twitter API


```
{
    "_id": {
      "$oid": "60ced9df9c7dd09e146307c5"
    },
    "created_at": "Thu Apr 08 12:12:58 +0000 2021",
    "id": {
      "$numberLong": "1380131552600985603"
    },
    "id_str": "1380131552600985603",
    "full_text": "RT @CNNChile: FMI propone un impuesto temporal a
        ricos para financiar necesidades derivadas de la pandemia
        https://t.co/6MkP77Gzxi",
    "truncated": false,
    "display_text_range": [
      {
        "$numberInt": "0"
      },
      {
        "$numberInt": "130"
      }
    ],
    "entities": {
      "hashtags": [],
      "symbols": [],
      "user_mentions": [
        {
          "screen_name": "CNNChile",
          "name": "CNN Chile",
          "id": {
```

```
            "$numberInt": "18248645"
          },
          "id_str": "18248645",
          "indices": [
            {
              "$numberInt": "3"
            },
            {
              "$numberInt": "12"
            }
          ]
        }
      ],
      "urls": [
        {
          "url": "https://t.co/6MkP77Gzxi",
          "expanded_url": "https://www.cnnchile.com/mundo/fmi-propone-
            impuesto-temporal-ricos_20210407/",
          "display_url": "cnnchile.com/mundo/fmi-prop...",
          "indices": [
            {
              "$numberInt": "107"
            },
            {
              "$numberInt": "130"
            }
          ]
        }
      ]
    },
    "source": "<a href=\"http://twitter.com/download/android\"
        rel=\"nofollow\">Twitter for Android</a>",
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
      "id": {
        "$numberLong": "1187042015382507520"
      },
      "id_str": "1187042015382507520"
    },
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
```

```
    "retweeted_status": {
      "created_at": "Wed Apr 07 20:17:52 +0000 2021",
      "id": {
        "$numberLong": "1379891192314351616"
      },
      "id_str": "1379891192314351616",
      "full_text": "FMI propone un impuesto temporal a ricos para
        financiar necesidades derivadas de la pandemia
        https://t.co/6MkP77Gzxi",
      "truncated": false,
      "display_text_range": [
        {
          "$numberInt": "0"
        },
        {
          "$numberInt": "116"
        }
      ],
      "entities": {
        "hashtags": [],
        "symbols": [],
        "user_mentions": [],
        "urls": [
          {
            "url": "https://t.co/6MkP77Gzxi",
            "expanded_url": "https://www.cnnchile.com/mundo/fmi-propone-
                impuesto-temporal-ricos_20210407/",
            "display_url": "cnnchile.com/mundo/fmi-prop...",
            "indices": [
              {
                "$numberInt": "93"
              },
              {
                "$numberInt": "116"
              }
            ]
          }
        ]
      },
      "source": "<a href=\"https://about.twitter.com/products/tweetdeck
        \" rel=\"nofollow\">TweetDeck</a>",
      "in_reply_to_status_id": null,
      "in_reply_to_status_id_str": null,
      "in_reply_to_user_id": null,
      "in_reply_to_user_id_str": null,
      "in_reply_to_screen_name": null,
      "user": {
```

```
        "id": {
          "$numberInt": "18248645"
        },
        "id_str": "18248645"
      },
      "geo": null,
      "coordinates": null,
      "place": null,
      "contributors": null,
      "is_quote_status": false,
      "retweet_count": {
        "$numberInt": "1640"
      },
      "favorite_count": {
        "$numberInt": "3324"
      },
      "favorited": false,
      "retweeted": false,
      "possibly_sensitive": false,
      "lang": "es"
    },
    "is_quote_status": false,
    "retweet_count": {
      "$numberInt": "1640"
    },
    "favorite_count": {
      "$numberInt": "0"
    },
    "favorited": false,
    "retweeted": false,
    "possibly_sensitive": false,
    "lang": "es",
    "datetime": {
      "$date": {
        "$numberLong": "1617883978000"
      }
    }
  }
```


Listing 1: Tweet object in BSON (MongoDB) format extracted from Twitter API