YouTube communication of Italian Political Parties

Maria Ascolese

March 11, 2022

Contents

1 Introduction

In this paper I use the `YouTube Data API v3` to collect data and metadata from the official YouTube channels of major Italian Parties in order to investigate content and language occurrences in YouTube videos.

The goal of this short analysis would be to compare Parties' topics and language of choice, and I believe that their YouTube channels might be accurate small representations of their communicative goals, as all video content on the channels has been necessarily chosen and published by the channel owner and, therefore, intentionality must be assumed (Vesnic-Alujevic & Van Bauwel, 2014).

2 Hypothesis

H1 Parties with common ideologies also show common keywords, emotional words and syntax.

3 Data

3.1 Channel data

I used the R package `Tuber` to collect YouTube videos data, as it consists of a number of functions useful to extract data from the YouTube platform via the YouTube Data API v3. I conducted the queries using each political party's ID in the `list_channel_videos` function, which resulted in obtaining metadata for each video published on every channel. The tool provided metadata for: - 20000 videos from **Movimento 5 Stelle**'s channel; - 9548 videos from **Lega Salvini Premier**'s channel; - 1955 videos from **Partito Democratico**'s channel; - 5524 videos from **Fratelli d'Italia**'s channel; - 60 videos from **Forza Italia**'s channel.

3.2 Video data

Since `Tuber`'s function to extract subtitles from YouTube videos has been recently disabled, I used the R package `youtubecaptions` instead. The function `get_caption` gives back a number of string characters for each video, which correspond to the unique captions appearing on the screen as the video plays. However, the `get_caption` function needs URLs, which I produced with a basic YouTube video URL and video IDs.

```
#Example of URLs produced for Fratelli d'Italia's videos

fdi_url <- str_c("https://www.youtube.com/watch?v=",
                 fdi_ch$contentDetails.videoId[1:5524])


#Example of for loop, which I used to get the captions for Partito Democratico's videos

for (i in pd_data$pd_url[1:1955]) {
  tryCatch({
    v <- get_caption(i, "it")
  }, error = function(e){})

  print(i)
  pd_captions <- append(pd_captions, v)
  Sys.sleep(1)
}
```

Scraping the video captions with these for-loops was a bit tricky. When I first scraped the Partito Democratico's captions, the for-loop was *extremely* slow and it even stopped a few times - that is, because some videos did not inclued captions. Then I added `tryCatch`, but I did not specify the `e` function, as I didn't want to risk to make the for-loop even slower.

The Partito Democratico's captions alone, with only 1955 videos, took 3 hours. Obviously, I got scared that the bigger channels (Lega, 9548 videos; Movimento 5 Stelle, 20000 videos) could rob me of days. So, I searched for possible reasons why my for-loop was so slow, and made a few experiments with the `Sys.time` to try and make it faster. I even considered conducting the loops in Amazon Web Services' Cloud, but I figured that it would not solve a thing: ultimately, the slowness depended on the function `get_caption`. I just accepted that I needed to leave the loops run for a couple of days.

When I finally got the captions, I then joined them in single string characters, obtaining whole transcriptions for each video of the channels.

4 Methods

4.1 Tokenization

After preparing the data in order to have clean, tidy datasets for each political party's channel, I finally came to the analysis, for which I used the `tidytext` package.

For each party's channel, I split the captions in words in order to remove stopwords (like conjunctions and auxiliary verbs, but also some adverbs that I manually added to the list), I removed encoding errors (like some accent marks) and some YouTube transcription errors (like numbers being transcripted with digits instead of words). When I finally ended up with clean lexemes of interest, for each party I obtained the `count` of each word and sorted them.

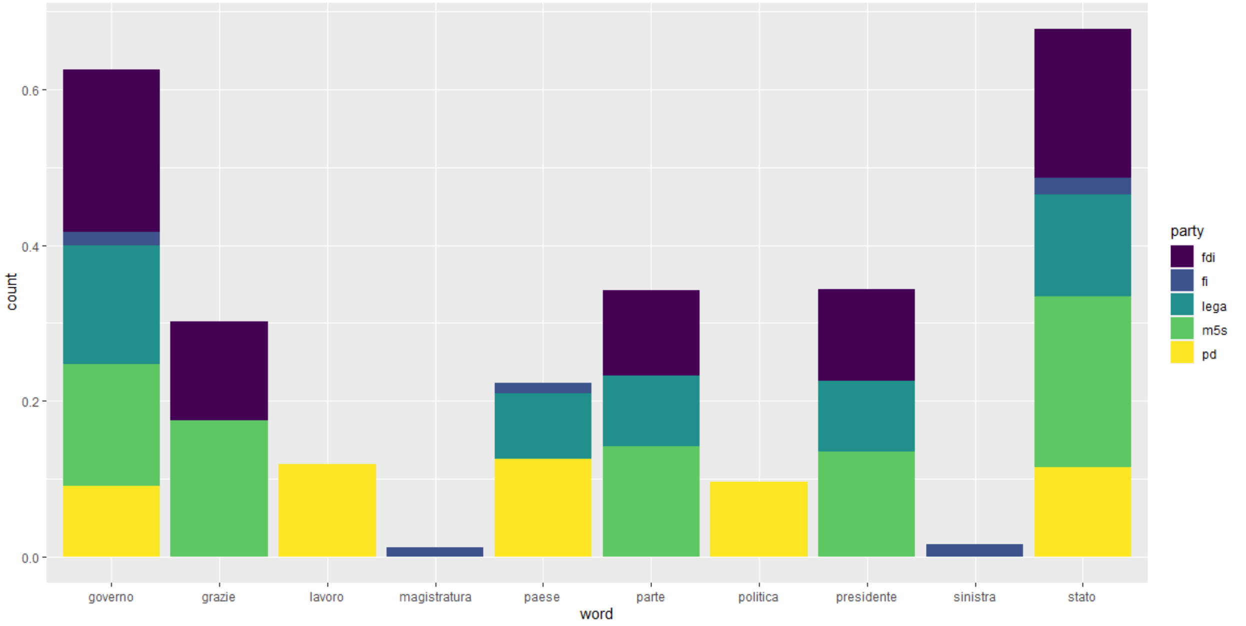
```r
#Example of tokenization conducted on Partito Democratico's transcriptions.
#I actually used two stopwords removal processes, which can be seen completely in the RScripts

text <- data_frame(Text = pd_dataset$text)

words <- text %>%
  unnest_tokens(output = word, input = Text)

wordcounts <- words %>%
  count(word, sort = TRUE)

for(i in 1:nrow(wordcounts)) {
  wordcounts$word[i] <- removeWords(wordcounts$word[i], sw_ita)

  print(i)
}

wordcounts <- wordcounts[-which(wordcounts$word == ""), ]
```

At a first glance, it appeared that the most frequent words were common between the parties: **governo**, **paese** and **lavoro**, to name a few. At this point, I was ready to visualize the data.

4.2 Visualization

In order to look into my hypothesis, I wanted to compare the frequency of words between parties. However, the number of videos uploaded by each party varies significantly: to make an example, the word **paese** occurs 63 times in Forza Italia's videos and 5794 times in Lega's videos, but it actually is the fifth most recurrent word in both channels. This great difference depends on the difference of content, as Forza Italia's channel only consists of 60 videos - against the 9548 videos of Lega's channel.

To make these numbers impartial and comparable, I decided to get the in-group relative frequency of occurrence of each word. For example, Forza Italia's **paese** percentage of appearance is **63/4646 = 0.013%**, which is comparable to Lega's **paese** appearance of **5794/68594 = 0.084%**. With these comparable results, I finally visualized the data.



5 Results

By comparing absolute and relative frequency of the words in each party's channel, I did not find any interesting similarity or significant difference. Actually, after removing stopwords, the most frequent words for each party were those of the very party's name (e.g., "lega" was the most recurring word for Lega, "forza" and "italia" were the most recurring words for Forza Italia, etc). After I removed these words too, as they weren't interesting to me, the parties' words turned out to be all very similar.

6 Conclusion

Ultimately, I could not really prove or reject my hypothesis. I believe that with more time and a deeper knowledge of text and content analysis, interesting phenomena could have been found in my data; I'm not sure if my hypothesis would have been necessarily proven to be true, but I feel that a lexical pattern does exist in the channels, at least on an in-group level. Perhaps, the fact that this research focused on a between-groups comparison was a bit of a stretch.

In the end, I can say that working with Italian political parties language is interesting and represents an exciting field to work on. Content patterns and lexical choices that I grasp from simply reading or listening to politicians' speeches become measurable and comparable, which is something that really fascinates me.

From a technical perspective, while working on this project I learned to deal with many new tools. Above all, I must name for-loops: only one month ago, they seemed the most advanced-level, unreachable tool to master, but now I can tell I finally grasped the concept and learned to write them on my own (at least, simple for-loops as the ones in my scraping RScript). Also, scraping was sort of exciting to me. A couple of times, I even needed to obtain some information from raw source pages: to me, it was fun to work on the naked content of familiar platforms as YouTube.

Concluding this project, my takeaway is that learning to scrape paves the way to great and exciting opportunities, as it allows you to virtually reach any data you're interested in. I believe that the APIs, particularly, opened a great deal of intriguing fields of research. The IBM Watson APIs, like the Watson Natural Language Understanding that I tried in order to conduct this research, seem like extremely powerful tools and, to be honest, are pretty easy to look into; and even if I did not end up using this last API, mostly because of time limitations, working on this project unveiled a whole world of research opportunities.

7 References

- GARZIA, D., *Personalization of politics between television and the internet: leader effects in the 2013 Italian parliamentary election*, Journal of Information Technology & Politics, 2017
- HANSON, G., HARIDAKIS, P. M., WAGSTAFF CUNNINGHAM, A., SHARMA, R., PONDER, J. D, *The 2008 Presidential Campaign: Political Cynicism in the Age of Facebook, MySpace, and YouTube*, Mass Communication and Society, 2014
- VESNIC-ALUJEVIC, L., VAN BAUWEL, S., *YouTube: A Political Advertising Tool? A Case Study of the Use of YouTube in the Campaign for the European Parliament Elections 2009*, Journal of Political Marketing, Ghent University, 2014