Chronicle of an Impeachment Foretold? Exploring the Reasons Why Peruvians Want Castillo Out

Alessandra Oshiro

4/7/2022

1. Context and Research Question

In 2021, Peru had presidential elections. The results of the tense electoral process were an accurate snapshot of the polarized political landscape of the country. Amid accusations of fraud, the conservative left-wing candidate Pedro Castillo was declared the winner, with an advantage of less than 1% over Keiko Fujimori, the conservative right-wing candidate. With only half of the country having voted for him, it is not surprising that Castillo has had a rough time being president. On the contrary, some sectors of the Peruvian population has been expressing their desire for Castillo's impeachment ever since he was declared the winner of the electoral race. Some of these demands have been encouraged by a fear campaign which emphasized Castillo's communist ideology, capitalizing on Peru's traumatic past with left-wing guerrillas and the experiences of countries like Venezuela. Nevertheless, after his first semester in power, it is also likely that these claims are based on a discontent with Castillo's performance or other events in Peruvian politics.

Given that the hopes for an impeachment is far from extinguished, this project aimed at exploring the main reasons why (some) Peruvians demand Pedro Castillo's impeachment. For this, the analysis will consist in identifying the most common topics raised by supporters of Castillo's impeachment, and assessing how present those topics are among those circles. Additionaly, given that a motion of impeachment was unexpectedly debated by the parliament when this project was being done, it will also include a comparison of the topic prevalence before and after the impeachment debate.

2. Data Source and Assessment

This project used Twitter data. A main reason to use social media was that, in contrast to newspapers, users do not need to be careful of stating strong value judgements, nor from using insults, slang, and other useful informal elements that are rarely seen in more serious publications. Moreover, by being open to anyone with internet connection, Twitter hosts more diverse opinions than what can be found in Peruvian traditional media. Certainly, social media is also more prone to fake information. Nevertheless, this was not an issue because the project required data about what Peruvians *believe* about Castillo's performance or politic ideology, rather than objective information about it.

Another shortcoming of Twitter is that certain populations might be overrepresented. As in other online platforms, the participation of young, male, and educated users is more common. Moreover, in the case of Peru, most of the tweets most likely come from Lima, the capital city, in contrast to the rural areas in the inner parts of the country. For a more accurate analysis, it would have been ideal to be able to control for factors like sex and gender, as well as to restrict the sample to certain areas or even districts to deal with the overrepresentation of some sectors. Nevertheless, this information is, unfortunately, not available in most tweets. Therefore, this issue needs to be taken into account when drawing conclusions from the data.

3. Data Collection

The data collection was done with the Twitter API and the `search_tweets` function from the `rtweet` package. To retrieve tweets that were unambiguous about their *support for Castillo's impeachment*, the query included explicit hashtags: #VacanciaPresidencialYa, #VacanciaPedroCastilloYa, #VacanciaPedroCastillo, #CastilloRenunciaYa, #FueraCastillo. An additional advantage of using these hashtags for scraping is that they are most likely exclusive to Peruvian Twitter users, which is helpful for dealing with the lack of geolocation data.

The project involved two tweet-scrapping periods, both using the same code and excluding retweets. The first one, which was the original plan, consisted in scrapping the tweets daily for a month (February, 15th to March, 15th). Initially, the code was run three times a day, requesting 500 tweets each time. Then, after receiving feedback from the class, the strategy was changed to running the code once a day, but requesting 5000 tweets. This second approach worked better for expanding the dataset, as more unique tweets remained after eliminating duplicates. As a result, this first dataset (`impeachment`) has 31562 observations, and includes 28 days (the APi could not find tweets for February, 16th for unknown reasons). The number of tweets per day is highly unbalanced (the plot can be found in the script), but that does not affect the analysis.

The second tweet-scrapping period was not part of the original plan, but was included after the last presentation in class, based on the feedback of my classmates. On March, 28th, the impeachment of Castillo was debated by the parliament, before whom he had to present a defense. Therefore, I was able to retrieve 7627 tweets from March, 26th to March, 30th. This second dataset (`post_impeachment`) covers the period of two days before the impeachment debate, the day of the impeachment debate, and two days after the impeachment debate. This was useful for comparing how the topics raised by those who wished for Castillo's impeachment before and after the impeachment debate happened. Certainly, it would have been better to have an uninterrupted dataset, with tweets from February, 15th to March, 30th. Nevertheless, this was not possible because of the restriction to retrieve historical tweets, and the fact that this second dataset was not really planned for.

4. Analytical Set-up

The set-up for the analysis included two different text analysis methods, as well as the standard data cleaning process that they require. For the first part of the analysis (i.e., identifying the main reasons behind the desire of Castillo's impeachment), LDA topic modeling was performed on the first dataset, which includes the tweets from February, 15th to March, 15th. Next, for the second part (i.e., calculating the daily topic proportion), a dictionary-approach was used on both the first and the second (March, 26th to March, 30th) dataset. The reason to choose these methods (and in that order) was that, with the LDA topic modeling, I would be able to get lists of most discriminant terms per topic. Then, these list would be used to calculate and compare the daily topic proportion before (February, 15th to March, 15th) and around (March 26th to 30th) the impeachment debate. Now, I give some details about each step, before discussing the results.

First, for the data cleaning process, most of the standard steps for text analysis were followed. The only exception was stemming, as it gave confusing and inconsistent results for Spanish words. However, the pre-processing did include converting all the text to lower case, as well as removing mentions, URLs, emojis, punctuation, and other irrelevant elements. Additionally, all accented letters were transformed into non-accented ones to avoid excluding spelling mistakes related to accentuation, which is common in Spanish-speaking social platforms. Next, the hashtags used as keywords were removed, as well as pre-made list of stop-words, and a custom one (`removable_words`). The `removable_words` list was constantly updated based on the general words (e.g., articles, some verbs, "jajaja", etc.) that came up in the topic modeling. Finally, all NAs and empty strings were dropped.

Then, for the first part of the analysis, `tm`'s `LDA` function was used to perform topic modeling on the `impeachment` dataset. Given that the method requires defining the number of clusters, k was set to 3. This choice was motivated by my expectations based on the Peruvian context: in the public discourse (at least before coming to Vienna), the issues being discussed in relation to the desire for Castillo's impeachment were 1) his communist ideology, 2) his incompetence for the role, and 3) corruption, which

is a very common theme in Peruvian politics. The beta scores resulting from the LDA model would be later used to build a list of most frequent terms per topic, as well as identifying the most discriminant terms through a pair-wise comparison of the beta-scores. For that, the filter was set to .0025 instead of the .001 used in class, as it was better as providing more meaningful terms. A last thing to mention about this first part of the analysis is that, To identify the reasons behind Peruvians' support for Castillo's impeachment, an unsupervised method is better than a supervised one because it allows for unexpected outcomes. With that, I mean that the topics that arise are not limited to our previous expectations of what to find. This was particularly useful given that, despite having some rough expectations based on my familiarity with the Peruvian political context, I was curious to find out whether other new narratives had become popular.

Finally, the second part of the analysis consisted in a dictionary approach. My intention was to use the lists resulting from the LDA topic modeling to calculate the daily topic proportion in both `impeachment` and the `post_impeachment` dataset to look for variations within and between them. For this calculation, a new binary variable would be created for each topic. Using the LDA lists as dictionaries, each tweet would be coded 1 if such topic was present (i.e., the tweet included any topic keyword), and 0 otherwise. It is important to note that this approach implies that the topics are not mutually exclusive, therefore, the proportion of topics had to be calculated separately. With this, I mean that the total number of daily tweets for each topic was calculated and divided by the total number of tweets for that day. Because of this, the sum of the topic proportions per day exceeds 100%.

5. Results

The nature of the research question allows for many different plots. Nevertheless, in the report I include only those which are the most relevant to illustrate the results of the analysis. The other plots can be seen by running the code provided in the script.

5.1 LDA Topic Modeling

As mentioned in the previous section, I expected the LDA topic modeling to provide lists of most discriminant terms, so that I could then use them as the dictionaries for the dictionary-approach. Nevertheless, the model was only partially successful in that regard. The main issue was that the resulting lists of most frequent, and more discriminant terms did not show well defined topics. With this, I mean that all the three topics included terms that referred to the same variety of topics. The difficulty can be better understood by looking at *Figure 1*, which shows the most discriminant topics between topic 1 and topic 2. Both include words related to protest and mobilization: "salgamostodos", "5mperu", "tomemoslacalle", "marchar" in topic 1; "20mperu", "marcha" in topic 2. Similarly, both include terms related to corruption and criminality: "criminal", "corrupcion", in topic 1; and "dinamicos", "ninos", "karelim" (names of groups and people involved in corruption scandals) in topic 2. The same outcome can be found in the list of most frequent terms per topic, and the other most discriminant terms plots.

Despite not providing well defined lists, the LDA topic modeling was still very useful for identifying other narratives behind the desire for Castillo's impeachment. On one hand, it helped me confirm that the three previously assumed topics (i.e., fear of communism, incompetence, and corruption) were present in the Twitter debate among the supporters of Castillo's impeachment. On the other, the appearance of unexpected words in the topic modeling evidenced the presence of topics that I was not aware of before the analysis. These were three: patriotism, democracy, and mobilization (relevant terms for this last one can be seen in *Figure 1*). In addition to that, by updating the `removable_words` list and rerunning the model several times, I was able to manually build dictionaries for each of the six topics. Therefore, I believe that, while the results of the LDA model were not the expected ones, they were still crucial for the completion of the project.

5.2 Dictionary Approach

As explained in the previous section, calculating the daily presence of each topic was straightforward, once the dictionaries were created. Admittedly, not all the topics show interesting trends. The plots for

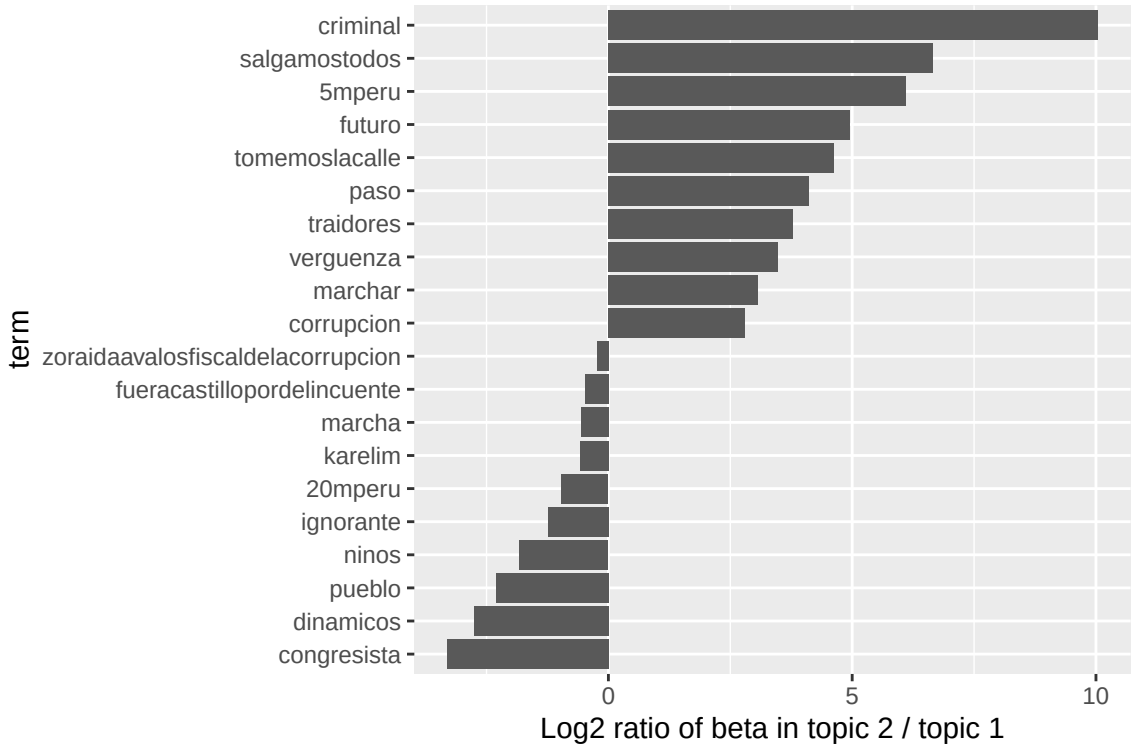

Figure 1: *Most Discriminant Terms Betweent Topic 1 and Topic 2*

each topic, both from the `impeachment` and the `post_impeachment` datasets can be seen in the script. Here, I include the plots of two topics that show interesting trends in the `impeachment` dataset, that is, the period between February, 15th and March, 15th. First, the left subplot in *Figure 2* shows how the proportion of tweets about corruption increase on February, 27th. This can be traced to the National Comptroller's announcement of the beginning of an investigation regarding a corruption scandal in which Castillo has been involved. Similarly, the right subplot in *Figure 2* presents a abrupt increase in the proportion of tweets about "mobilization". This peak can also be traced to a particular event. In March, 5th there was a big mobilization to demand the impeachment of Pedro Castillo. This explains the big presence of words related to social mobilization and protest in the LDA topic modeling.

The mobilization topic is also interesting, however, when doing the comparison between the tweets before the impeachment debate (`impeachment` dataset), and those around the debate (`post_impeachment` dataset). *Figure 3* shows the average daily percentage of each topic in each dataset. It is interesting to see that, while the tweets between February, 15th to March, 15th (`impeachment` dataset) had show a high percentage of the "mobilization" topic, the tweets between March, 26th - 30th (`post_impeachment` dataset) show a minimal presence of such topic. Naturally, this can be explained by the particular event mentioned above.

As more general trends, it is also interesting to note that, contrary to what I had expected, the topic of corruption is more prevalent than that of communism in both periods. Moreover, as *Figure 4* shows, it seems to have peaked during the day of the impeachment debate (March, 28th) together with the topic of incompetence. Nevertheless, as *Figure 3* illustrates, the issue of Castillo's communist ideology is always present as a reason for people to justify his impeachment.

6. Conclusions

This project aimed at exploring the reasons behind Peruvians' demands for the impeachment of Pedro Castillo. As shown in the previous sections, the text analysis was able to deliver some answers to the research question. In addition to the three topics I expected to find (i.e., fear of communism, incompetence, and corruption), three smaller ones were identified (i.e., mobilization, patriotism, and democracy). The time series analysis was useful to explain certain trends in the variation in the daily prevalence of

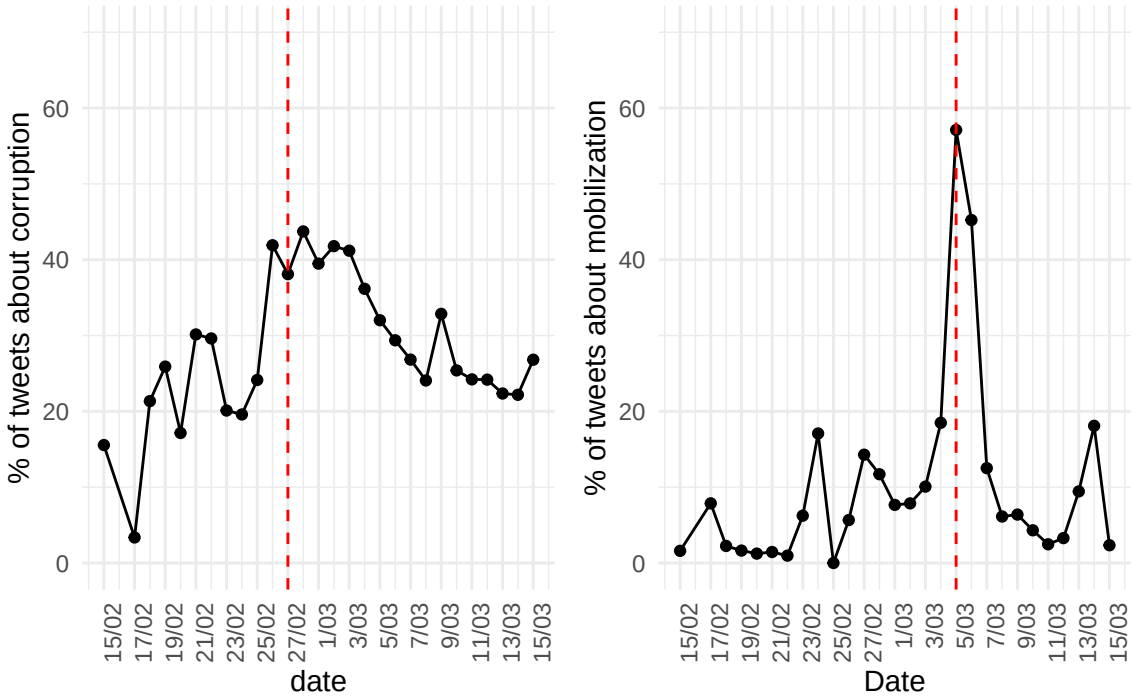

Figure 2: *Daily Percentage of Corruption (Left) and Mobilization (Right) Topics Between February, 15th and March, 15th. The announcement of the investigation (Left) and the big protest (Right) are signaled with the red line.*

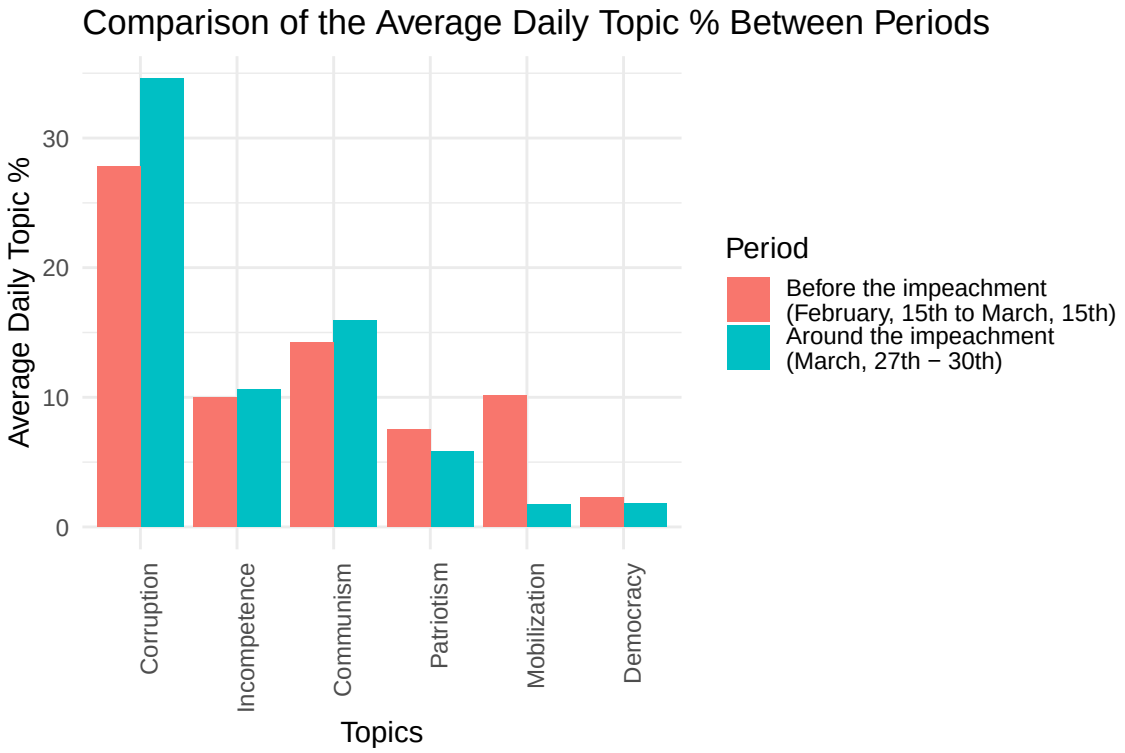

Figure 3: *Comparison of the Average Daily Topic % Between Periods*

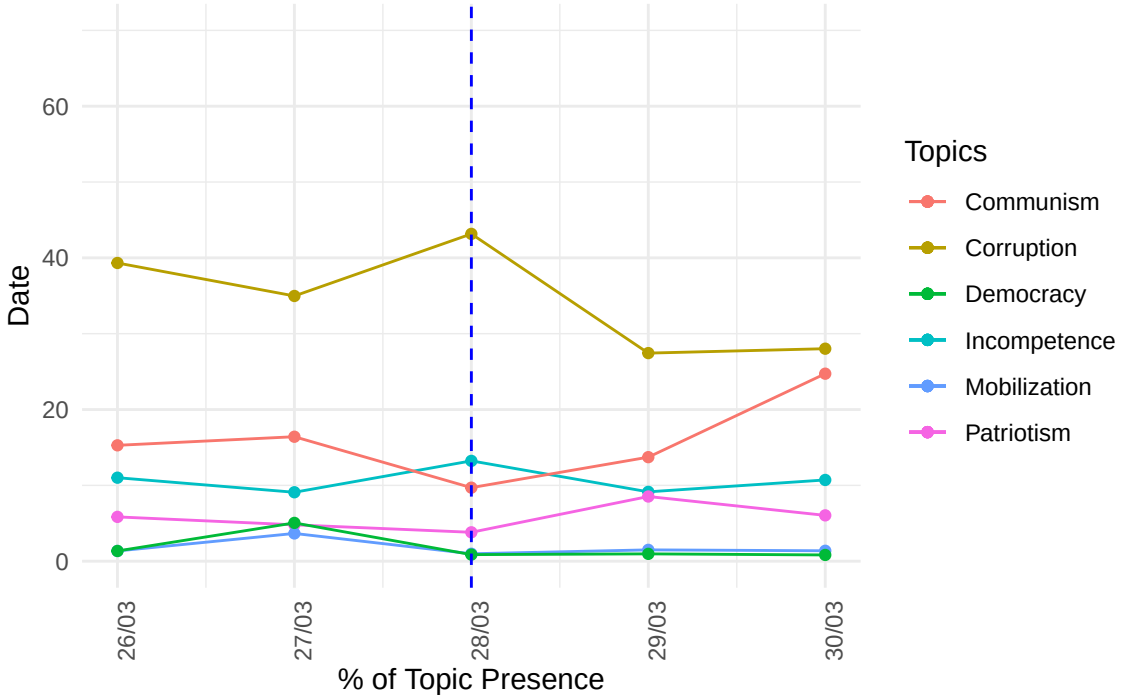

Figure 4: *Daily % of Topics Around the Impeachment (March, 27th - 30th)*

topics. Moreover, in general terms, both the `impeachment` dataset and the `post_impeachment` dataset show that corruption is the most important issue for those who want Castillo out of the presidential seat. Moreover, this topic seems to have increased in the period that I have referred to as "around the impeachment debate" (March 27th - 30th).

From a methodological point of view, I think I would need to look for other text analysis methods that work better on short, informal, and "all-over-the-place" texts such as tweets. Although the analysis still gave some interesting insights, perhaps the topic modeling could have been more useful with another method. Finally, it would be worth continuing this analysis with more recent data, as the Peruvian political landscape has become very turbulent in the past days. I was not able to include more recent tweets because of time limitations, but I believe that an extention of this topic analysis could result in an interesting timeline of what is, perhaps, an impeachment foretold.